Statistics Surveys
Vol. 6 (2012) 142–228
ISSN: 1935-7516
DOI: 10.1214/12-SS102

A survey of Bayesian predictive methods for model assessment, selection and comparison*†

Aki Vehtari and Janne Ojanen

Aalto University
Department of Biomedical Engineering and Computational Science (BECS)
e-mail: Aki.Vehtari@aalto.fi; Janne.Ojanen@aalto.fi

Abstract: To date, several methods exist in the statistical literature for model assessment, which purport themselves specifically as Bayesian predictive methods. The decision theoretic assumptions on which these methods are based are not always clearly stated in the original articles, however. The aim of this survey is to provide a unified review of Bayesian predictive model assessment and selection methods, and of methods closely related to them. We review the various assumptions that are made in this context and discuss the connections between different approaches, with an emphasis on how each method approximates the expected utility of using a Bayesian model for the purpose of predicting future data.

AMS 2000 subject classifications: Primary 62-02; secondary 62C10.
Keywords and phrases: Bayesian, predictive, model assessment, model selection, decision theory, expected utility, cross-validation, information criteria.

Received October 2012.

Contents

*This is an original survey paper.
†This paper was accepted by Elja Arjas.

1. Introduction

The aim of this survey is to provide a unified decision theoretic review of Bayesian predictive model assessment, selection and comparison methods, and of methods closely related to them. Bayesian decision theory gives a natural definition for the assessment of the *predictive performance* of a statistical model as well as a comparison of several models by their predictive performance as formal decision problems.

In science the usefulness of a theory is tested by performing predictions for observations made in an experiment able to falsify the said theory. Significant discrepancies between observations and predictions suggest that the theory, or in our narrower view the statistical model, is not useful. In decision theory the correspondence between predictions and observations is described by a utility function, whose values are computable given predictions and observations. The *predictive performance of a model* is defined to be the utility evaluated at one or several future observations. The definition of predictive performance as utility is equivalent to *generalization ability*, a concept often used in the machine learning literature.

When the future observations are not available the predictive performance can be estimated by computing the *expected predictive performance* (expected utility) given a belief model for the future observations. Expected predictive performance is a useful quantity in assessing a single model. Indeed, if a model does not give reasonable predictions, there is usually not much sense in trying to infer on its parameters. Furthermore, a set of models can be compared against each other according to their expected predictive performance. Most of the methods reviewed in this survey have been advocated as tools for the common task of model selection, but we also discuss their use for the purpose of assessing the predictive performance of a model.

In the Bayesian statistical framework all aspects that appear relevant to the modeling problem should be described by probability models. For an answer, one finally integrates away all uncertain quantities, with respect to their conditional distribution given the data. In prediction problems the key quantity arising from the Bayesian theory is the posterior predictive distribution, that is, the distribution of the yet unobserved future observations conditioned on the data. It is a generally held view that one should use models that are rich enough to capture all essential uncertainties including, when in doubt, the model structure. Generally speaking, we agree with G. E. Box's famous quote *"All models are wrong, but some are useful"*. Even rich models are wrong in the sense that they do not fully correspond to the mechanisms in Nature that generated the data, but their usefulness can be assessed by evaluating their predictive performance or by some other model criticism approach.

A common opinion, and one shared by the authors, following from adopting the Bayesian statistical framework and using a rich enough model, is that as long as one is happy with the results from model criticism and predictive performance assessment, there is no need for model selection. Model selection can be a useful tool in tackling practical modeling problems, even though selecting

a more restricted model leads to ignoring uncertainties inherent in the initial model specification. For example, one may ask whether the predictive performance of a simple parametric model is practically as good as the performance of a complex non-parametric model. By using such a simple alternative it would be possible to reduce future measurement costs of the explanatory variables values and also alleviate communication of the model's essential features to other interested parties. In some cases one may also be willing to trade predictive performance against lower costs resulting from less demanding data collection and computation requirements. Also, when performing explanatory variable selection a common informal use of the model selection methods is to assess the predictive relevance of the covariates.

The framework for assessing and selecting models based on their predictive performance can be applied in different ways, for example, by defining different utility functions and prediction scenarios. Moreover, often the practical application of the formal decision theoretic concepts requires theoretical and computational approximations. The reviewed methods are presented in a common notation in order to make comparisons easier, especially since the decision theoretic assumptions on which these methods are based are not always clearly stated in the original articles. We also voice our own opinions on how we think the predictive framework should be applied. Finally, we stress that one should be aware that the behavior of many of the methods based on the predictive framework is not very well known in many applications, for example, when a model is selected from a very large set of candidate models. While some results exist, giving solid general advice on all aspects of predictive model assessment, selection and comparison is impossible in absence of sufficient quantitative comparisons of the presented methods.

In Section 2 the notation for a Bayesian predictive model and the concept of the actual belief model are defined. In Section 3 the Bayesian decision theoretic framework for model assessment and selection is reviewed. Important issues related to practical model comparison are discussed in Section 4. Predictive methods for model assessment, selection and comparison are reviewed in Section 5. The paper concludes with a discussion in Section 6.

2. Bayesian predictive model

We consider a prediction problem with an explanatory variable (covariate, input variable, predictor) x and an outcome variable (response, target, output variable) y. The same notation is used interchangeably for scalar and vector-valued quantities. The observed data are denoted by $D = \{(x_i, y_i)\}_{i=1}^n$ and the future observation by $(\tilde{x}, \tilde{y})$. An abbreviation $y_{(1:n)} = (y_1, \ldots, y_n)$ is used to avoid clutter in formulas.

From a predictivist Bayesian point of view (Bernardo and Smith, 1994) the main interest in statistical inference is inference about observable quantities such as the future observation $\tilde{y}$. Parametric models and updating beliefs about model parameters with Bayes's theorem provides a convenient framework for

determining the distribution of future observations. Given a model specification M, a Bayesian model consists of a statistical model $p(y, x|\psi, M)$ for observations and of a prior distribution $p(\psi|M)$ for the model parameters. We use the same notation for both the discrete and continuous distributions. In prediction problems it is common to specify the statistical model separately for y conditional on x,

$$p(y, \theta|x, M) = p(y|x, \theta, M)p(\theta|x, M), \tag{1}$$

where the distribution of y given x is parametrized with θ, and for x,

$$p(x, \varphi|M) = p(x|\varphi, M)p(\varphi|M), \tag{2}$$

where φ is the parameter of the distribution of x. Often the explanatory variable x is assumed to be given and the focus of the prediction problem is the conditional model in Eq. (1). This is the case with most of the methods reviewed in Section 5. In the following we treat the explanatory variable as a known quantity, and come back to the implications of conditioning on either fixed or random x in Section 4.3. Conditioning on observed data D by the Bayes' theorem results in the posterior distribution $p(\theta|D, M)$ for the model parameters, which in turn can be used to determine the *posterior predictive distribution*

$$p(\tilde{y}|\tilde{x}, D, M) = \int p(\tilde{y}|\tilde{x}, \theta, M)p(\theta|\tilde{x}, D, M)d\theta \tag{3}$$

describing beliefs about the future observation given the observed data D and the model M.

The Bayesian framework in itself is not sufficient to guarantee that a model is adequate for its designed purpose. For example, a grossly misspecified model may describe the actual problem very poorly. Assessing the adequacy of a model is often referred to as model criticism. Although predictive performance of a model is an important (if not the most important) aspect of model criticism, a multitude of different model criticism approaches and tools exists (see, e.g., Gelman et al., 1995; O'Hagan, 2003, and references therein). While model criticism is beyond the scope of this survey, its importance cannot be over-emphasized.

In complex real-world modeling situations a simple parametric model is often not flexible enough for building a satisfying belief model. Model averaging and non-parametric models are often used as a means of obtaining richer models. In a situation in which a set of alternative models $\{M_k\}_{k=1}^K$ and a corresponding prior $p(M_k)$ on that set have been specified, one can integrate over the models and thereby arrive at the *Bayesian model averaging (BMA)* (e.g. Hoeting et al., 1999) predictive distribution

$$p_{\mathrm{BMA}}(\tilde{y}|\tilde{x}, D) = \sum_{k=1}^{K} p(\tilde{y}|\tilde{x}, D, M_k)p(M_k|D), \tag{4}$$

where $p(M_k|D)$ are the posterior probabilities of the models M_k. In case of nested models, for example in covariate selection, where the encompassing model

can be reduced to any submodel by setting the parameters to specific values, the BMA predictive distribution over all the submodels can be equivalently formulated by placing a discrete prior probability for the said specific values and integrating over the parameters (e.g., George and McCulloch, 1993; Brown, Vannucci and Fearn, 1998). A rich class of belief models can be obtained when, instead of considering a finite number of alternative parametric models, a continuum of non-parametric models is specified by defining a prior on a suitable function space (e.g., O'Hagan and Forster, 2004, Ch. 13).

When a rich enough model, describing well the knowledge about the modeling problem and capturing the essential prior uncertainties, is constructed and there are no substantial deficiencies found in model criticism phase, we follow Bernardo and Smith (1994) and call such a model the *actual belief model*, and denote it by M_*. In other words, the predictive distribution $p(\tilde{y}|\tilde{x}, D, M_*)$ is a quantitatively coherent representation of our subjective beliefs about the unobserved future data. We also use the term *reference model* for M_* especially in a model selection context.

3. Predictive model assessment and selection as decision problems

The Bayesian framework offers a way of representing and revising beliefs. Inference on an unknown quantity, whether it is a future observation or a parameter of a statistical model, can be represented as a decision problem where a decision to choose a specified inference action is based on beliefs about the unknown quantity. Our formulation of predictive model assessment and selection as decision problems follows the ideas presented by Bernardo and Smith (1994) and Key, Pericchi and Smith (1999). For a general introduction to the Bayesian decision theory see references (Berger, 1985; Raiffa and Schlaifer, 2000; Robert, 2001, and references therein). A related discussion on model assessment and selection can be found, for example, in references (O'Hagan and Forster, 2004; O'Hagan, 2003; Bayarri, 2003; Gelfand, 2003).

In the context of statistical inference the components of the decision problem are the following:

- $a \in \mathcal{A}$, available decisions, actions or answers to the inference problem;
- $\omega \in \Omega$, the unknown states of the world;
- $u(a, \omega) : \mathcal{A} \times \Omega \to \mathbb{R}$, a utility function attaching a reward to each answer to an inference problem (decision) a when a state of the world ω obtains;
- $p(\omega|D)$, a specification of the current beliefs about the state of the world, represented as the posterior distribution conditioned on observations D.

Observing the state of the world ω allows to observe the utility for any a, and in particular, to determine the *optimal answer* $\hat{a}$ to an inference problem by maximizing the *observed utility*. In practice, the state of the world ω cannot be directly observed, or if it can, the decision must be made before the observations are available. The optimal decision under uncertainty about ω can be determined by maximizing the *expected utility*, that is, the expectation of the utility function

taken over the distribution $p(\omega|D)$. The observed utility is typically available only in experiments where generating new data is easy and cheap such as, for example, in a computer simulation with artificially created data.

Decision theory provides a unifying framework for describing the majority of the predictive model selection methods reviewed in Section 5. In Section 3.1 prediction is formulated as a decision problem, and the components of the prediction task are discussed in some detail. Predictive model assessment (evaluating the expected utility of the actual belief model) is discussed in Section 3.2, while in Section 3.3 predictive model selection (choosing a single model from a set of candidate models based on their estimated predictive performance) is considered. An alternative formulation of the prediction problem in terms of model parameters is presented in Section 3.4. Finally, in Section 3.5 some model selection approaches and concepts that are closely related to Bayesian predictive model selection are presented for the purpose of providing background information for many commonly used non-Bayesian predictive approaches. The outline of the Section is illustrated in Fig. 1.

3.1. Prediction as an inference task

We define prediction as a decision problem with the following components:

- the state of the world is a future observation $\tilde{y} \in \mathcal{Y}$;
- an answer $a \in \mathcal{A}$ to an inference problem in a prediction task is *a prediction for the future observation*, whose exact nature depends on the utility function and the specification of $\mathcal{A}$;
- utility function $u(a, \tilde{y})$, which defines a reward for predicting the unknown future observation $\tilde{y}$ with a;
- belief about the future observation, described by the posterior predictive distribution $p(\tilde{y}|D, M_*)$ of the actual belief model M_*.

In a prediction problem one aims to give as good a prediction[1] a as possible for an unknown future observation $\tilde{y}$, a set of several future observations $\tilde{y}_{(1:\tilde{n})}$ or some function of the future observation. The optimal prediction action $\hat{a}$ obtains from maximizing the expected utility. The expected utility depends on the utility function as well as on the actual beliefs concerning the future data, described by the predictive distribution $p(\tilde{y}|D, M_*)$. For the time being, it is assumed that expectations and other integrals with respect to $p(\tilde{y}|D, M_*)$ are readily available; implications from specifying the properties of M_* are considered in Section 4 and a number of practical definitions are discussed in Section 5.

Bayesian statistical decision theoretic literature is often concerned with finding the optimal decision under uncertainty over parameters θ. As discussed in Section 3.4, it is possible to reduce the prediction problem into this more common decision theoretic formulation by taking θ to be the unknown state of the world and defining a suitable utility function. In the spirit of the predictivist Bayesian view we prefer the definition directly in terms of $\tilde{y}$ in this survey, so

[1] An equivalent term "forecast" is often used in, for example, meteorology and economics.

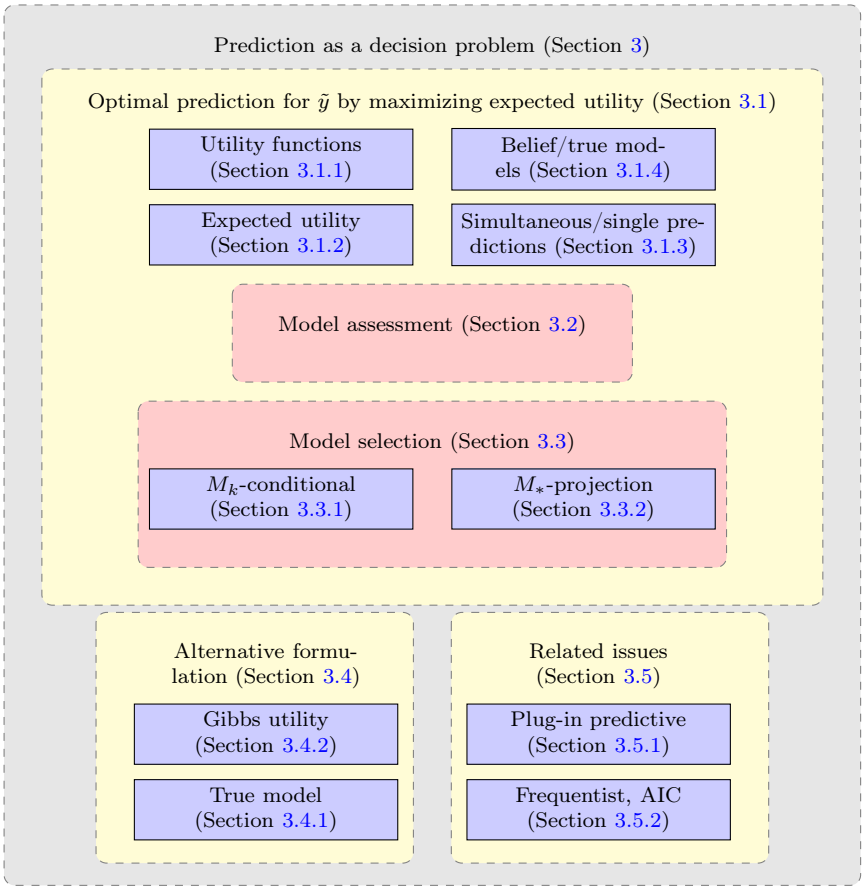

FIG 1. *Content for Section 3.*

that the value of the utility function is in principle observable; it can be evaluated by observing the future data. This is not the case when the prediction problem is defined in terms of θ, as typically the model parameters cannot be observed. Moreover, due to model-specific definitions of $p(\tilde{y}|D, M_*)$ some of the methods in Section 5 are straightforward to present in the $\tilde{y}$-formalism, whereas presenting them in terms of the θ-formalism would be unnecessarily complicated.

Without loss of generality the target of the prediction task is defined to be a single future observation $\tilde{y}$, unless explicitly stated otherwise. Also, in order to simplify the treatment in this Section only outcomes y are considered, and the additional considerations involved in the conditional modeling of y given an explanatory variable x are discussed in Section 4.3.

3.1.1. Utility functions for prediction

A utility function $u(a, \tilde{y})$ assigns a reward for the prediction action a when the future observation $\tilde{y}$ obtains. In statistical decision theory, it is common to define a *loss or cost function* $l(a, \tilde{y})$ instead of a utility function $u(a, \tilde{y})$. Without a loss of generality, one may write $l(a, \tilde{y}) = v(\tilde{y}) - u(a, \tilde{y})$, where $v(\tilde{y})$ is an arbitrary fixed function independent of a. In such a formalism, the optimal a can be obtained by minimizing the expected loss. Although we prefer terminology from the utility theory we also use the loss formalism when appropriate.

In a prediction task, the answer a is a prediction for the future observation. In *point prediction* (predictive point estimation or point forecasting) a single value $a \in \mathcal{A} \subseteq \mathcal{Y}$ representing the unknown future observation is reported. In *probabilistic prediction* (probabilistic forecasting) the aim is to report inferences about $\tilde{y}$ in such a way that the full uncertainty over $\tilde{y}$ is taken into account. In probabilistic prediction, the possible answers are probability distributions $a(\tilde{y}) \in \mathcal{A}$, where $\mathcal{A} \subseteq \mathcal{F}$ is a possibly somehow restricted set of probability distributions for $\tilde{y}$. In order to avoid unnecessary clutter in notation, the same symbol a is used to denote both the point predictions and the probabilistic predictions. The nature of a should be clear from the context, but if there is a possibility of confusion, it will be explicitly specified whether a is a point estimate or a distribution.

Preferably, the utility function u is specifically tailored for the application at hand, and it measures as correctly as possible the benefit (or cost) of predicting future data with the model. For example, Miyamoto (1999) reviews quality-adjusted life years (QALY) utility functions for combined survival duration and health quality, and Fouskakis and Draper (2008); Fouskakis, Ntzoufras and Draper (2009) discuss an example in which monetary utility is placed for the data collection costs as well as for the accuracy of predicting the mortality rate in a health policy problem. However, often explicit benefit or cost information is not available and the predictive performance of a model is assessed by utility functions commonly found to be appropriate in reporting scientific inference. Typically, these utility functions assign larger rewards (or smaller losses) to predictions close to future observations, with closeness defined in a specific mathematical sense.

Common utility functions in point prediction are scoring functions such as squared error, absolute error or absolute percentage error. Scoring functions are commonly formulated as loss functions in the literature. A good review of the most common scoring functions is presented by Gneiting (2011), who also discusses the desirable properties for scoring functions in prediction problems. We use the squared error as an example utility function for point prediction, because the squared error and its derivatives seem to be the most common scoring functions in predictive literature (Gneiting, 2011).

In probabilistic prediction the appropriate utility functions are scoring rules, such as quadratic, logarithmic and zero-one score, whose properties are reviewed by Gneiting and Raftery (2007) and Bernardo and Smith (1994) (who use the term score function). Bernardo and Smith (1994) argue that suitable scoring

rules for prediction are local and proper: a scoring rule is proper if it is maximized by the actual belief model, $a = p(\tilde{y}|D, M_*)$ and strictly proper if it is uniquely maximized by $p(\tilde{y}|D, M_*)$; and local if the value of the utility function depends on the unknown $\tilde{y}$ only through the value $a(\tilde{y})$. Propriety of the scoring rule ensures that the decision maker reports his true beliefs honestly, while locality incorporates the possibility that bad predictions for some $\tilde{y}$ may be judged more harshly than others. The logarithmic score proposed by Good (1952) is a good example of a utility function for probabilistic prediction. The logarithmic score is the unique (up to an affine transformation) local and proper score function (Bernardo, 1979), and appears to be the most commonly used utility function in model selection.

3.1.2. Expected utility and optimal decisions

In order to obtain the optimal prediction maximizing utility, one needs to be able to evaluate the value of the utility function depending on the considered future observation. As access to new observations is typically restricted, observed utility cannot be used as the basis for determining the optimal prediction $\hat{a}$. Instead, the observed utility can be estimated by the *expected utility*

$$\bar{u}_*(a) = \int u(a, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{5}$$

with the expectation taken over the posterior predictive distribution of the actual belief model $p(\tilde{y}|D, M_*)$, which describes the uncertainty of the future observation $\tilde{y}$ conditioned on the observed data D. We write the subscript in $\bar{u}_*$ to explicitly remind that the expectation is taken over the model M_*; in the following Sections expectations with respect to predictive distributions of other models are also encountered. Expected utility is a reasonable estimate for the observed utility if $p(\tilde{y}|D, M_*)$ is a good proxy for an actual set of future observations.

The decision to choose the optimal prediction $\hat{a}$, whether it is a point prediction or a probabilistic prediction, is made by maximizing the expected utility

$$\hat{a} = \arg\max_{a \in \mathcal{A}} \int u(a, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}. \tag{6}$$

The resulting *maximized expected utility* is then given by

$$\bar{u}_*(\hat{a}) = \int u(\hat{a}, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}. \tag{7}$$

For scoring rules the maximized expected utility function is sometimes referred to as information measure or entropy function. In this context the maximized expected utility can be written as $\bar{u}(\hat{a}, p)$, where $\hat{a} = \hat{a}(\tilde{y})$ refers to optimal prediction under model $p = p(\tilde{y}|D, M_*)$. An associated discrepancy or divergence function is defined as $d(a, p) = \bar{u}(p, p) - \bar{u}(a, p)$. Because the optimal prediction

$\hat{a}$ can be obtained by minimizing $d(a, p)$, discrepancies or divergences can be directly used as loss functions. For more information on information measures and discrepancies or divergences as loss functions, see Robert (1996); Bernardo (2005a,b); Gneiting (2011); Grünwald and Dawid (2004).

As logarithmic score or the squared error are used in the majority of the methods in Section 5, we use them as examples for illustrating the optimal predictions and the values of the maximized expected utility.

Logarithmic utility function proposed by Good (1952) is a widely-used scoring rule for probabilistic prediction when the unknown state of the world is the future observation $\tilde{y}$. Given any prediction $a(\tilde{y})$ the utility function is defined as the *logarithmic score*,

$$u(a, \tilde{y}) = \log a(\tilde{y}), \tag{8}$$

that is, the logarithm of the value of the probability distribution at the observation $\tilde{y}$. The logarithmic score is a strictly proper and a local score function, and therefore a good choice as a utility function for prediction (Robert, 1996).

The answer to an inference problem is to choose the optimal prediction $\hat{a}(\tilde{y})$ from the set of all probability distributions $\mathcal{F}$. The expected utility

$$\bar{u}_*(a) = \int \log a(\tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y} \tag{9}$$

is maximized by $\hat{a}(\tilde{y}) = p(\tilde{y}|D, M_*)$. That is, because $p \in \mathcal{F}$ the p-optimal (or equivalently, M_*-optimal) prediction is the predictive distribution of the actual belief model itself. The maximized expected utility for the logarithmic utility function

$$\bar{u}_*(\hat{a}) = \int \log p(\tilde{y}|D, M_*) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{10}$$

is the *negative entropy* of the predictive distribution $p(\tilde{y}|D, M_*)$. The associated divergence function is the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951)

$$d_{\mathrm{KL}}\left\{p(\tilde{y}|D, M_*), a(\tilde{y})\right\} = \mathrm{KL}(p||a) = \int \log \frac{p(\tilde{y}|D, M_*)}{a(\tilde{y})} p(\tilde{y}|D, M_*) d\tilde{y}. \tag{11}$$

Minimizing the KL-divergence with respect to $a \in \mathcal{F}$ gives an equivalent result as maximizing the expected logarithmic utility in Eq. (9). The KL-divergence is often used as a loss function in model selection literature. The connection of the logarithmic utility function and the KL divergence illustrates the intuitive fact that the utility of prediction is high when the prediction is close to the predictive distribution of the actual belief model.

The logarithmic utility function is a good choice in prediction problems where properties such as asymmetry or tail thickness of the predictive distribution may be important, because the optimal probabilistic prediction is the posterior predictive distribution of the actual belief model.

Squared error is an example of a utility function in situations in which the decision action is to choose a point prediction $a \in \mathcal{Y}$ for the future observation $\tilde{y}$. Squared error and its many derivatives are common choices in the statistical literature. It may be also considered as a quadratic approximation to the general class of convex loss functions.

The squared error utility function depending on a point prediction a can be defined through the *squared error* or *quadratic loss*

$$s(a, \tilde{y}) = (a - \tilde{y})^2 \tag{12}$$

as $u(a, \tilde{y}) = -s(a, \tilde{y})$. However, we do not convert the quadratic forms into utility functions because the squared error is typically presented as a loss function in the literature.

The optimal point prediction $\hat{a}$ minimizing the expected loss (negative expected utility)

$$\bar{s}_*(a) = \int (a - \tilde{y})^2 \, p(\tilde{y}|D, M_*) d\tilde{y} \tag{13}$$

can be shown to be the posterior predictive mean

$$\hat{a} = \mathbb{E}\left[\tilde{y}|D, M_*\right] = \int \tilde{y} \; p(\tilde{y}|D, M_*) d\tilde{y} \tag{14}$$

and the expected loss for the optimal prediction $\hat{a}$ is

$$\bar{s}_*(\hat{a}) = \int \left(\tilde{y} - \mathbb{E}\left[\tilde{y}|D, M_*\right]\right)^2 p(\tilde{y}|D, M_*) d\tilde{y} = \operatorname{var}\left[\tilde{y}|D, M_*\right], \tag{15}$$

the *variance of the posterior predictive distribution* $p(\tilde{y}|D, M_*)$.

The squared error loss results in a point estimate that incorporates information about the location of the predictive distribution. In other words, formulating a prediction problem with the squared error loss is equivalent to regarding the two first central moments of the predictive distribution as the important information for predicting future observations. That is, only the location and scale of the predictive distribution are considered important, while other properties, such as skewness or kurtosis, do not directly affect the evaluation of the expected utility.

3.1.3. Single and simultaneous prediction

Although in principle it makes no difference whether the unknown state of the world is defined to be a single future observation $\tilde{y}$ or a set of $\tilde{n}$ future observations $\tilde{y}_{(1:\tilde{n})}$, it is useful to make a difference between *single prediction* where the uncertainty of a single future observation is described by the *marginal predictive distribution* $p(\tilde{y}|D, M_*)$ and *simultaneous prediction*, where the uncertainty about $\tilde{n}$ future observations is described by the *joint predictive distribution* $p(\tilde{y}_1, \ldots, \tilde{y}_{\tilde{n}}|D, M_*)$. The theory so far has been formulated for predicting a

single future observation, but the theoretical framework for solving the simultaneous prediction problem is exactly the same. The chain rule representation of the joint predictive distribution

$$p(\tilde{y}_1, \ldots, \tilde{y}_{\tilde{n}}|D, M_*) = \prod_{i=1}^{\tilde{n}} p(\tilde{y}_i|\tilde{y}_{(1:i-1)}, D, M_*), \tag{16}$$

where $\tilde{y}_{(1:0)} = \emptyset$ and $\tilde{y}_{(1:1)} = \tilde{y}_1$, illustrates that simultaneous prediction is equivalent to $\tilde{n}$ consecutive single predictions with the posterior distribution updated after each new observation.

In practice, instead of a simultaneous prediction several single predictions are often made using the marginal predictive distributions $p(\tilde{y}_j|D, M_*)$, $j = 1, \ldots, \tilde{n}$. From Eq. (16) it is evident that generally the joint predictive distribution for a sample of size $\tilde{n}$ is different from the product of $\tilde{n}$ marginal predictive distributions

$$\prod_{i=1}^{\tilde{n}} p(\tilde{y}_i|\tilde{y}_{(1:i-1)}, D, M_*) \neq \prod_{j=1}^{\tilde{n}} p(\tilde{y}_j|D, M_*). \tag{17}$$

The methods based on the marginal predictive distributions instead of simultaneous prediction are commonly used, because 1) $\tilde{y}_{(1:\tilde{n})}$ are not observed in the immediate future and thus updating of the posterior is not possible during prediction – for example, an automatic digit recognition system for postal codes in letter addresses does not get instant information about the correct digit classification, 2) $\tilde{n}$ is unknown – model assessment or selection would be affected by an arbitrary selection of $\tilde{n}$, so instead average performance for single prediction can be estimated, 3) for many models the marginal predictive distributions are easier to compute than the joint predictive distribution, 4) some utility functions do not make a difference between marginal and joint predictions, and 5) an approximation made during the estimation of the predictive performance makes the difference between the marginal and joint predictions to disappear.

3.1.4. On belief models and true models

In the predictive Bayesian approach, the optimal decisions are made by maximizing the expected utility. The uncertainty over the future observations is described by the data-dependent actual belief model $p(\tilde{y}|D, M_*)$. In a strict subjective Bayesian view there is only one set of data D available and all inference is conditioned on these observations.

Sometimes a concept of the *true model* representing the actual data generating machinery is proposed. In a typical theoretical treatment only very general properties of the true models are specified. For example, the observations are assumed to be independently subject to the same probability distribution $p_t(\cdot)$. We wish to emphasize the difference between the true model and a Bayesian belief model. The Bayesian belief model is the result of learning from data under

uncertainty. That is, a probabilistic model is used to represent both the inherent uncertainties and the lack of information in the modeling task. Furthermore, we are not required to assume that the target of the modeling task is random. On the other hand, the properties of the true model are specified by the modeller a priori, and they are not learned from the data. Many model selection approaches have been formulated based on the idea of locating a model close to the true model, for example, in the KL-divergence sense. However, in order to obtain an operational approach any computation involving the concept of a true model needs to be approximated somehow: either the true model is estimated from the data or represented as a proxy sample of the observations. Under such approximations the boundary between the true model and the actual belief model becomes blurred.

It is our view that postulating the existence of a true model and the associated probability distribution $p_t(\cdot)$ should not be done in order to provide a way of constructing practical operational statistical inference or model selection approaches. Rather, assuming a true model allows us to study the theoretical properties of the subjective data-driven approaches based on belief models. For example, knowing $p_t(\tilde{y})$, the true distribution of the future observation, we may define the *generalization utility*

$$\bar{u}_t(\hat{a}) = \int u(\hat{a}, \tilde{y}) p_t(\tilde{y}) d\tilde{y} \tag{18}$$

for assessing the predictive performance of the M_*-optimal prediction $\hat{a}$ over all possible future observations. Furthermore, existence of the true model allows to consider statistically how variations in the observed data set D affect the predictions and the corresponding expected utilities. That is, one can take expectations over $p_t(D)$ of an expected utility such as

$$\mathbb{E}_D \left[\bar{u}_*(\hat{a}|D) \right] = \int \bar{u}_*(\hat{a}|D) p_t(D) dD, \tag{19}$$

where the dependence on the observations in the expected utility in Eq. (5) is explicitly shown, and define quantities such as bias and variance of $\bar{u}_*(\hat{a}|D)$. These *frequency properties* are considered in more detail in Section 4.6.

In the extreme case the true model $p_t(\cdot)$ is a known and completely specified quantity. For example, in simulation experiments knowledge on the data-generating machinery allows to evaluate the true generalization utility as well as the frequency properties of the utility estimate for any model either analytically or by sample-based Monte Carlo (MC) approaches.

If the existence of the true model has been postulated, an intuitive idea in statistical learning is that with increasing number of observations the data-dependent model should resemble the data-generating distribution more and more closely. In this sense it may be argued that the actual belief model is an estimate for the unknown true model. Formalizing these ideas requires defining certain properties of the true model in relation to the proposed statistical models. We follow the categorization in (Watanabe, 2009). In the case of a *realizable*

and regular true model it may be assumed that the true model is included in the proposed model space $p_t(\cdot) \in \{p(\cdot|\theta, M) : \theta \in \Theta\}$; that is, a unique parameter θ_0 exists such that $p_t(\cdot) = p(\cdot|\theta_0, M)$. If, in addition, the Fisher's information matrix is positive definite the true model is also *regular* for $p(\cdot|\theta, M)$. In a case of realizable and regular true model, the posterior distribution for the belief model parameters converges to a single point, and we may assess, for example, the asymptotic properties of how the Bayesian learning framework estimates the underlying true model. However, the true model $p_t(\cdot)$ may be *unrealizable*, which means that although there is no unique parameter such that $p_t(\cdot) = p(\cdot|\theta_0, M)$, nevertheless a particular parameter value minimizing $d_{\mathrm{KL}}\{p_t(\cdot), p(\cdot|\theta, M)\}$ exists. Furthermore, for a *singular* true model there is a set Θ_0 such that for the parameter values $\theta \in \Theta_0$ the KL-divergence $d_{\mathrm{KL}}\{p_t(\cdot), p(\cdot|\theta, M)\}$ is minimized. For singular models the posterior distribution does not converge to a single point, but to an algebraic or analytical set. Many common statistical models (for example, mixture models, latent variable models, Bayes networks, hidden Markov models and neural networks) are singular. Singular learning theory (Watanabe, 2009) provides a way of assessing the theoretical predictive properties of belief models in the singular case.

3.2. Model assessment: predictive performance of the actual belief model M_*

In this survey, the term *predictive model assessment* refers to evaluating the predictive performance of the actual belief model as the maximized expected utility. The definition holds regardless of whether the the main focus is in the estimation of the prediction performance or the prediction task is considered as a subcomponent of a more comprehensive decision problem. After satisfactory model criticism the model M_* can be considered to adequately represent the uncertainties involved in the prediction task, and the beliefs about the future observation can be described by the posterior predictive distribution $p(\tilde{y}|D, M_*)$. Obtaining the optimal prediction $\hat{a}$ and evaluating the maximized expected utility $\bar{u}(M_*, \hat{a})$, where the model label is now written explicitly, proceeds exactly as described in Section 3.1. Model assessment is presented as a stylized decision theoretic problem in Fig. 2.

When the prediction task is a part of a larger decision theoretic problem the maximized expected utility is typically used to guide the decision maker in choosing a course of action which is not directly involved with assessing the predictive performance of the actual belief model. When the prediction task is not a subcomponent of a broader formal decision problem, the purpose of model assessment is to report to the application expert the optimal prediction under the model and the corresponding assessment of the predictive performance of the said model.

Even in the latter case there is inherently an informal decision problem involved in reporting the expected utility of the model. Typically the application expert uses the estimate of the predictive performance to decide whether large

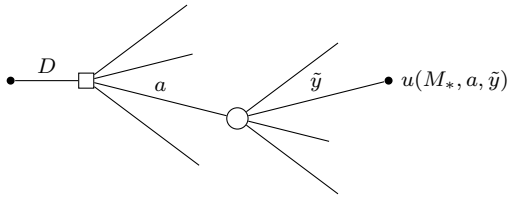

FIG 2. *Predictive model assessment represented as a stylized decision theoretic problem. Given the actual belief model M_* the only decision is to choose the prediction a maximizing the expected utility under $p(\tilde{y}|D, M_*)$. Random nodes are represented by circles and decision nodes by squares.*

enough benefits can be obtained by predicting future observations with the model. For example, in case of monetary utility the reduction in costs or increase in profits resulting from better predictions may not be significant enough to warrant taking the predictive model into use. In such case there is an inherent comparison to some baseline activity, such as "make predictions completely at random", "use constant prediction for future regardless of obtained information" or "keep doing things in the old way". To alleviate this comparison, the predictive performance should be reported in such a way that the application expert can understand the significance of the result in light of *external information* that is not included in the model (e.g., Gelman et al., 2003).

When one of the common utility functions for reporting scientific inference is used, the resulting expected utilities such as the average logarithmic scores can be rather unintuitive. Although such utility functions may not be useful in model assessment, they can still be highly useful in model selection and comparison.

3.3. Model selection: prediction with a candidate model M_k other than the actual belief model M_*

Predictive model selection refers to a decision problem where a single model with the best predictive performance is selected from a set of *candidate models* $\{M_k\}_{k=1}^K$. The unknown state of the world is again the future observation $\tilde{y}$ and the beliefs about the future observation are described by the posterior predictive distribution $p(\tilde{y}|D, M_*)$ of the actual belief model. The formal decision problem involves two sequential decisions: after selecting the model M_k the decision maker selects a prediction $a_k \in \mathcal{A}_k$ depending on the selected model. The predictive performance of a candidate model M_k is given by the maximized expected utility

$$\bar{u}(M_k, \hat{a}_k) = \int u(M_k, \hat{a}_k, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{20}$$

where $\hat{a}_k$ is the optimal prediction under the model M_k. Model selection is done by ranking the candidate models based on their expected utilities, so that the

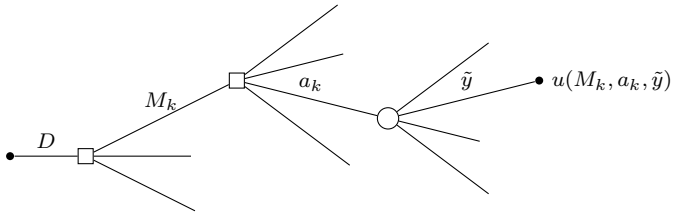

FIG 3. *Formal representation of the predictive model selection task as a decision problem. The choice of the candidate model M_k is followed by a choice of the model-dependent prediction a_k. The predictive distribution of the actual belief model $p(\tilde{y}|D, M_*)$ describes the beliefs about the future observation.*

optimal model choice results from a maximization

$$\hat{M} = \arg\max_k \bar{u}(M_k, \hat{a}_k). \tag{21}$$

Predictive model selection is presented as a formal decision problem in Fig. 3.

Utility functions can be tailored individually for each specific model selection problem. However, instead of specifying an application-specific utility function a more common approach is to rely on utility functions suitable for statistical inference. The most widely-used examples in the literature are the logarithmic score and the squared error. The lack of a clear interpretation for the expected utility values is not a serious issue as long as the main goal in model selection is to order the candidate models with respect to their predictive performance. Utility functions can also be constructed by adding a term depending on the model structure to the utility depending on the prediction accuracy,

$$u^{(c)}(M_k, a_k, \tilde{y}) = u(M_k, a_k, \tilde{y}) + c(M_k). \tag{22}$$

For example, in input variable selection $c(M_k)$ may be a term favoring models with a small number of covariates and penalizing more complex models with a larger number of covariates.

A complete specification of model selection as a decision problem requires defining the space of possible predictions $\mathcal{A}_k$ for each model M_k. Roughly speaking, the effect of the model M_k is taken into account in two ways. In the *reference predictive* model selection (Section 3.3.1) the predictions are M_k-optimal over a common $\mathcal{A}_k = \mathcal{A}$ whereas in the *projection predictive* model selection (Section 3.3.2) predictions are M_*-optimal with each $\mathcal{A}_k$ restricted in a meaningful way by the respective M_k. Although both approaches adhere to the same formal decision theoretic framework, the distinction is useful because the reference predictive approach requires a complete definition of priors $p(\theta_k|M_k)$ for every candidate model, whereas the projection predictive approach does not.

3.3.1. Reference predictive model selection: M_k-optimal prediction

The reference predictive model selection presented by Bernardo and Smith (1994) follows the general decision theoretic outline illustrated in Fig. 3. The prediction space for any model M_k is the space of all probability distributions $\mathcal{A}_k = \mathcal{F}$. However, the optimal prediction $\hat{a}_k$ for a model M_k is determined by a maximization

$$\hat{a}_k = \arg\max_{a_k \in \mathcal{F}} \bar{u}_k(M_k, a_k) = \arg\max_{a_k \in \mathcal{F}} \int u(M_k, a_k, \tilde{y}) p(\tilde{y}|D, M_k) d\tilde{y}, \tag{23}$$

where the expectation is taken with respect to the posterior predictive distribution of the model M_k obtained by the standard Bayesian treatment. That is, the optimal prediction is selected as if the predictive distribution $p(\tilde{y}|D, M_k)$ described the actual beliefs about $\tilde{y}$.

Given the model M_k and the M_k-optimal prediction $\hat{a}_k$ the expected utility of the model M_k is defined as

$$\bar{u}_*(M_k, \hat{a}_k) = \int u(M_k, \hat{a}_k, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{24}$$

where $p(\tilde{y}|D, M_*)$ describes the actual beliefs for the future observation. The model with the best predictive performance is obtained by Eq. (21).

The optimal prediction for each candidate model is obtained without knowledge of any predictive properties of M_*, while the predictive performance for each model is computed as an expectation over $p(\tilde{y}|D, M_*)$. The term reference predictive reflects the role of M_* as a common yardstick for comparing the candidate models: M_* acts as a reference whose predictive properties are sought after in the candidate models.

The reference predictive approach requires a complete definitions of priors $p(\theta_k|M_k)$ for each candidate model. For example, in case of nested models a coherent specification of such priors may be difficult.

Example: logarithmic utility function The M_k-optimal prediction results from maximizing the expected utility

$$\bar{u}_k(M_k, a_k) = \int \log a_k(\tilde{y}) p(\tilde{y}|D, M_k) d\tilde{y} \tag{25}$$

with respect to all possible probability distributions, $a_k \in \mathcal{F}$. As discussed in Section 3.1.2 the optimal prediction is the posterior predictive distribution of the model M_k, $\hat{a}_k(\tilde{y}) = p(\tilde{y}|D, M_k)$. The expected utility for the model M_k with the M_k-optimal prediction is given by

$$\bar{u}_*(M_k, \hat{a}_k) = \int \log p(\tilde{y}|D, M_k) p(\tilde{y}|D, M_*) d\tilde{y}. \tag{26}$$

The maximized expected utility in Eq. (26) is equivalent up to a constant independent of M_k (negative entropy of the actual belief model) to the negative

Kullback-Leibler divergence between the predictive distribution of the actual belief model and the predictive distribution of the candidate model. Thus the maximization of the expected utility is equivalent to minimizing the Kullback-Leibler divergence between the actual belief model and the predictive distribution of model M_k.

Example: squared error Under the squared error loss function the M_k-optimal point prediction results from minimizing the expected loss

$$\bar{s}_k(M_k, a_k) = \int (\tilde{y} - a_k)^2 p(\tilde{y}|D, M_k) d\tilde{y} \tag{27}$$

over $a_k \in \mathcal{Y}$. The optimal point prediction can be shown to be the posterior predictive mean conditional on the model M_k,

$$\hat{a}_k = \int \tilde{y}\ p(\tilde{y}|D, M_k) d\tilde{y} = \mathbb{E}[\tilde{y}|D, M_k]. \tag{28}$$

The expected loss under the actual belief model M_* is given by

$$\bar{s}_*(M_k, \hat{a}_k) = \int (\tilde{y} - \mathbb{E}[\tilde{y}|D, M_k])^2 p(\tilde{y}|D, M_*) d\tilde{y}. \tag{29}$$

Straightforward manipulation of the expected loss in Eq. (29) leads to

$$\bar{s}_*(M_k, \hat{a}_k) = \mathrm{var}\,[\tilde{y}|D, M_*] + \left(\mathbb{E}[\tilde{y}|D, M_*] - \mathbb{E}[\tilde{y}|D, M_k]\right)^2, \tag{30}$$

from which it is evident that the model M_k minimizing the expected loss is the one whose predictive mean is closest to the predictive mean of the actual belief model in the squared error sense. The expected utility is the variance of the predictive distribution of the reference model (as in Equation (15)) plus the squared difference between the predictive means.

3.3.2. Projection predictive model selection: M_*-projected prediction

In *projection predictive* model selection the optimal prediction $\hat{a}_k$ under a candidate model M_k is the M_*-optimal prediction over $\mathcal{A}_k$, which is a set of possible predictions restricted by the model structure M_k. The optimal prediction $\hat{a}_k$ is obtained by maximizing the expected utility

$$\hat{a}_k = \arg \max_{a_k \in \mathcal{A}_k} \int u(M_k, a_k, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{31}$$

where the expectation is taken with respect to the posterior predictive distribution of the actual belief model. In other words, the prediction $\hat{a}_k$ is M_*-optimal, whereas in the reference predictive approach the predictions were M_k-optimal. The resulting maximized expected utility is given by

$$\bar{u}_*(M_k, \hat{a}_k) = \int u(M_k, \hat{a}_k, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{32}$$

which differs from Eq. (24) only by the definition of $\hat{a}_k$.

The key component in the projection predictive approach is the definition of $\mathcal{A}_k$. For example, in probabilistic prediction the space $\mathcal{A}_k$ can be restricted to parametric probability distributions $\{p(\tilde{y}|\theta_k, M_k) : \theta_k \in \Theta_k\}$, so that selecting the optimal prediction $\hat{a}(\tilde{y})$ becomes equal to selecting the optimal point estimate $\hat{\theta}$.

A major difference to the reference predictive approach in Section 3.3.1 is the possibility to avoid defining priors $p(\theta_k|M_k)$ for the candidate models M_k by treating the parameters of the candidate model as decision variables. Also, there are model selection approaches related to the projection predictive framework, which directly project the posterior distribution of the parameters of the actual belief model onto the parameter space of the candidate models.

Example: predictive point estimation with the logarithmic utility function Given a logarithmic utility function the optimal prediction $\hat{a}_k(\tilde{y})$ for the model M_k is determined by maximizing the expected utility

$$\bar{u}_*(M_k, a_k) = \int \log a_k(\tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{33}$$

where the expectation is taken with respect to the posterior predictive distribution of the actual belief model. The maximization is performed over the set of parametric models defined by M_k so that $a_k(\tilde{y}) \in \mathcal{A}_k = \{p(\tilde{y}|\theta_k, M_k) : \theta_k \in \Theta_k\}$. The maximization can be written equivalently in terms of the parameter as

$$\hat{\theta}_k = \arg\max_{\theta_k \in \Theta_k} \int \log p(\tilde{y}|\theta_k, M_k) p(\tilde{y}|D, M_*) d\tilde{y}. \tag{34}$$

Given the optimal prediction $\hat{a}_k(\tilde{y}) = p(\tilde{y}|\hat{\theta}_k, M_k)$ the expected utility for the model M_k is

$$\bar{u}_*(M_k, \hat{\theta}_k) = \int \log p(\tilde{y}|\hat{\theta}_k, M_k) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{35}$$

where the parameter is now explicitly written as the decision variable. In other words, the point estimate $\hat{\theta}_k$ is such that the parametric distribution $p(\tilde{y}|\hat{\theta}_k, M_k)$ is as close as possible to the posterior predictive distribution of the actual belief model in the KL-divergence sense. The predictive point estimation approach is illustrated in Fig. 4.

Example: predictive posterior approximation with logarithmic utility function The M_*-optimal prediction is determined by maximizing the expected utility

$$\bar{u}_*(M_k, a_k) = \int \log a_k(\tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{36}$$

where $a_k(\tilde{y}) \in \mathcal{A}_k$. The definition $\mathcal{A}_k = \left\{\int p(\tilde{y}|\theta_k, M_k) q(\theta_k) d\theta_k : q \in \mathcal{Q}\right\}$ requires specifying a set of posterior projections $q(\theta_k)$ belonging to a suitable

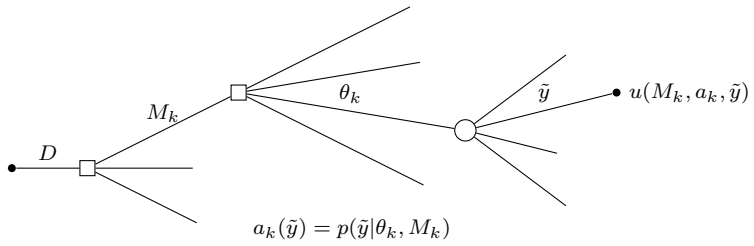

FIG 4. *Predictive point estimation: projection predictive model selection with the M_*-optimal selection for the model parameter θ_k.*

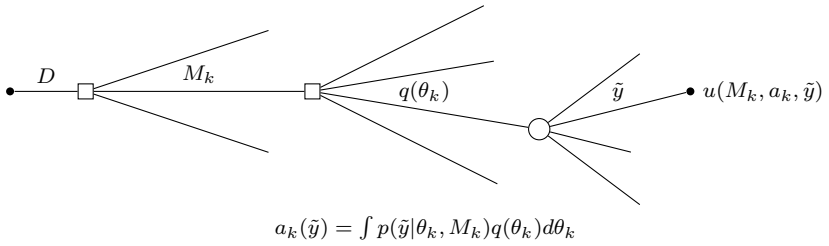

FIG 5. *Predictive model selection with optimal $q(\theta_k)$ projected from the actual belief model.*

restricted family of probability distributions $\mathcal{Q}$. For example, $\mathcal{Q}$ could consist of Gaussian distributions. The expected utility maximization can be written in terms of $q(\theta_k)$, so that the optimal posterior projection is given by

$$\hat{q}(\theta_k) = \arg \max_{q(\theta_k) \in \mathcal{Q}} \int \log \left(\int p(\tilde{y}|\theta_k, M_k) q(\theta_k) d\theta_k \right) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{37}$$

and the corresponding maximized expected utility is defined as

$$\bar{u}_*(M_k, \hat{q}) = \int \log \left(\int p(\tilde{y}|\theta_k, M_k) \hat{q}(\theta_k) d\theta_k \right) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{38}$$

where $\hat{a}_k(\tilde{y}) = \int p(\tilde{y}|\theta_k, M_k) \hat{q}(\theta_k) d\theta_k$ is the optimal prediction. The procedure is illustrated in Fig. 5.

The M_*-optimal posterior projection $\hat{q}(\theta_k)$ is not an approximation for the posterior distribution $p(\theta_k|D, M_k)$ of the model M_k. Instead, $\hat{q}(\theta_k)$ contains properties that are important in producing a prediction approximating the properties of the predictive distribution of the actual belief model. For example, in input variable selection the prediction $\hat{a}_k(\tilde{y})$ may contain information about input variables included in the structure of M_* but not M_k; the standard Bayesian treatment of the model M_k would disregard all information about variables not included in the model M_k.

Maximization with respect to the posterior projection $q(\theta_k)$ is not trivial, and we are not aware of any successful applications of this principle. See Section 5.4 for additional discussion.

3.4. Alternative formulation for prediction as an inference task: reduction to θ

Commonly in Bayesian statistical decision theory (see, for example, (Berger, 1985; Robert, 2001)) the components of a decision problem are the following:

- the unknown state of the world is the parameter $\theta \in \Theta$ of the sampling model $p(y|\theta, M_*)$;
- a decision or an answer to an inference problem $a \in \mathcal{A}$;
- utility function $u(a, \theta)$;
- beliefs about the unknown state of the world are described by the posterior distribution $p(\theta|D, M_*)$.

A prediction task can be reduced to a decision problem involving θ by specifying a suitable utility function assigning a reward to a prediction for a future observation $\tilde{y}$ when the state of the world θ obtains. A suitable utility function for predictive model selection can be formed, for example, as the expectation of a scoring function or scoring rule over the sampling distribution

$$u(a, \theta) = \int \tilde{u}(a, \tilde{y}) p(\tilde{y}|\theta, M_*) d\tilde{y}. \tag{39}$$

The optimal prediction can be obtained by maximizing the expected utility

$$\bar{u}_*(a) = \int u(a, \theta) p(\theta|D, M_*) d\theta. \tag{40}$$

For example, with the logarithmic score $\tilde{u}(a, \tilde{y}) = \log a(\tilde{y})$ the expected utility

$$\bar{u}_*(a) = \int \left[\int \log a(\tilde{y}) p(\tilde{y}|\theta, M_*) d\tilde{y} \right] p(\theta|D, M_*) d\theta \tag{41}$$

equals Eq. (9) for any probabilistic prediction a. As a trivial consequence the same optimal prediction $\hat{a}(\tilde{y}) = p(\tilde{y}|D, M_*)$ results from both formulations. We give an example of a model selection approach depending on an unknown θ in Section 3.4.1.

Mathematically the difference between the two formulations is merely the order of integration. In fact, the θ-formulation is probably more widely used in the literature. Especially when $u(a, \theta)$ is available analytically, calculating the posterior expectation (for example, using posterior samples) can be simpler than taking expectations over the predictive distribution. However, as stated in Section 3.1, we prefer the formulation based directly on the posterior predictive distribution, because it allows the utility function to be observable and, as discussed further in Section 4.1, provides notation for a larger class of predictive methods where the θ-formulation is less natural.

It is also possible to give a decision theoretic formulation of a model selection problem depending on the unknown θ in such a manner that the selection of prediction for the selected model M_k is not required (Gibbs utility, Section 3.4.2), or that the predictive properties of the selected M_k are not considered at all (zero-one utility, Section 3.4.3).

3.4.1. Predictive model selection when the unknown state of the world is the parameter of the actual belief model

When the unknown state of the world is a parameter of the sampling model the predictive model selection problem is typically formulated with the Kullback-Leibler utility function

$$u(M_k, a_k, \theta_*) = -\int \log\left(\frac{p(\tilde{y}|\theta_*, M_*)}{a_k(\tilde{y})}\right) p(\tilde{y}|\theta_*, M_*) d\tilde{y}, \tag{42}$$

which describes the utility of choosing a model M_k and a prediction a_k when the sampling model is $p(\tilde{y}|\theta_*, M_*)$. It is common to use the loss function form, so that the KL-divergence loss describes the loss of selecting a model different from the unknown sampling model; in the realizable regular case (Section 3.1.4) one may think of the loss of selecting a model different from the true model. The information about the unknown state of the world is described by the posterior distribution $p(\theta_*|D, M_*)$, whatever the set of candidate models is.

A special case arises when a prior $p(M_k, \theta_k)$ can be placed on *all possible* model specifications, especially when the set of candidates models is the same $\{M_k\}_{k=1}^K$. The information about the unknown state of the world is contained in the posterior distribution $p(M_k, \theta_k|D)$. Equivalently, it may be stated that the actual beliefs about the future observations are described by the BMA predictive distribution, Eq. (4).

Example: parametric point estimation with Kullback-Leibler divergence utility Given a sampling model $p(y|\theta_*, M_*)$, a prior $p(\theta_*|M_*)$ and observations D the unknown state of the world is described by the posterior distribution $p(\theta_*|D, M_*)$. The model selection problem, as illustrated in Fig. 6, requires selecting a model M_k and subsequently the parameter value θ_k so that the parametric model $p(\tilde{y}|\theta_k, M_k)$ is close to the sampling model depending on an unknown parameter θ_*. The closeness of the prediction to the sampling model is defined by the negative Kullback-Leibler divergence utility function

$$u(M_k, \theta_k, \theta_*) = -\int \log\left(\frac{p(\tilde{y}|\theta_*, M_*)}{p(\tilde{y}|\theta_k, M_k)}\right) p(\tilde{y}|\theta_*, M_*) d\tilde{y}. \tag{43}$$

Maximizing the expected utility

$$\bar{u}_*(M_k, \theta_k) = -\int \left[\int \log\left(\frac{p(\tilde{y}|\theta_*, M_*)}{p(\tilde{y}|\theta_k, M_k)}\right) p(\tilde{y}|\theta_*, M_*) d\tilde{y}\right] p(\theta_*|D, M_*) d\theta_* \tag{44}$$

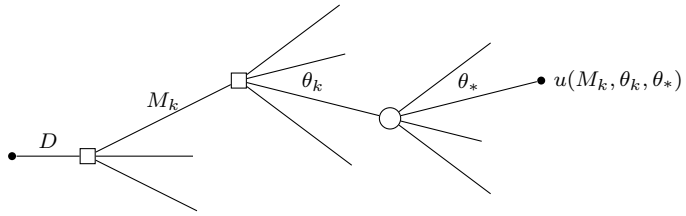

FIG 6. *Predictive model selection with the negative KL-divergence utility under the unknown parameter θ_* for the sampling model $p(y|\theta_*, M_*)$.*

with respect to θ_k results in the optimal point estimate $\hat{\theta}_k$ for the model M_k, and the model with the largest maximized expected utility $\bar{u}_*(M_k, \hat{\theta}_k)$ is the optimal model choice. Changing the order of integration in Eq. (44) and dropping terms constant with respect to M_k and θ_k leads to expected utility

$$\bar{u}_*(M_k, \theta_k) = \int \log p(\tilde{y}|\theta_k, M_k) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{45}$$

where $p(\tilde{y}|D, M_*) = \int p(\tilde{y}|\theta_*, M_*) p(\theta_*|D, M_*) d\theta_*$ is the predictive distribution of the actual belief model. The result is identical to that obtained in the parametric predictive point estimation example in Eq. (34), so that these approaches are equivalent with respect to the resulting prediction and model choice.

Example: Kullback-Leibler divergence utility and M_k-optimal prediction selection Consider a sampling model $p(\tilde{y}|\theta_{k'}, M_{k'})$ and a prior specification $p(\theta_{k'}|M_{k'})p(M_{k'})$, with a set of K discrete model structures $\{M_{k'}\}_{k'=1}^{K}$. Given observations D beliefs about the unknown state of the world are described by the posterior distribution $p(\theta_{k'}, M_{k'}|D)$.

A two step model selection approach, similar as in Section 3.3.1, is based on the negative Kullback-Leibler divergence utility function

$$u(M_k, a_k, M_{k'}, \theta_{k'}) = -\int \log\left(\frac{p(\tilde{y}|\theta_{k'}, M_{k'})}{a_k(\tilde{y})}\right) p(\tilde{y}|\theta_{k'}, M_{k'}) d\tilde{y}. \tag{46}$$

between the unknown sampling model $p(\tilde{y}|\theta_{k'}, M_{k'})$ and a prediction $a_k(\tilde{y})$ depending on a candidate model M_k. As illustrated in Fig. 7 the selection of a candidate model M_k is followed by a subsequent selection the prediction a_k. The M_k-optimal prediction following from maximizing the utility

$$\bar{u}_k(M_k, a_k, M_k) = -\int \left[\int \log\left(\frac{p(\tilde{y}|\theta_k, M_k)}{a_k(\tilde{y})}\right) p(\tilde{y}|\theta_k, M_k) d\tilde{y}\right] p(\theta_k|D, M_k) d\theta_k \tag{47}$$

is seen to be the Bayesian posterior predictive distribution, $\hat{a}_k(\tilde{y}) = p(\tilde{y}|D, M_k)$. Given the M_k-optimal prediction $\hat{a}_k(\tilde{y})$ the expected utility for the model M_k

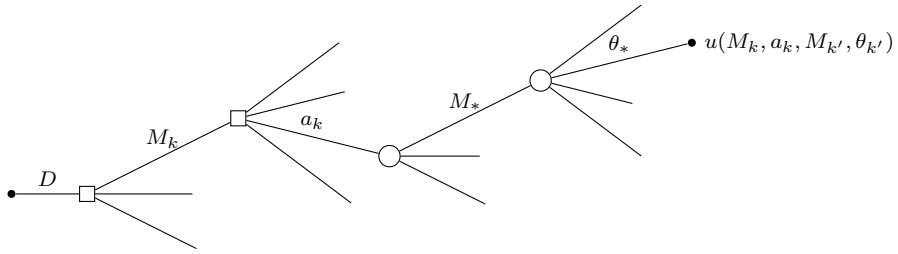

FIG 7. *Predictive model selection with the negative KL-divergence utility under the unknown sampling model* $p(y|\theta_{k'}, M_{k'})$.

is obtained as the posterior-averaged KL-divergences between the sampling model with unknown parameters and the posterior predictive distribution of the model M_k,

$$\bar{u}_*(M_k, \hat{a}_k) = \\ -\sum_{k'=1}^{K} \left[\int d_{\mathrm{KL}} \left\{ p(\tilde{y}|\theta_{k'}, M_{k'}), p(\tilde{y}|D, M_k) \right\} p(\theta_{k'}|D, M_{k'}) d\theta_{k'} \right] p(M_{k'}|D). \quad (48)$$

After a straightforward manipulation of the resulting expected utility and dropping the terms constant with respect to model M_k one can show that the model maximizing the utility in Eq. (48) is the same as the model maximizing the utility in Eq. (26) when the actual belief model M_* is the BMA predictive distribution $p_{\mathrm{BMA}}(\tilde{y}|D)$, Eq. (4). A generalization of this approach based on α-divergences has been proposed by Trottini and Spezzaferri (2002).

3.4.2. Model selection with the Gibbs utility

An often-used utility in Bayesian model selection is the negative KL-divergence from the actual belief model $p(\tilde{y}|D, M_*)$ to the prediction, Eq. (11). In the same spirit as in projection predictive methods in Section 3.3.2, the predictions may be restricted to the parametric form $p(\tilde{y}|\theta_k, M_k)$, so that the utility function can be written directly as

$$u(M_k, \theta_k) = -\int \log \left(\frac{p(\tilde{y}|D, M_*)}{p(\tilde{y}|\theta_k, M_k)} \right) p(\tilde{y}|D, M_*) d\tilde{y}. \quad (49)$$

Instead of selecting a point estimate for θ, that is, selecting the optimal prediction $p(\tilde{y}|\hat{\theta}_k, M_k)$, one may be interested in the average predictive performance of the model M_k, as illustrated in Fig. 8. The average predictive performance is defined as the expected utility, with the expectation over the unknown parameter θ_k taken with respect to the model-conditional posterior distribution,

$$\bar{u}(M_k) = -\int \left[\int \log \left(\frac{p(\tilde{y}|D, M_*)}{p(\tilde{y}|\theta_k, M_k)} \right) p(\tilde{y}|D, M_*) d\tilde{y} \right] p(\theta_k|D, M_k) d\theta_k. \quad (50)$$

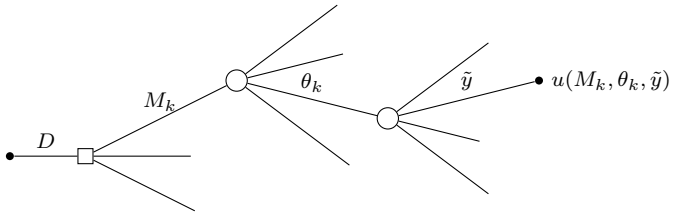

FIG 8. *Model selection with Gibbs utility: the predictive performance of the model M_k is evaluated as the posterior expected predictive performance of the parametric models $p(\tilde{y}|\theta_k, M_k)$.*

Dropping the terms constant with respect to model M_k results in the *Gibbs utility* (terminology follows Watanabe (2009), although Watanabe uses loss functions instead utilities)

$$\bar{u}_*^G(M_k) = \int \left[\int \log p(\tilde{y}|\theta_k, M_k) p(\tilde{y}|D, M_*) d\tilde{y} \right] p(\theta_k|D, M_k) d\theta_k \tag{51}$$

$$= \int \left[\int \log p(\tilde{y}|\theta_k, M_k) p(\theta_k|D, M_k) d\theta_k \right] p(\tilde{y}|D, M_*) d\tilde{y}. \tag{52}$$

The Gibbs utility measures the predictive performance of the candidate model M_k as the average predictive performance of the parametric probability distributions indexed by M_k. The significant difference to the model selection approaches based on the logarithmic utility function described in Section 3.3 is the lack of selection of unique prediction a_k: the Gibbs utility cannot be used to select a model and to identify a unique prediction action for the future data given the selected model. In other words, Gibbs utility can be employed to select a model based on the average predictive performance, but maximization of the expected utility will not tell the user how to actually predict the future observations.

Mathematically the difference in the Gibbs utility in Eq. (52) and the expected logarithmic predictive density in Eq. (26) is the order of logarithm and the inner integration. From Jensen's inequality it is evident that the expected logarithmic predictive density is lower bounded by the Gibbs utility.

The form of the Gibbs utility is often computationally simpler than the expected logarithmic predictive density in Eq. (26). For example, for observation models in the exponential family it is easier to take expectations of log likelihood in Eq. (52), while the expectations of the logarithm of the predictive distribution in Eq. (26) can be mathematically more involved.

3.4.3. Zero-one utility on the model space

If the unknown state of the world is assumed to be a model specification $M_{k'}$ belonging to the exhaustive set of models $\{M_{k'}\}_{k'=1}^K$, and the beliefs about the unknown model are described by the posterior distribution $p(M_{k'}|D)$, then the

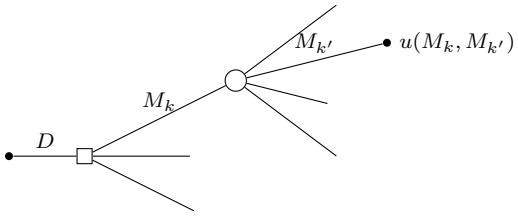

FIG 9. *Model selection with the zero-one utility: the optimal model choice is the model M_k with the highest posterior probability $P(M_k|D)$.*

zero-one utility function defined on the model space

$$u(M_k, M_{k'}) = \begin{cases} 1 & \text{if } M_{k'} = M_k \\ 0 & \text{if } M_{k'} \neq M_k \end{cases} \tag{53}$$

describes the utility of selecting the model M_k when the unknown model structure is $M_{k'}$. When $\{M_{k'}\}_{k'=1}^K = \{M_k\}_{k=1}^K$ the posterior distribution can be written in terms of the candidate models by

$$p(M_{k'}|D) = \begin{cases} p(M_k|D) & \text{if } M_{k'} = M_k \\ 0 & \text{if } M_{k'} \neq M_k, \end{cases} \tag{54}$$

because the unknown model $M_{k'}$ is believed to be among the set of candidate models. The expected utility for selecting the model M_k is seen to be

$$\bar{u}(M_k) = \int u(M_k, M_{k'}) p(M_{k'}|D) dM_{k'} = p(M_k|D). \tag{55}$$

The optimal model selection under the zero-one utility function is to choose the model with highest posterior probability. This is equivalent to Bayes factor (Kass and Raftery, 1995) model selection in case of equal prior probabilities for models (see Section 5.6). It is noteworthy that with the zero-one utility function model selection is not based on the predictive properties of the candidate models, nor does solving the decision problem result in a unique optimal prediction.

3.5. Other closely related concepts

Many model selection formulations in the literature, an obvious example being the non-Bayesian approaches, do not fit into the presented Bayesian predictive framework. Nevertheless, these methods may still be based on the predictive properties of the models, and they may be closely related to the Bayesian methods described in this survey.

3.5.1. Plug-in predictive distribution and deviance

A plug-in[2] predictive distribution

$$p(\tilde{y}|\hat{\theta}_k(D), D, M_k) \approx \int p(\tilde{y}|\theta_k, M_k)p(\theta_k|D, M_k)d\theta_k = p(\tilde{y}|D, M_k) \tag{56}$$

is a common approximation to the fully Bayesian predictive distribution, which results from using a point estimate for all or some of the parameters instead of integrating over the full posterior distribution. Replacing the full predictive distribution of the candidate model by the plug-in predictive distribution results in a plug-in utility. For example, in case of the logarithmic utility function the corresponding plug-in utility is given by

$$\bar{u}_*(M_k, \hat{a}_k) \approx \bar{u}_*^P(M_k, \hat{\theta}_k(D)) = \int \log p(\tilde{y}|\hat{\theta}_k(D), D, M_k)p(\tilde{y}|D, M_*)d\tilde{y}. \tag{57}$$

Plug-in quantities are often encountered especially with information criteria presented in Section 5.5.

In contrast to the predictive point estimates presented in Sections 3.3.1 and 3.4.1, the plug-in estimator $\hat{\theta}_k$ does not need to follow from decision-theoretic optimality. In fact, the plug-in point estimate can be chosen based on a different utility function (for example, posterior mode corresponding to zero-one utility, or posterior mean corresponding to squared error) than the one that is actually used to evaluate the expected utility (for example, logarithm of the predictive distribution).

In Bayesian applications, data-dependent plug-in estimates are typically used for hyperparameters, while the lower-level parameters are integrated over. Good examples of such approaches are empirical Bayes (EB) (Carlin and Louis, 1996), type-II maximum likelihood (ML-II) (Berger, 1985) and evidence framework (MacKay, 1992), which differ mainly in how the hyperparameters are estimated.

With complex models, an optimization-based plug-in approach can be computationally relatively simple while a full integration over all the unknown parameters may be infeasible. In practical modeling problems, ignoring the uncertainty related to the estimated parameters can be informally justified if the predictions by the model are not significantly affected.

A specific example of a loss function based on plug-in logarithmic score function is the *deviance* function

$$\mathrm{Dev}(D; \hat{\theta}) = 2\log C(D) - 2\log p(D|\hat{\theta}_k(D), M_k), \tag{58}$$

where the function $C(D)$ does not depend on the candidate model. For example, in the context of generalized linear models the function $C(D)$ is the maximum achievable log likelihood $\log p(D|\tilde{\theta})$ of the full model, representing the best attainable data fit (McCullagh and Nelder, 1989). However, candidate models are

[2]The point estimate $\hat{\theta}(D)$ is sometimes called *plug-in* estimate, as it is plugged into the model, sometimes without a solid theoretical justification.

often compared by the difference of the respective deviances, so that the constant terms in the deviance function cancel out. Deviance is used mostly in frequentist literature as it is closely connected to the likelihood ratio statistic. For consistency, in this review we replace the deviance function by the plug-in log-score.

3.5.2. On frequentist formulation of model assessment and selection

Reviewing the vast frequentist and other non-Bayesian literature on model assessment and selection is outside the scope of this survey. Some of the results from the related theory can also be utilized in the Bayesian framework. As necessary, in the following sections the most interesting and relevant frequentist results are treated with references for further reading. Mostly we follow the framework introduced, for example, by Akaike (1974) and Burnham and Anderson (2002).

In frequentist statistics, the predictions of a model M_k are usually based on a point estimate such as the maximum likelihood estimate (MLE) $\hat{\theta}(D)$. The quality of the predictions can be assessed by evaluating the expected utility

$$\bar{u}_t(M_k, \hat{\theta}_k) = \mathbb{E}_{\tilde{y}, D_n}\left[u(M_k, \hat{\theta}_k(D_n), \tilde{y})\right], \tag{59}$$

where the expectation is taken not only over the future unknown observation $\tilde{y}$, but also over all the possible n-sized sets of observations D_n with respect to an unknown true distribution p_t. In this survey, we use a common notation for all approaches, although the terminology in the frequentist statistical decision theory literature is typically different from the Bayesian literature.

Under certain regularity conditions, the sampling distribution of the estimator $\hat{\theta}_k$ and the Bayesian posterior distribution of θ_k approach asymptotically the same Gaussian distribution, and the MLE $\hat{\theta}_k$ and mode of the posterior distribution converge to the same value (see, e.g., Gelman et al., 1995, and references therein). However, in the non-asymptotic case replacing the sampling distribution of the estimator $\hat{\theta}_k$ with the posterior distribution of θ_k does not produce equivalent results, as illustrated in the following example.

Example: Kullback-Leibler divergence A common formulation for many non-Bayesian predictive model selection approaches is based on the Kullback-Leibler divergence between an unknown true distribution of the observations, $p_t(y)$ and the candidate model $p(y|\theta_k, M_k)$ (Akaike, 1974; Burnham and Anderson, 2002). The parameters of the model are estimated from the observed data D. The expected prediction error is the expectation of the KL divergence loss over all possible training data sets following the same unknown distribution $p_t(y)$,

$$l_t(M_k, \hat{\theta_k}) = \mathbb{E}_{D_n}\left[\int \log\left(\frac{p_t(\tilde{y})}{p(\tilde{y}|\hat{\theta}_k(D_n), M_k)}\right) p_t(\tilde{y}) d\tilde{y}\right]. \tag{60}$$

The expectation over the training sets can be written in terms of the sampling distribution of the estimator $g(\hat{\theta}_k)$. Ignoring terms constant with respect to the model M_k, the expected prediction error can be written equivalently as the expected prediction utility

$$\begin{aligned}\bar{u}_t(M_k, \hat{\theta}_k) &= \mathbb{E}_{\hat{\theta}_k}\, \mathbb{E}_{\tilde{y}} \left[\log p(\tilde{y}|\hat{\theta}_k, M_k)\right] \\ &= \int \left[\int \log p(\tilde{y}|\hat{\theta}_k, M_k) p_t(\tilde{y}) d\tilde{y}\right] g(\hat{\theta}_k) d\hat{\theta}_k. \end{aligned} \tag{61}$$

The similarity of Eq. (61) to the Gibbs utility in Eq. (52) may be one reason why the Gibbs utility has been used so extensively in Bayesian model selection.

In practical model selection approaches, both the expectations in Eqs. (60)–(61) must be approximated, as they are taken over an unknown distribution $p_t(y)$. Examples of the resulting information criteria are discussed in Section 5.5.

4. Predictive model comparison in practice

Predictive model assessment and selection, as presented in Section 3, are rather straightforward decision problems when the actual belief model $p(\tilde{y}|D, M_*)$ is readily available and utility-related computations over $p(\tilde{y}|D, M_*)$ can be performed. How the actual belief model is defined depends on assumptions underlying the prediction task at hand. A common categorization presented by Bernardo and Smith (1994) into $\mathcal{M}$-closed, $\mathcal{M}$-completed and $\mathcal{M}$-open views is useful in clarifying the strength of statements we are willing to make about $p(\tilde{y}|D, M_*)$.

We use the term *model comparison*[3] in practical model selection context where we need to take into account external information that is difficult to formulate either in terms of probabilistic models or within decision theoretic framework. In particular, the specific definition of $p(\tilde{y}|D, M_*)$, limited amounts of data, the nature of possible input variables and the computational approaches that are employed all affect how the expected utility is estimated. While the goal is still to identify the model with the best predictive performance defined in terms of expected utility, the properties of the expected utility estimate need to be considered when constructing a practical utility-based approach for comparing different models. The outline of the section is illustrated in Fig. 10.

4.1. $\mathcal{M}$-closed, $\mathcal{M}$-completed and $\mathcal{M}$-open views

In the *$\mathcal{M}$-closed* view it is possible to either enumerate all possible model structures $\{M_k\}_{k=1}^K$ and place a prior distribution $p(M_k)$ over them or specify a non-parametric model with a prior distribution on a suitable function space. This is equivalent to stating a belief that one of the candidate models is the

[3] The authors are aware that the term "model comparison" is often used interchangeably with model selection.

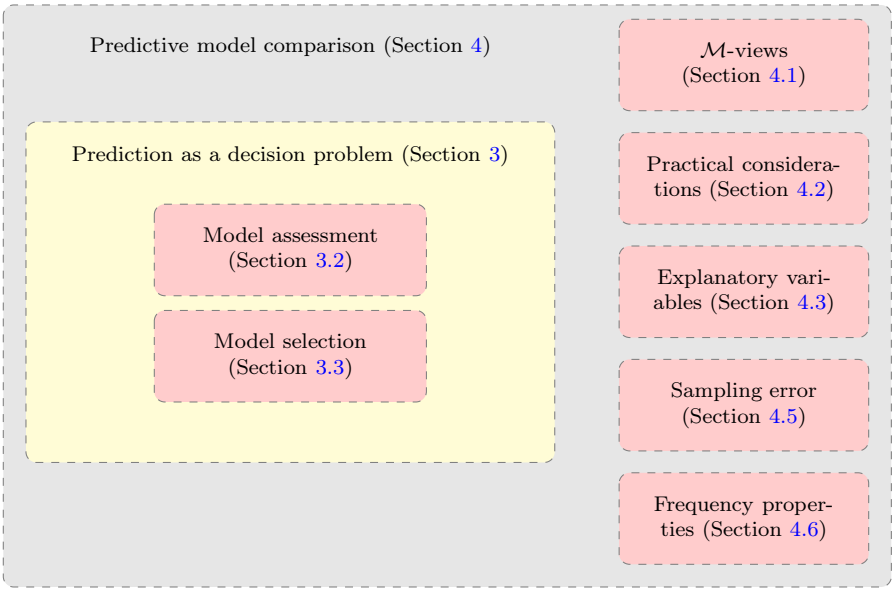

FIG 10. *Content for Section 4 and its relation to Section 3.*

"true model generating the data", under uncertainty about which of the candidate model is the said "true model". If the number of alternative models is countable, the actual belief model of the future observations is constructed as the Bayesian model averaging predictive distribution $p(\tilde{y}|D, M_*) = p_{\mathrm{BMA}}(\tilde{y}|D)$, Eq. (4). Literally taken, the $\mathcal{M}$-closed view is appropriate only when it is known for certain that the true data generating real world mechanism is among a finite set of models. Situations where this applies are not often encountered; one example could be a computer simulation where the observations are actually generated by one of the candidate models. Although it is difficult to find situations in which the strict interpretation of the $\mathcal{M}$-closed view holds, Bayesian model averaging has been shown to have good predictive performance (Raftery and Zheng, 2003). Thus, often it is not too unreasonable to proceed with the $\mathcal{M}$-closed view as if one believed in it, that is, by placing prior weights on a limited set of well-defined alternative models.

In the $\mathcal{M}$-*completed* view one forms a rich enough model M_* whose predictions are considered to best reflect the uncertainty in the prediction task. In typical modeling problems, it is impossible to come up with an exhaustive list of possible candidate models, with one of them being guaranteed to be the true data generating model, and place an explicit model prior. Instead of the BMA predictive distribution, the predictive distribution of the actual belief model $p(\tilde{y}|D, M_*)$ is considered to be the best available description of the uncertainty of future data.

In the $\mathcal{M}$-*open* view, the aim is to avoid constructing explicitly the actual belief model, as there is a strong conviction under the current background infor-

mation that any such model would not reflect well the properties of future data. In this case, it may be more appropriate to assess the predictive performance of the candidate models under minimal modeling assumptions rather than being confident about the realism of one's current predictive model. It is possible to resort to thinking that while it is not possible to correctly specify the distribution of the future data, it is still possible to obtain pseudo Monte Carlo samples from it (Bernardo and Smith, 1994). Such thinking leads to sample re-use methods such as cross-validation. The decision theoretic formulation in terms of model parameters, discussed in Sections 3.4 and 3.4.1, is not suitable for describing the $\mathcal{M}$-open case.

The categorization into $\mathcal{M}$-closed, -completed or -open views should not be understood in an overly strict sense, as there are approaches combining properties from different categories as well as ones that cannot be classified in the above sense.

4.2. Expected utility estimation in practice

$\mathcal{M}$-closed and $\mathcal{M}$-completed views both lead to defining the expected utility for model M_k with the optimal optimal prediction action $\hat{a}_k$ directly as Eq. (20),

$$\bar{u}_*(M_k|D) = \int u(M_k, \hat{a}_k, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}, \tag{62}$$

where the uncertainty related to future data is described by the actual belief model $p(\tilde{y}|D, M_*)$. Much of the diversity found in predictive model assessment and selection methods comes from the differences in the specification of $p(\tilde{y}|D, M_*)$, whose implications are further explored in Section 5. Although calculating the integral in Eq. (62) in practice may require an approximative approach, such as numerical integration or a Monte Carlo solution, the expected utility is obtained in a rather straightforward fashion. Also, the estimate in Eq. (62) is highly dependent on the quality of the actual belief model M_*; misspecified beliefs about the future observations may be lead to poor expected utility estimates.

$\mathcal{M}$-open view corresponds to avoiding the explicit specification of $p(\tilde{y}|D, M_*)$ by re-using observations D as a proxy for the predictive distribution of the actual belief model. If samples $\{\tilde{y}_j\}_{j=1}^{\tilde{n}}$ independent of D can be obtained, for example, as a separate test data set, the expected utility $\bar{u}_*(M_k|D)$ could be approximated as

$$\bar{u}_{\text{test}}(M_k|D) = \frac{1}{\tilde{n}} \sum_{j=1}^{\tilde{n}} u(M_k, \hat{a}_k, \tilde{y}_j). \tag{63}$$

Assessing the utility in this way is referred to as external validation (Gelman et al., 2003), and also as the *test utility* (or when loss functions are used, test

error) especially in the machine learning literature. In absence of additional test data set, the naïve approach would be to use exact replicates of data, $\dot{y}_i = y_i$. The dot-notation is used to emphasize that even though $\dot{y}_i$ has the same value as y_i, it represents a realization of a different random quantity. The resulting estimate

$$\bar{u}_{\text{train}}(M_k|D) = \frac{1}{n}\sum_{i=1}^{n} u(M_k, \hat{a}_k, \dot{y}_i), \tag{64}$$

is often referred to as the *training utility* (or when loss functions are used, training loss or training error). Training utility is a biased estimate of test utility, since replicates $\{\dot{y}_i\}_{i=1}^n$ are not independent from observations $D = \{y_i\}_{i=1}^n$.

As the naïve example shows, the actual way of implementing the sample re-use has a significant effect on the expected utility estimate. One way of improving the quality of a training utility based estimate is to introduce a sample re-use strategy aimed at reducing the effect of dependence between $\{\dot{y}_i\}_{i=1}^n$ and D. This leads to methods such as cross-validation, where the observations D are divided in various ways to get independent proxies for D and $\tilde{y}$. For example, in leave-one-out cross-validation (LOO-CV) with logarithmic utility function the expected utility is estimated by the *LOO-CV utility*

$$\bar{u}_{\text{LOO}}(M_k|D) = \frac{1}{n}\sum_{i=1}^{n} u(M_k, \hat{a}_k, y_i|D_{(\setminus i)}), \tag{65}$$

where each $D_{(\setminus i)}$ (observations not including y_i) serves as a proxy for D and y_i as a proxy for $\tilde{y}$ in turn. However, each $D_{(\setminus i)}$ contains fewer observations than D, which makes the LOO-CV estimate in Eq. (65) a biased estimate of the expected utility. Also, because each $D_{(\setminus i)}$ is different from D, additional variance is introduced in the estimate. The variations of cross-validation are discussed in more detail in Section 5.1.3.

The expected utility estimate can also be improved by estimating and correcting for the bias in the training utility, which leads to information criteria type approaches. Formally, the effect of bias is reduced by introducing a bias correction term to the training utility,

$$\bar{u}_{\text{IC}}(M_k|D) = \frac{1}{n}\sum_{i=1}^{n} u(M_k, \hat{a}_k, \dot{y}_i) + \text{bias correction}. \tag{66}$$

Information criteria are discussed in Section 5.5.

4.3. Expected utility estimation when explanatory variables x are included

Throughout Sections 3 and 4 explanatory variables x were left out from equations in order to keep the notation simpler. Conditional modeling of $y|x$ depends

on the assumptions related to the explanatory variables x. The decision problem as well as the practical expected utility estimation need to be defined in a slightly different way, depending on whether the future explanatory variable $\tilde{x}$ is assumed to be random, unknown, fixed, deterministic or controlled quantity, or even a mixture of the said quantities.

x and $\tilde{x}$ are random refers to a case where observed x and not yet observed $\tilde{x}$ are assumed to be exchangeable random quantities. This case is typical in observational studies where the explanatory variables are not controlled. Commonly the distribution of x is assumed to be stationary in time, but covariate shift can be taken into account, for example, by weighting methods (e.g. Shimodaira, 2000; Sugiyama, and Müller, 2005; Sugiyama, Krauledat and Müller, 2007).

Adopting an $\mathcal{M}$-close or $\mathcal{M}$-completed view requires placing a prior $p(x, \tilde{x})$ for the explanatory variables, so that the expected utility is defined as

$$\bar{u}_*(M_k|D) = \int u(M_k, \hat{a}_k, \tilde{y}, \tilde{x}) p(\tilde{y}, \tilde{x}|D, M_*) d\tilde{y} d\tilde{x}, \tag{67}$$

where the explanatory variables are now explicitly shown in the utility function.

In practice, modeling the distribution of x is usually more difficult than conditional modeling of $y|x$, making it common to settle for an $\mathcal{M}$-open view approach. If an $\mathcal{M}$-open approach is used for both $y|x$ and x, the expected utility can be estimated for example by training utility, Eq. (64), as

$$\bar{u}_*(M_k|D) \approx \frac{1}{n} \sum_{i=1}^{n} u(M_k, \hat{a}_k, \dot{y}_i, \dot{x}_i). \tag{68}$$

It is also possible take an $\mathcal{M}$-closed or $\mathcal{M}$-completed view with respect to $y|x$ and an $\mathcal{M}$-open view with respect to x. The expected utility estimate is then given by

$$\bar{u}_*(M_k|D) \approx \frac{1}{n} \sum_{i=1}^{n} \left[\int u(M_k, \hat{a}_k, \tilde{y}, \dot{x}_i) p(\tilde{y}|\dot{x}_i, D, M_*) d\tilde{y} \right]. \tag{69}$$

With random x and $\tilde{x}$ it is very likely that the values of $\tilde{x}$ are different from x. In such a case it may be interesting to evaluate the *out-of-sample performance*, that is, the expected utility at locations $\tilde{x}$ which are not necessarily among the observations $(x_1, \ldots, x_n)$. In the full $\mathcal{M}$-closed or $\mathcal{M}$-completed approaches the out-of-sample estimate comes naturally as the expectation is taken also over $p(\tilde{x}|D, M_*)$. In $\mathcal{M}$-open view based sample re-use approaches, hold-out and cross-validation predictive methods (Section 5.1) can be used for estimating the out-of-sample performance.

Typically $\tilde{x}$ is present in the utility function only as a condition for the optimal prediction $\hat{a}_k$, such as in $p(\tilde{y}|\tilde{x}, D, M_k)$ for the logarithmic utility or in $\mathbb{E}[\tilde{y}|\tilde{x}, D, M_k]$ for the squared error. Thus $\tilde{x}$ can be omitted from the utilities in subsequent sections to avoid clutter in notation.

x and $\tilde{x}$ are fixed when $(\tilde{x}_1, \ldots, \tilde{x}_n)$ equal to $(x_1, \ldots, x_n)$. That is, x and $\tilde{x}$ are known constants, and only the conditional part $y|x$ has any uncertainty. Typical examples of such a case are found in spatial epidemiology where, for example, locations x_i of counties do not change, and future observations are of the form $(x_i, \tilde{y}_i)$.

As future x are known, it would be logical to compute the expected utility using simultaneous prediction, but as single prediction is typically simpler to obtain, it is a more often used and useful proxy for the more complicated estimate. Taking an $\mathcal{M}$-closed or $\mathcal{M}$-completed view with respect to $y|x$ and an $\mathcal{M}$-open view with respect to x leads to estimating the expected utility as

$$\bar{u}_*(M_k|D) = \frac{1}{n}\sum_{i=1}^{n}\left[\int u(M_k, \hat{a}_k, \tilde{y})p(\tilde{y}_i|x_i, D, M_*)d\tilde{y}_i\right]. \tag{70}$$

Here the sum over fixed values x_i collects the utilities following from individual single predictions to a single statistic, while in Eq. (69) for unknown x the sum results from the Monte Carlo expectation based on the pseudo samples $\dot{x}_i$ from $p(\tilde{x})$.

In fixed x case, there is no need to estimate the out-of-sample performance at new x values, but depending on the prediction task we may consider prediction of $\tilde{y}_i$ given either all of D which includes y_i as above, or with $D_{(\setminus i)}$ with y_i removed. For example, in spatial data analysis these prediction tasks correspond to predicting for ith area given the observations in all areas either by including, or excluding, the ith area itself. These predictions can be quite different if observations with different x_i are independent or weakly dependent, that is, $p(\tilde{y}_i|x_i, D, M)$ and $p(\tilde{y}_i|x_i, D_{(\setminus i)}, M)$ are not similar.

In $\mathcal{M}$-open sample re-use methods, it is not possible to separate samples (x_i, y_i) and to get an independent proxy for $\tilde{y}$. Thus, for example, in leave-one-out cross-validation only the $p(\tilde{y}_i|x_i, D_{(\setminus i)}, M)$ prediction scenario is possible. Moreover, in a fixed x case the uncertainty is only with respect to $y|x$, yet the sample re-use methods treat the fixed explanatory variables $\tilde{x}$ as random. This error does not affect the estimated expected utility, but affects the assessment of the sampling error associated with the estimate.

$\tilde{x}$ is deterministic when $(\tilde{x}_1, \ldots, \tilde{x}_{\tilde{n}})$ are deterministic quantities, but different from $(x_1, \ldots, x_n)$. A typical example is a time series prediction where the future time points are known, but not yet observed.

Similarly to the case of fixed explanatory variables, the $\tilde{x}$ are constant, and the uncertain part of the model is the conditional model $p(\tilde{y}|\tilde{x}, D, M_*)$. However, with deterministic explanatory variables there is a need to estimate the out-of-sample performance with a structure specific to $\tilde{x}$. In the full $\mathcal{M}$-open case cross-validation variants (see Section 5.1.3) can be used to take into account the out-of-sample performance. In the $\mathcal{M}$-closed and $\mathcal{M}$-completed approaches it is trivial to estimate the expected utility given the known deterministic values

$(\tilde{x}_1, \ldots, \tilde{x}_{\tilde{n}})$ as

$$\bar{u}_*(M_k|D) = \frac{1}{n}\sum_{i=1}^{\tilde{n}} \left[\int u(M_k, \hat{a}_k, \tilde{y}) p(\tilde{y}|\tilde{x}_i, D, M_*) d\tilde{y} \right]. \tag{71}$$

x is controlled or $\tilde{x}$ is partially controlled when the explanatory variables are determined by experimental design or controlled otherwise, and the outcome variables $y|x$ are considered to be observational quantities. In this case, knowing x does not give information about either random or controlled $\tilde{x}$. A typical case is an industrial or medical design of experiment.

If $\tilde{x}$ are random, additional information can be used to form $p(\tilde{x})$. For example, a large number of x's may have been observed, but due to a high measurement cost, the associated outcome variables y have been observed only for a smaller number of x's selected by design of experiment. In absence of such additional information needed to form $p(\tilde{x})$, or when $\tilde{x}$ are also controlled, one choice is to consider how well the model performs with a fixed x.

If x's are random observations, it is possible that $\tilde{x}$ are at least partially controlled. For example, in a production process there may be randomness due to weather, amount and quality of ingredients, varying temperature, and so on. After observing the behaviour of the process it may be desired to control some of the previously uncontrolled quantities which have an effect on the process. Although the future values $\tilde{x}$ will be restricted, the performance of the model needs to be assessed over a wider range of possible x-values in order to make a good control decision. With respect to that decision problem the $\tilde{x}$ can be considered fixed with no uncertainty.

4.4. Model comparison

While selecting a model with the best predictive performance in terms of expected utility is a simple concept in principle, the computational and approximative steps taken in estimating the expected utility lead us to consider the sampling error (Section 4.5) and the frequency properties (Section 4.6) of the expected utility estimates. In practice the expected utilities of the candidate models are often compared in a pairwise manner, using subjective assessment of the significance of the difference in the respective expected utilities.

$\mathcal{M}$-closed or $\mathcal{M}$-completed view Given the belief model $p(\tilde{y}|D, M_*)$ the difference between the expected utilities of any two candidate models, $\bar{u}_*(M_k|D)$ and $\bar{u}_*(M_j|D)$, equals the expectation of the difference,

$$\begin{aligned}\bar{u}_*(M_j|D) - \bar{u}_*(M_k|D) &= \int u(M_j, \hat{a}_j, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y} - \int u(M_k, \hat{a}_k, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y} \qquad (72)\\ &= \int \left[u(M_j, \hat{a}_j, \tilde{y}) - u(M_k, \hat{a}_k, \tilde{y}) \right] p(\tilde{y}|D, M_*) d\tilde{y}. \qquad (73)\end{aligned}$$

The latter form may be preferable if the expectation over $p(\tilde{y}|D, M_*)$ needs to be approximated.

With a strictly proper scoring rule it is evident that under the $\mathcal{M}$-closed and $\mathcal{M}$-completed views the actual belief model M_* will always be preferred. For example, given $M_j = M_*$ and the logarithmic utility function the expected difference of the utilities is the KL-divergence from M_* to M_k, a strictly positive quantity for all $M_k \neq M_*$. However, the aim in model comparison is often identification of a simpler model that is sufficiently close to M_* in terms of predictive performance, even though the said aim is not represented in the utility function and therefore not properly included into the decision theoretic framework. The quantification of this practical difference in the expected utilities of the models compared is referred to as *calibration* of the model comparison method. Calibration is discussed in Section 5.7. Again, it is worth stressing that the results from model comparison in the $\mathcal{M}$-closed and $\mathcal{M}$-completed approaches depend on the quality of the actual belief model M_*.

$\mathcal{M}$-open view In the $\mathcal{M}$-open approaches issues related to the significant difference in the expected utilities of the models compared are the same as in the $\mathcal{M}$-closed or $\mathcal{M}$-completed case. However, because the actual belief model is represented by a finite proxy sample it is possible that there is considerable uncertainty whether one model is better than another model. It is possible to compute the corresponding uncertainty measure by estimating the sampling error of the expected utility differences (see Section 4.5). These uncertainty measures can also be used to guide in choosing, for example, the simplest model which is not significantly worse than some larger, more complex model.

Selection induced bias A model selection procedure based on maximizing the expected utility estimated with re-used samples suffers from a phenomenon called *selection induced bias*. A model selection procedure using a criterion conditional on the training data, such as the estimated expected utility, fits to the observed data. Even if the procedure is based on an approach giving unbiased expected utility estimates for any particular *model*, the data-fitted model selection procedure causes the expected utility estimate of *the selected model* to be biased (see, for example, Stone, 1974; Rencher and Pun, 1980; Reunanen, 2003; Vehtari and Lampinen, 2004; Shen, Huang and Ye, 2004; Varma and Simon, 2006; Cawley and Talbot, 2010).

If the number of candidate models is very large (for example, the number of models grows exponentially as the number of observations n grows, or the number of covariates $p \gg \ln(n)$ in covariate selection) a model selection procedure can strongly overfit to the data. For example, Birgé and Massart (2007) demonstrate this for a penalized least-squares criterion and Arlot and Celisse (2010) provide references in relation to (non-Bayesian) cross-validation methods. It is possible to estimate the selection induced bias and obtain unbiased estimates (see section 5.1.3). This does not, however, prevent the model selection procedure from possibly overfitting to the observations and consequently selecting models with suboptimal predictive performance.

4.5. Sampling error of the expected utility estimate

If the expected utility is calculated as a Monte Carlo estimate, the associated Monte Carlo error depends on the number of samples from $p(\tilde{y}|D, M_*)$. In the $\mathcal{M}$-open sample re-use approaches the size of the proxy sample is fixed and n is typically relatively small which may result in a substantial sampling error, whereas the Monte Carlo error resulting from a Monte Carlo integration over the posterior distribution of parameters can be reduced by increasing the number of samples. As an example of the latter case, Zhu and Carlin (2000) estimate the Monte Carlo variance of Deviance Information Criterion (DIC) arising from Monte Carlo sampling of the parameter posterior.

A typical problem in Monte Carlo methods is that the variance of the Monte Carlo error can be substantial when the Monte Carlo integral is calculated from samples drawn from a distribution with thick tails. This can happen when the proxy sample contains rare observations: as only a restricted subset of all possible values of the future observation $\tilde{y}$ can be represented by the proxy sample, the expected utility estimate may be sensitive with respect to the actual observed data. Choice of a particular utility function can also amplify the effect of rare observations in the proxy sample, as demonstrated in a non-Bayesian cross-validation setting by Leung (2005).

Straightforward sample re-use approaches do not include any assumptions on the distribution of the future observations. Expected utility estimates with a smaller variance for the Monte Carlo error could be obtained by Bayesian Monte Carlo (Rasmussen and Ghahramani, 2003), where additional smoothness assumptions are made. The improved estimates come with the cost of additional computation as well as statements regarding the properties of the distribution $p(\tilde{y}|D, M_*)$.

Variance of the sampling error The estimate of the expected utility can be written as the sample average

$$\bar{u}_*(M_k|D) \approx \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} u(M_k, \tilde{y}_i|D) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} u_i, \tag{74}$$

where u_i is the utility for a sample $\tilde{y}_i$. The variance of the sampling error is

$$\mathrm{Var}_{\mathrm{MC}}(u) = \frac{\widehat{\mathrm{Var}}(u)}{\tilde{n}}, \tag{75}$$

where

$$\widehat{\mathrm{Var}}(u) \approx \frac{1}{\tilde{n}-1} \sum_{i=1}^{\tilde{n}} \left(\bar{u}_*(M_k|D) - u_i\right)^2. \tag{76}$$

Eq. (75), and a more robust quantile-based variance estimate, were proposed for cross-validation in a non-Bayesian setting by Breiman et al. (1984, ch. 11).

Both are often adequate approximations even if the distribution of u_i's is not Gaussian.

A pairwise comparison for models M_k and $M_{k'}$ can be made by calculating the difference of the expected utilities as

$$\bar{u}_*(M_k|D) - \bar{u}_*(M_{k'}|D) \approx \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (u_i(M_k, \tilde{y}_i|D) - u_i(M_{k'}, \tilde{y}_i|D)), \tag{77}$$

as both $u(M_k, \tilde{y}_i|D)$ and $u(M_{k'}, \tilde{y}_i|D)$ depend on the same sample $\tilde{y}_i$. A variance estimate for the difference can be computed in a similar fashion. The expected difference and variance can be used to approximate the probability for the sign of the difference, which in the case of similar expected utility estimates can be used as an additional indicator for significance of the difference.

Bayesian bootstrap If the uncertainty related to $\bar{u}_*(M_k|D)$ cannot be described well with a Gaussian distribution, Vehtari and Lampinen (2002) proposed to use a non-parametric Bayesian bootstrap (BB) (Rubin, 1981) approach based on the Dirichlet distribution. Sampling from the Dirichlet distribution g gives BB samples from the distribution of the distribution of u and thus samples of any parameter of this distribution can be obtained. For example, with $\bar{u} = \mathbb{E}[u]$, for each BB sample b the mean of u is calculated as if $g_{i,b}$ were the probability that $u = u_i$; that is, $\bar{u}_b = \sum_{i=1}^{n} g_{i,b} u_i$. The distribution of the values of $\bar{u}_b; b = 1, \ldots, B$ is the BB distribution of the mean $\bar{u}$. For important properties of Bayesian bootstrap, see Lo (1987); Weng (1989); Mason and Newton (1992). The assumption that all possible distinct values of (x, y) have been observed is usually wrong, but with moderate n and not very thick tailed distributions, inference should not be very sensitive to this.

4.6. Frequency properties

In the following, we use the notation $\hat{u} \approx \bar{u}$ for the method-specific utility estimate to emphasize that the estimate for the expected utility $\hat{u}$ for model M_k is conditioned on the specification $p(\tilde{y}|D, M_*)$, the training data $D = \{y_i\}_{i=1}^{n}$ as well as on method-specific approximations.

Although in the Bayesian framework, inference and decisions are always conditioned on the fixed observed data, there are certain reasons for considering the frequency properties of the methods for estimating the expected utility. Frequency properties can be used to describe the long run behavior of the utility estimation method or a model selection procedure in repeated use, and should be taken into consideration when the estimated expected utility is used as a basis for model selection or comparison.

Although the repeated use of the model assessment and selection methods in practice means using the said methods for data sets in different modeling problems, the analysis of the frequency properties is usually simplified by considering a single distribution p_t from which both the observed data D_n of size n and the

unobserved future data $\tilde{y}$ are generated. The expected utility $\bar{u}$ can be regarded as a random quantity, as varying realizations of D_n cause both $p(\tilde{y}|D_n, M_*)$ and $p(\tilde{y}|D_n, M_k)$ to vary. Additionally, the method-specific expected utility estimate $\hat{u}$ may include approximation errors and stochastic variation.

The frequency properties can be considered in a finite sample case and asymptotically ($n \to \infty$). While asymptotic properties are often interesting, in practice the finite sample performance is more important in model assessment, selection and comparison, as realistic model selection problems involve limited amount of data. Moreover, frequency properties in finite samples typically have a larger impact on the comparison results. In the following, the most common frequency properties – bias, variance, efficiency and consistency – are considered.

4.6.1. Bias, variance and efficiency

Bias is the expected difference between a method-specific estimate $\hat{u}(M|D_n)$ and the expected utility $\bar{u}_t(M|D_n)$ given a known p_t,

$$\text{bias}[\hat{u}] = \mathbb{E}_{D_n}\left[\hat{u}(M|D_n) - \bar{u}_t(M|D_n)\right], \tag{78}$$

where the expectation is over the repeated sampling of the training data set D_n and the "true" utility

$$\bar{u}_t(M|D_n) = \int u(M, \hat{a}, \tilde{y}) p_t(\tilde{y}) d\tilde{y} \tag{79}$$

follows from evaluating the expected predictive performance over the known p_t (with all the optimal decisions $\hat{a}$ resulting from maximization with respect to $p(\tilde{y}|D_n, M_*)$).

The bias is a result of $p(\tilde{y}|D_n, M_*)$ differing from $p_t(\tilde{y})$ and of the method-specific approximation errors. In practice, $p_t(\tilde{y})$ is unknown. For some models and methods finite sample or asymptotic results can be computed analytically, but often simulations from a known constructed $p_t(\tilde{y})$ are used to obtain empirical results of the order of bias.

Variance

$$\text{var}\left[\hat{u}\right] = \mathbb{E}_{D_n}\left[\left(\hat{u}(M|D_n) - \mathbb{E}_{D_n}\left[\hat{u}(M|D_n)\right]\right)^2\right] \tag{80}$$

is a measure of the variability of the estimate $\hat{u}(M|D_n)$. For some models and methods finite sample or asymptotic results can be computed analytically, but as variance does not depend on $p_t(\tilde{y})$, it can also be estimated by different sample re-use techniques (e.g., Section 5.7.1).

Efficiency can be measured, for example, with the mean square error

$$\text{eff}[\hat{u}] = E_{D_n}\left[\left(\hat{u}(M|D_n) - E_D \bar{u}_t(M|D_n)\right)^2\right], \tag{81}$$

The mean square error will be small when both the bias and the variance are small.

In model assessment a large bias is undesirable as it leads to systematically wrong expected utility estimates. Large variance, in turn, is a signal of unreliable estimates. Unbiasedness in estimation is often desired, but efficiency can be considered to be more important in model assessment.

In model selection even a considerable bias may not be a problem. For example, a constant bias does not affect model selection in any fashion, so that a utility estimation method with a large constant bias but a small variance can be considered to be better than an unbiased one with a large variance. Furthermore, it has been observed that an estimate biased towards favoring simpler models can lead to better model selection performance than an unbiased estimate with a large variance (Cawley and Talbot, 2010). Such a complexity-dependent bias can be considered as implicit complexity prior information inherent in the model selection, even though no explicit complexity-related terms were introduced into the utility function.

The effect of the dependencies in sample re-use Training utility, a simple $\mathcal{M}$-open sample re-use approach, uses proxy samples for the future data that are not independent of the observed data, as discussed in section 4.2. The resulting optimistic bias can be reduced by alternative sample re-use techniques such as cross-validation (section 5.1) or using a bias correction as in information criteria (section 5.5). Many of the methods proposed in the literature, however, do not remove the optimistic bias completely.

The effect of the sample size to the expected utility estimate (learning curve) Some of the $\mathcal{M}$-open sample re-use based methods (for example, hold-out and cross-validation approaches) condition the predictions on only part of the data D to avoid the above-mentioned dependency bias. That is, only $m < n$ data points are used to compute the predictive distribution. In such cases it is useful to consider a theoretical construction called the *learning curve* in the machine learning literature (see, e.g. Rasmussen and Williams, 2006, Ch 7.). The learning curve is related to the *posterior convergence rate* in Bayesian literature. The learning curve describes the expected predictive performance of the model given the size of the training data $0 \leq m \leq n$, where the expectation is taken over all the possible training datasets generated from $p_t(\tilde{y})$. Usually the expected predictive performance increases monotonically as m increases. At any specific m, there is a systematic bias in the expected predictive performance when compared to results obtained with the full data consisting of n observations. Moreover, the steepness of the learning curve can vary for models of different complexity, so that a comparison between such models with $m < n$ training data points may not give the same ordering of the predictive performance as at the n training data points. This bias can be estimated and, for example, corrected cross-validation estimates have been presented (section 5.1.3).

The effect of model bias In $\mathcal{M}$-completed and $\mathcal{M}$-closed cases and information criteria, either M_* or M_k is used as model for the future data distribution. In practice, modeling will almost always introduce bias, which may however be

reasonably small if the model is good. The effect of the deterministic approximative integration when forming posterior predictive distributions can also be considered to be part of model bias.

The effect of the data realization The data realization D_n affects the predictive distribution, but it also affects the estimate of the future data distribution through $p(\tilde{y}|D_n, M_*)$ and $p(\tilde{y}|D_n, M_k)$ in the $\mathcal{M}$-completed and $\mathcal{M}$-closed approaches, and through re-used samples in the $\mathcal{M}$-open case. $\mathcal{M}$-open methods have generally higher variance than $\mathcal{M}$-completed and $\mathcal{M}$-closed methods due to their high variance in sample re-use (section 4.5). Different models also have different sensitivity to variations in the data, with more robust models producing less variable predictions. This effect can not be easily separated from the overall variation.

The effect of stochastic methods If the construction of the predictive distribution or expected utility estimate involves stochastic methods, such as Monte Carlo, this will produce additional variability. This variability can often be reduced to be small compared to other variations (see, e.g., Vehtari and Lampinen, 2002).

4.6.2. Consistency

Consistency is often mentioned as a desirable property in model selection. A consistent model selection procedure will select the "true model" among an exhaustive set of candidates as $n \to \infty$. The definition makes sense only when a true model is assumed (section 3.4.1). As an asymptotic property, consistency is not that important in a small sample case.

Consistency of model selection procedures is sometimes characterized by the so-called AIC-BIC dilemma (Yang, 2005). Roughly speaking, the dilemma means that the model selection methods are efficient either in the sense of prediction, or consistent in the sense of selecting the "true model". From the maximum expected utility predictive point of view, efficiency is the more important property.

Watanabe (2010a) shows that under certain regularity conditions for both regular and singular models, Bayesian cross-validation (Section 5.1.3) and widely applicable information criterion (WAIC, Section 5.5) are asymptotically equal and asymptotically converge to the true utility. Watanabe's results seem to be the only current fully Bayesian asymptotic results. However, frequentist non-asymptotic and asymptotic results (under different regularity conditions) on model assessment and selection can provide useful tools and hints at corresponding fully Bayesian results.

5. Methods for predictive model assessment and selection

In this Section, we review specific predictive model assessment and selection methods proposed in the literature within the unifying decision theoretic frame-

work presented in the previous Sections. The methods are divided into the following categories:

- $\mathcal{M}$-open treatment for both $\tilde{y}|\tilde{x}$ and $\tilde{x}$
- $\mathcal{M}$-closed/completed treatment for $\tilde{y}|\tilde{x}$ and $\mathcal{M}$-open for $\tilde{x}$
- $\mathcal{M}$-closed/completed treatment for both $\tilde{y}|\tilde{x}$ and $\tilde{x}$
- Information criteria
- Projection approaches
- Prior predictive or marginal likelihood

Methods that do not fit perfectly into these categories are presented in the category with the most similar approaches. If x and $\tilde{x}$ are not random quantities (see Section 4.3), the focus of modelling is solely on the conditional model for $\tilde{y}|\tilde{x}$. The following notation is used in describing sample re-use methods. An index set corresponding to all the observations is defined as $I = \{1, 2, \ldots, n\}$. An index set $I_s = \{i_1, \ldots, i_j\}$ defines a subset of j observations $D_{(I_s)} = \{(x_i, y_i)\}_{i \in I_s}$, and $D_{(\setminus I_s)} = \{(x_i, y_i)\}_{i \in I \setminus I_s}$ refers to a subset containing the remaining observations not indexed by I_s.

An abbreviation $\bar{u} \approx \ldots$ is used to emphasize that we are computing an estimate for some "ideal" quantity such as $\bar{u}_*(M)$, $\bar{u}_t(M)$ or $\bar{u}_{\text{test}}(M)$. Also, we aim to avoid clutter in notation by using the same symbol without spelling out all the details of the particular method (for example, a reference predictive approach with the BMA actual belief model for single prediction with deterministic x).

The logarithmic utility function Eq. (8) is used as the default utility function in the following examples. The logarithmic utility function is used especially often in model selection methods: on one hand it follows from the Bayesian decision theoretic considerations (Section 3.1) and on the other hand from the information-theoretic considerations based on the KL divergence (Section 3.5.2). The logarithmic forms presented in Fig. 11 are routinely encountered in the reviewed model selection methods (the terminology follows Watanabe (2009), while the notation is different).

5.1. $\mathcal{M}$-open treatment for both $\tilde{y}|\tilde{x}$ and $\tilde{x}$

As discussed in Sections 4.1 and 4.2, in the $\mathcal{M}$-open approaches the actual belief models $p(\tilde{y}|\tilde{x}, D, M_*)$ and $p(\tilde{x}|D, M_*)$ are not explicitly defined, but instead samples representing the distribution of the future observations $(\tilde{y}, \tilde{x})$ are assumed to be available. Although a finite number of samples does not necessarily represent continuous distributions well, expectations can be estimated often with sufficient accuracy, as discussed in more detail in Section 4.5. If the dimensionality of the data is large compared to the number of observations, it is possible that a finite number of samples do not cover well the area with a substantial density, thereby producing less reliable estimates (Jonathan, Krzanowski and McCarthy, 2000). In practice, in absence of an additional independent sample it is possible to re-use the observations in D, preferably in a way introducing a minimal sample re-use bias to the expected utility estimate. Sample re-use methods are robust in the sense that they are independent of the model assumptions

Bayes

generalization $$\bar{u}_t(M_k|D) = \int \log p(\tilde{y}|D, M_k) p_t(\tilde{y}) d\tilde{y}$$

reference $$\bar{u}_*(M_k|D) = \int \log p(\tilde{y}|D, M_k) p(\tilde{y}|D, M_*) d\tilde{y}$$

training $$\bar{u}_{\text{train}}(M_k|D) = \frac{1}{n}\sum_{i=1}^{n} \log p(\dot{y}_i|D, M_k)$$

Gibbs

generalization $$\bar{u}_t^G(M_k|D) = \int \left[\int \log p(\tilde{y}|\theta_k, M_k) p(\theta_k|D, M_k) d\theta_k \right] p_t(\tilde{y}) d\tilde{y}$$

reference $$\bar{u}_*^G(M_k|D) = \int \left[\int \log p(\tilde{y}|\theta_k, M_k) p(\theta_k|D, M_k) d\theta_k \right] p(\tilde{y}|D, M_*) d\tilde{y}$$

training $$\bar{u}_{\text{train}}^G(M_k|D) = \frac{1}{n}\sum_{i=1}^{n} \int \log p(\dot{y}_i|\theta_k, M_k) p(\theta_k|D, M_k) d\theta_k$$

Plug-in

generalization $$\bar{u}_t^P(M_k|D) = \int \log p(\tilde{y}|\hat{\theta}_k(D), M_k) p_t(\tilde{y}) d\tilde{y}$$

reference $$\bar{u}_*^P(M_k|D) = \int \log p(\tilde{y}|\hat{\theta}_k(D), M_k) p(\tilde{y}|D, M_*) d\tilde{y}$$

training $$\bar{u}_{\text{train}}^P(M_k|D) = \frac{1}{n}\sum_{i=1}^{n} \log p(\dot{y}_i|\hat{\theta}_k(D), M_k)$$

FIG 11. *Different forms that the logarithmic utility function can take in model selection problems.*

in the predictive model. On the other hand, existing prior information about the properties of future observations is ignored.

5.1.1. Posterior predictive

In the posterior predictive approach the utility function depending on the predictive distribution is evaluated at the observations. The "dot-notation" is used to emphasize that the exact replicates $(\dot{x}_i, \dot{y}_i)$ of observations (x_i, y_i) included in D are considered to be realizations from the future data distribution.

In single prediction each of the n observations is predicted individually. The estimate for the expected utility is simply the training utility Eq. (68). With logarithmic utility function the expected utility estimate is

$$\bar{u} \approx \bar{u}_{\text{train}}(M) = \frac{1}{n}\sum_{i=1}^{n} \log p(\dot{y}_i|\dot{x}_i, D, M), \tag{82}$$

and with the squared error loss function the expected loss is estimated by

$$\bar{s} \approx \bar{s}_{\text{train}}(M) = \frac{1}{n}\sum_{i=1}^{n}\left(\dot{y}_i - \mathbb{E}[\tilde{y}|\dot{x}_i, D, M]\right)^2. \tag{83}$$

In simultaneous prediction, all the future observations are treated as a random vector, and the prediction results can be different from single prediction when the joint predictive distribution is not equal to the product of independent marginal predictive distributions. While it may be reasonable to assume that the future observations are independent given the model parameters, integrating over the posterior distribution of the unknown parameters typically results in dependencies in the multivariate predictive distribution. With logarithmic utility function the estimated expected utility is given by

$$\bar{u} \approx \log p(\dot{y}_1, \ldots, \dot{y}_n|\dot{x}_1, \ldots, \dot{x}_n, D, M) = \log p(\dot{y}_{(1:n)}|\dot{x}_{(1:n)}, D, M). \tag{84}$$

The replicate sample $\{(\dot{x}_i, \dot{y}_i)\}_{i=1}^n$ is not independent of D in Eqs. (82)–(84). Conditioning the predictive model and evaluating the utility $\bar{u}$ with the same data $D = \{(x_i, y_i)\}$ gives over-optimistic results, which are reflected as a bias in the estimate for the expected utility. Furthermore, out-of-sample performance cannot be assessed as there is no evaluation outside the observed covariate values $x_{(1:n)}$ conditioning the predictive distribution.

The amount of optimism depends on how much the inclusion of a single observation pair (x_i, y_i) changes the posterior. Thus, if n is very large compared to the effective number of parameters p_{eff} of the model, the optimism may be negligible. Typically this is not the case with rich models. On the contrary, complex models tend to fit well to observations. Training utility as a measure for predictive performance may lead to favoring over-fitting and consequently selecting maximally complex models, a behaviour that is typically undesirable for a model selection strategy. Therefore, even if n is large, it is still safer to use, for example, the hold-out predictive approach (Section 5.1.2), which is equally simple to implement and has a similar computational burden.

A number of posterior predictive methods have been proposed, even though lately the posterior predictive approach has not been recommended because of the disadvantages of using the same data for both training and testing. Well-known examples of the simultaneous prediction include the posterior Bayes factor (Aitkin, 1991)

$$\text{PoBF}(M_1, M_2) = \frac{p(\dot{y}_{(1:n)}|\dot{x}_{(1:n)}, D, M_1)}{p(\dot{y}_{(1:n)}|\dot{x}_{(1:n)}, D, M_2)}, \tag{85}$$

written in a form intended for comparing two candidate models, and the M-criterion (Laud and Ibrahim, 1995)

$$\text{M-crit}(M) = \left(p(\dot{y}_{(1:n)}|\dot{x}_{(1:n)}, D, M)\right)^{-1/n}, \tag{86}$$

which has been scaled to units of a single outcome variable $\tilde{y}$. Gutiérrez-Peña and Walker (2001) use the single prediction posterior predictive estimate (Eq. (82))

for the $\mathcal{M}$-open case. The single prediction training utility appears also as a part of the widely applicable information criterion (WAIC) (Watanabe, 2010b) and the simultaneous prediction training utility as a part of a criterion by Ando and Tsay (2010). These two criteria also include a bias correction for the optimism.

Goodness-of-fit testing with Bayesian posterior predictive p-values (Guttman, 1967; Rubin, 1984; Meng, 1994; Bayarri and Berger, 1999, 2000; Robins, van Der Vaart and Ventura, 2000) and posterior predictive model checking (Gelman et al., 1995; Gelman, Meng and Stern, 1996) are examples of using the predictive distribution for *model criticism*. These approaches are useful for revealing inconsistencies between a model and data, but they are not recommended for model selection, because these procedures suffer from the same overfitting to the observations as the general posterior predictive approach.

5.1.2. Hold-out predictive

Hold-out predictive approach, also known as the *validation*, *test* or *partial predictive* method, may be seen as a compromise between the posterior predictive (training utility) and external validation (test utility) approaches discussed in Section 4.2. The observations are divided into a training set of n_t observations, indexed by $I_t = \{i_1, \ldots, i_{n_t}\}$, and a hold-out set of the remaining $n_h = n - n_t$ observations indexed by $I_h = \{i_{n_t+1}, \ldots, i_n\}$. The training set $D_{(I_t)}$ is used to compute the predictive distribution, and the hold-out set $D_{(I_h)}$ is used in evaluating the utility. Variations using a very small training set are referred to as partial methods by O'Hagan and Forster (2004).

In single prediction the estimated expected utility takes the form

$$\bar{u} \approx \frac{1}{n_h} \sum_{i \in I_h} \log p(y_i | x_i, D_{(I_t)}, M) \tag{87}$$

and in simultaneous prediction the form

$$\bar{u} \approx \log p(y_{(I_h)} | x_{(I_h)}, D_{(I_t)}, M). \tag{88}$$

It is evident from these equations that the hold-out predictive approach is equivalent to *test error* Eq. (63) in the case where a part of the data are held out for testing the model performance.

The hold-out approach allows for out-of-sample predictions for explanatory variables x outside the training data. As non-overlapping training set $D_{(I_t)}$ and hold-out set $D_{(I_h)}$ can be considered to be independent so that the optimism of the posterior predictive approach is avoided, but now a learning curve related bias is introduced as the predictive distribution is conditioned on $n_t < n$ observations. The choice of the split of the data into I_t and I_h is arbitrary as there are $\binom{n}{j}$ possible splits. It is reasonable to assume that there is variation in the expected utility estimate depending on the specific choice for a split. However, the hold-out method can be considered to be robust if there are enough observations in both the training and hold-out set.

The use of separate training and hold-out sets may be considered as decades old modeling folklore. Recently hold-out method has been used, for example, in covariate selection by Draper and Fouskakis (2000) with an application-specific monetary utility function and by Fearn, Brown and Besbeas (2002) with a single prediction logarithmic utility function, while O'Hagan (2003) gives an example of a simultaneous prediction with a Mahalanobis distance-type squared error loss function,

$$\bar{s} \approx \left(y_{(I_h)} - m_h\right)^T V_h^{-1} \left(y_{(I_h)} - m_h\right), \tag{89}$$

where $m_h = \mathbb{E}\left[\tilde{y}_{(I_h)}|x_{(I_h)}, D_{(I_t)}, M\right]$ is the mean of the joint hold-out predictive distribution for $\tilde{y}_{(I_h)}$ and $V_h = \mathrm{Cov}\left[\tilde{y}_{(I_h)}|x_{(I_h)}, D_{(I_t)}, M\right]$ is the covariance matrix.

5.1.3. Cross-validation predictive

Cross-validation (CV) methods for model assessment and comparison have been proposed by several authors: for early accounts see Stone (1974); Geisser (1975) and for the Bayesian cross-validation see Geisser and Eddy (1979); Gelfand, Dey and Chang (1992); Gelfand and Dey (1994); Bernardo and Smith (1994); Gelfand (1996). In the following we consider some of the most common CV approaches for Bayesian model selection. Many related variations and their properties in are reviewed by Arlot and Celisse (2010) in a non-Bayesian context.

Cross-validation can be considered to be an extension of the hold-out approach. The basic idea in the cross-validation approaches is to split the data into cross-validation sets indexed by $I_1, \ldots, I_K$. Each data subset $D_{(I_k)}$ is used as a validation set in turn, while the remaining sets form a training set $D_{(\setminus I_k)}$. In Bayesian setting it is desirable that the training sets $D_{(\setminus I_k)}$ are as similar to D as possible, which leads naturally to leave-one-out (LOO) CV.

Robustness is an advantage of the cross-validation approaches: the double use of data is not as severe as in the posterior predictive approach, as the same observations are never used simultaneously for training and evaluating the expected utility. Cross-validation methods also estimate the out-of-sample predictive performance. A major setback in CV is the larger computational burden than in the posterior or hold-out predictive approaches, although different approximations can be used to reduce the computational cost.

LOO-CV The leave-one-out cross-validation is a CV variant where each observation takes the role of the validation set in turn, which leads to a natural single prediction approach. With logarithmic utility function the expected utility estimate is given as

$$\bar{u} \approx \bar{u}_{\mathrm{LOO}}(M) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i|x_i, D_{(\setminus i)}, M), \tag{90}$$

and with squared error loss function as

$$\bar{s} \approx \bar{s}_{\mathrm{LOO}}(M) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbb{E}[\tilde{y}|x_i, D_{(\setminus i)}, M]\right)^2, \tag{91}$$

where

$$p(\tilde{y}|x_i, D_{(\setminus i)}, M) = \int p(\tilde{y}|x_i, \theta, M)p(\theta|D_{(\setminus i)}, M)d\theta \tag{92}$$

is the leave-one-out predictive density.

Watanabe (2010a) assumes a true model under fairly general regularity conditions for both D and $\tilde{y}$ and shows that the logarithmic LOO-CV utility is unbiased in the sense that for regular and singular statistical models $p(y|M)$ the expected logarithmic LOO-CV utility is asymptotically equal to the expected true utility in single prediction (expectation taken over all training datasets). With finite data the error is of order $o(1/n)$, so that when n is not very small, LOO-CV can be considered as nearly unbiased.

However, in model selection the selection induced bias (Section 4.4) makes the expected utility estimate of the selected model biased even when using the (nearly) unbiased CV-estimates. The model selection induced bias can be taken into account by the double/nested/2-deep cross-validation (e.g. Stone, 1974; Jonathan, Krzanowski and McCarthy, 2000) or making an additional bias correction (Tibshirani and Tibshirani, 2009).

The computational burden of the LOO-CV approach can be overwhelming especially with large datasets, as the posterior distribution of the parameters must be computed separately for each of the n LOO-CV predictive distributions. For certain models such as linear models and Gaussian processes with fixed hyperparameters the LOO-CV utility can be computed analytically if the observation model is Gaussian (see, e.g., Shao, 1993; Orr, 1996; Peruggia, 1997; Sundararajan and Keerthi, 2001) or approximated efficiently with expectation propagation (Opper and Winther, 2000; Rasmussen and Williams, 2006) or Laplace approximation (Held, Schrödle and Rue, 2010). In a more general case the number of required computational operations can be reduced, for example, with importance-sampling (IS) LOO-CV or k-fold CV.

IS-LOO-CV Importance sampling leave-one-out cross-validation by Gelfand, Dey and Chang (1992) is a computationally efficient way of approximating the LOO-CV with Monte Carlo sampling. In a straightforward Monte Carlo approach, the LOO predictive density can be approximated as

$$\begin{aligned} p(\tilde{y}|\tilde{x}, D_{(\setminus i)}, M) &= \int p(\tilde{y}|\tilde{x}, \theta, D_{(\setminus i)}, M)p(\theta|D_{(\setminus i)}, M)d\theta \\ &\approx \frac{1}{T}\sum_{t=1}^{T} p(\tilde{y}|\tilde{x}, \theta^{(\setminus i),t}, D_{(\setminus i)}, M), \end{aligned} \tag{93}$$

where $\theta^{(\setminus i),t}$ are samples from the leave-one-out posterior $p(\theta|D_{(\setminus i)}, M)$. Instead of sampling from the LOO-posteriors for each i separately, the IS-LOO-CV approach uses the full posterior $p(\theta|D, M)$ as an importance sampling distribution

for each $p(\theta|D_{(\backslash i)}, M)$. In short, this means that the samples θ^t from the full posterior can be weighted using importance sampling weights

$$\frac{p(\theta^t|D_{(\backslash i)}, M)}{p(\theta^t|D, M)} \propto \frac{1}{p(y_i|x_i, \theta^t, D_{(\backslash i)}, M)} = w^{(\backslash i),t}. \tag{94}$$

The expected utility estimate is then given by

$$\bar{u} \approx \bar{u}_{\text{IS-LOO}}(M) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{t=1}^{T} \frac{w^{(\backslash i),t} p(y_i|x_i, \theta^t, M)}{\sum_{t=1}^{T} w^{(\backslash i),t}}\right) \tag{95}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{\frac{1}{T}\sum_{t=1}^{T} \frac{1}{p(y_i|x_i,\theta^t,M)}}\right). \tag{96}$$

General conditions of the convergence of the importance sampling are given by Geweke (1989). The reliability of the importance sampling can be estimated by examining the variability of the importance weights. For simple models the variance of the importance weights may be computed analytically. For example, the necessary and sufficient conditions for the variance of the case-deletion importance sampling weights to be finite for a Bayesian linear model are given by Peruggia (1997). Epifani, MacEachern and Peruggia (2008) extend the results and provide analytical results for generalized linear models and Michaelis-Menton models to assess whether the estimators satisfy a central limit theorem. In the general case, an efficiency estimate of the importance sampling can be computed from the obtained weights (see Newton and Raftery, 1994; Gelman et al., 1995, ch. 10; Peruggia, 1997; Vehtari and Lampinen, 2002), but this approach can not prove convergence.

Most often resampling in IS-LOO-CV is made with replacement. Gelman et al. (1995) and Stern and Cressie (2000) recommend resampling without replacement to reduce the variance due to highly variable importance weights. Skare, Bølviken and Holden (2003) show that resampling without replacement gives a smaller total variance error.

Bhattacharya and Haslett (2007) describe a useful variation of IS-LOO-CV for inverse problems (they also use resampling without replacement). The tails of the full posterior are usually thinner than the LOO posterior tails, Bhattacharya and Haslett (2007) propose to use one of the LOO distributions as importance distribution and propose ways to measure which of the LOO distributions would be central, that is, close to every other LOO distribution. The same idea could be used for forward models to improve the importance sampling.

Bornn, Doucet and Gottardo (2010) improve reliability by using sequential Monte Carlo and tempered sequence of distributions from the full posterior to the leave-one-out posteriors.

Plummer's (2008) penalized loss function method uses the IS-LOO-CV approach with the Gibbs log-score.

k-fold-CV In the k-fold-CV approach the data are split into $k \ll n$ subsets, or folds, (usually $k \sim 10$), each of which is in turn used as the validation set while

the remaining data are used for model training. The k-fold-CV can be used to reduce the computation time in single prediction, as only k posterior evaluations are needed instead of n in the LOO-CV. The expected utility estimate of the k-fold CV is given by

$$\bar{u} \approx \bar{u}_{\text{k-CV}}(M) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i, D_{(\setminus I_{k(i)})}, M), \tag{97}$$

where $k(i)$ refers to indices of observations in the same CV fold as the observation (y_i, x_i).

The k-fold-CV is always conditioned on less than n observations, which leads to a biased expected utility estimate. The bias can be estimated and a corrected CV estimate can be computed (e.g., Burman, 1989; Fushiki, 2011).

Splitting the data in different ways leads to variability in the results (e.g. Chakrabarti and Ghosh, 2007; Arlot and Celisse, 2010) which, however, may be small compared to other variabilities (Vehtari and Lampinen, 2002). The variability due to a specific data-split can be reduced by computing all the possible data splits or repeating the data splitting randomly and averaging the results. This is rarely done due to the resulting increase in computation time.

The sampling error for k-fold-CV can be estimated (e.g., Dietterich, 1998) by first computing the expected utility for each fold

$$\bar{u}^{(j)} = \frac{1}{n_j}\sum_{i\in I_j} \log p(y_i|x_i, D_{(\setminus I_j)}, M), \quad j = 1, \ldots, k, \tag{98}$$

where I_j are the n_j indices in the jth fold, and then by a variance estimate

$$\widehat{\text{Var}}(u) = \frac{1}{k-1}\sum_{j=1}^{k} \left(\bar{u}_{\text{k-CV}} - \bar{u}^{(j)}\right)^2. \tag{99}$$

Compared to Eqs. (74)–(76) the $\bar{u}^{(j)}$'s are closer to Gaussian, but a drawback is an increased variance in the variance estimate itself.

The k-fold-CV can be used for block simultaneous prediction ($n_{block} < n$), but similarly as with any CV approach, full simultaneous prediction is not possible. Block simultaneous prediction can be useful for hierarchical models.

For time series with unknown finite range dependencies, the k-fold-CV can be combined with the h-block-CV proposed by Burman, Chow and Nolan (1994). Instead of just leaving the ith point out, additionally a block of h cases from either side of the ith point is removed from the training data for the ith point. The value of h depends on the dependence structure, and it could be estimated for example from autocorrelations. Burman and Nolan (1992) show that $h = 0$ could be used in the case of stationary Markov process and squared error loss function (or a utility well approximated by a quadratic form) (see also Akaike, 1973). However, in real world problems the exact properties of the process are usually not known. The approach could also be applied in other models with finite range dependencies (e.g., spatial models), by removing a block of h cases from around the ith point.

Other variations Other variations of CV have been proposed to achieve specific desirable properties. In non-Bayesian settings smaller training datasets are often useful. Sometimes it is desired that $D_{(\setminus I_k)}$ would be as independent as possible, to simulate the effect of conditioning on an unknown dataset D_n in Equation (60). In such cases, specific cross-validation or bootstrap methods can be used. Using smaller training sets is sometimes combined with exhaustive, repeated, Monte Carlo or balanced incomplete design data splitting to reduce the random splitting variance (e.g., Dietterich, 1998; Nadeau and Bengio, 2000; Arlot and Celisse, 2010). Taking the expectation over the training datasets has also been proposed as an answer to the question "Given two learning methods A and B and training data D, which *method* will produce a more accurate *model* when trained on new training data D_n of the same size as D?" (Dietterich, 1998). An extreme case in machine learning is to use completely independent training and validation sets, which requires large amounts of data (Rasmussen et al., 1996; Neal, 1998).

Smaller training sets produce a bias which depends on the training set size and the learning curves of the models. This bias provides additional penalty for more complex models which may be helpful in model selection if no other penalty or prior favoring simpler models is used (Arlot and Celisse, 2010).

To our best knowledge, there is not yet an asymptotic consistency result for the Bayesian cross-validation in *model selection*, but results from non-Bayesian literature (Yang, 2005, 2007; Arlot and Celisse, 2010) imply that the usual Bayesian LOO-CV would not be consistent, although an alternative Bayesian CV could be constructed which could be consistent with the sufficient condition depending on the convergence rates of the models and the asymptotic data division.

5.1.4. Approximative CV for hierarchical models

The computational burden of the LOO-CV can be alleviated with the approximative approaches in hierarchical models, where the full posteriors conditional on the full data D are used in some levels of the models instead of the cross-validation posteriors.

Cross-validation of only lower level in hierarchical model Consider a case of a hierarchical model in which the observation depends on a latent value and hyperparameters according to a model $p(y_i|f_i,\theta)$. The parameters of the model have a joint prior $p(\mathbf{f},\theta) = p(f_1,\ldots,f_n,\theta)$. For some models such as the Gaussian process the LOO-CV density

$$p(y_i|x_i, D_{(\setminus i)}, \theta, M) = \int p(y_i|f_i,\theta,M)p(f_i|x_i, D_{(\setminus i)},\theta,M)df_i \tag{100}$$

conditioned on the hyperparameters can be either computed analytically or approximated efficiently, for example, with expectation propagation or Laplace

approximation. The full LOO-CV predictive density follows from integrating over the LOO-posterior of the hyperparameters. Instead, an approximation

$$p(y_i|x_i, D_{(\setminus i)}, M) = \int p(y_i|x_i, D_{(\setminus i)}, \theta, M)p(\theta|D_{(\setminus i)}, M)d\theta$$
$$\approx \int p(y_i|x_i, D_{(\setminus i)}, \theta, M)p(\theta|D, M)d\theta \tag{101}$$

can be used. In the approximation the posterior distribution of the hyperparameters $p(\theta|D, M)$ is computed using all data, but the hyperparameter-conditioned predictive distributions are computed using $D_{(\setminus i)}$. This computation time saving approximation is a reasonable alternative if removing (x_i, y_i) has only a small impact on $p(\theta|D, M)$ but a larger impact on $p(y_i|x_i, \theta, D, M)$.

Ghosting Another example of a mixed CV and posterior predictive approach is the *ghosting* approach by Marshall and Spiegelhalter (2003) who consider hierarchical models and predictive p-value checks to compare whether observed values are in the extreme tails of the predictive distributions. For latent variable models the LOO predictive density can be written as

$$p(y_i|x_i, D_{(\setminus i)}, M) = \int p(y_i|x_i, f_i, \theta, M)p(f, \theta|D_{(\setminus i)}, M)d\theta df. \tag{102}$$

Marshall and Spiegelhalter approximate the second term in the integral to get

$$p(y_i|x_i, D_{(\setminus i)}, M) \approx \int p(\dot{y}_i|\dot{x}_i, f_i, \theta, M)p(f_i|f_{\setminus i}, \theta, M)p(f_{\setminus i}, \theta|D, M)d\theta df, \tag{103}$$

where $\dot{x}_i$ and $\dot{y}_i$ are exact replicates as x_i and y_i are included in D. Data are used twice, but the bias is smaller than in the posterior predictive approach as y_i does not directly affect f_i.

5.2. $\mathcal{M}$-closed/completed treatment for $\tilde{y}|\tilde{x}$ and $\mathcal{M}$-open for $\tilde{x}$

Predictive methods with an explicit model for the conditional part $p(y|x)$ and a sample re-use approach for $p(x)$ are close to the full $\mathcal{M}$-closed/completed treatment while avoiding the difficulty in modeling the distribution of x. Many of the methods in this section deviate from the full $\mathcal{M}$-closed/completed treatment for $p(y|x)$, but still they use at least partially an explicit model for the conditional part $p(y|x)$.

5.2.1. Reference predictive

In the $\mathcal{M}$-closed and $\mathcal{M}$-completed views beliefs about future observations given the explanatory variables are represented by the actual belief model $p(\tilde{y}|\tilde{x}, D, M_*)$. The common term *reference model* is used for all the various definitions of the

actual belief model M_*, and the following model selection approaches can be grouped together under the heading of *reference predictive* methods. The reference predictive approach takes different forms depending not only on the utility function but more importantly on the reference model M_*, which can be an encompassing model, a full model including all the explanatory variables, the Bayesian model average predictive model, or a non-parametric model.

For example, with logarithmic utility function the expected utility in single prediction is

$$\bar{u} \approx \bar{u}_*(M_k) = \frac{1}{n} \sum_{i=1}^{n} \int \log p(\tilde{y}|\dot{x}_i, D, M_k) p(\tilde{y}|\dot{x}_i, D, M_*) d\tilde{y} \tag{104}$$

and in simultaneous prediction

$$\bar{u} \approx \bar{u}_*(M_k) = \int \log p(\tilde{y}_{(1:n)}|\dot{x}_{(1:n)}, D, M_k) p(\tilde{y}_{(1:n)}|\dot{x}_{(1:n)}, D, M_*) d\tilde{y}. \tag{105}$$

The expected utility of each candidate model M_k is computed with respect to the reference model M_* serving as a common yardstick for comparing the predictive performance of the competing models. From the Eqs. (9) and (11) it is clear that the maximization of the expected logarithmic utility function corresponds to minimizing the Kullback-Leibler divergence

$$d_{\mathrm{KL}} \left\{ p(\tilde{y}|\dot{x}, D, M_*), p(\tilde{y}|\dot{x}, D, M_k) \right\} = \int \log \frac{p(\tilde{y}|\dot{x}, D, M_*)}{p(\tilde{y}|\dot{x}, D, M_k)} p(\tilde{y}|\dot{x}, D, M_*) d\tilde{y} \tag{106}$$

from the reference model M_* to a candidate model M_k. In other words, the reference predictive approach in model selection equals to searching for a candidate model with the predictive distribution similar to the reference model. Corresponding (unconditional) squared error loss function results can be found in Section 3.3.1.

In order to estimate the out-of-sample performance at locations $\tilde{x}$ not necessarily included in the training sample $(x_1, \ldots, x_n)$ cross-validation predictive densities could be used instead of posterior predictive densities used in Eq. (104).

If the reference model does not describe the future data sufficiently well, it is obvious that the expected utility estimates are likely to be biased. Also, if the reference model has already been overfitted to the training data, the selection process favors also overfitted models. However, the reference predictive model selection process itself does not cause additional fitting to the data, as the degree of fit to the data is determined by the reference model. That is, there is no selection induced bias in the reference predictive approach, which has been demonstrated by Vehtari and Lampinen (2004).

If the goal is to discover a simple model, an additional cost c_k for model complexity can be included in the utility function. Without penalty for the complexity, the model maximizing the utility would be the reference model itself. Instead of adding an additional cost for model complexity, it is also possible

to choose the simplest model for which expected utility is not practically significantly different from the expected utility of the reference model. Practical significance can be based on expert information or calibration of the expected utility (Section 5.7.2).

BMA reference model In the Bayesian predictive criterion by San Martini and Spezzaferri (1984) the reference model is the Bayesian model average, so that the expected utility in single prediction for model M_k is given by

$$\bar{u}_*(M_k) = \frac{1}{n}\sum_{i=1}^{n}\int \log p(\tilde{y}|\dot{x}_i, D, M_k)p_{\mathrm{BMA}}(\tilde{y}|\dot{x}_i, D)d\tilde{y}. \tag{107}$$

The criterion requires assigning priors to the set of possible models $\{M_k\}_{k=1}^K$, and computing the BMA predictive density.

BMA as the reference model has been criticized by an argument that setting a prior on the model space requires assuming that one of the models is the true model, that is, assuming the $\mathcal{M}$-closed view. However, from the $\mathcal{M}$-completed point of view averaging over models in the BMA predictive density is analogous to integrating over continuous parameters: what matters is whether the predictive model is rich enough to describe actual beliefs about the future observations.

Encompassing reference model Ibrahim and Laud (1994) and Laud and Ibrahim (1995) consider explanatory variable selection with a designed x and a replicated experiment (Section 4.3). For a reference model they propose the encompassing model (all covariates included) or the intercept model (no covariates included).

Instead of the logarithmic utility function equivalent to the KL divergence from the actual belief model to the candidate model they proposed the K-criterion

$$\begin{aligned}\text{K-crit}(M_k) = {} & d_{\mathrm{KL}}\left\{p(\tilde{y}|\dot{x}, D, M_*), p(\tilde{y}|\dot{x}, D, M_k)\right\} \\ & + d_{\mathrm{KL}}\left\{p(\tilde{y}|\dot{x}, D, M_k), p(\tilde{y}|\dot{x}, D, M_*)\right\}.\end{aligned} \tag{108}$$

The second term in the sum of the two Kullback-Leibler divergences corresponds to the estimation of the expected utility of the reference model M_* with respect to each candidate model M_k. As there is usually less confidence in the candidate models being accurate descriptions for the future observations, the K-criterion is not so much a useful measure of the predictive performance of the candidate models, but more a measure of how different predictions the models M_* and M_k make.

Non-parametric reference models Various non-parametric models – obtained by setting a specific prior on a function space – are often chosen as reference models. Simpler parametric models, for example, the candidate models, can be expanded into non-parametric models in order to obtain a rich enough

model to serve as a reference model. Gutiérrez-Peña (1997) proposed to use a Gaussian process model centered on one of the candidate parametric models and Trottini and Spezzaferri (2002) proposed a mixture of Gaussian processes each centered on a different model as the reference model. The likelihood in a Gaussian process model is usually parametric, and the term semiparametric is sometimes used for such a combination of non-parametric and parametric models. However, the likelihood could also be non-parametric or some other fully non-parametric model could also be used.

A fully non-parametric model is not automatically a good reference model. For example, Gutiérrez-Peña and Walker (2001) propose a Dirichlet process as a non-parametric model reference model; the expectation of this non-informative fully non-parametric model equals to the single prediction posterior predictive estimate (Eq (82)). Gutiérrez-Peña and Walker (2005) proposed a Dirichlet process mixture in a related projection approach (see section 5.4) and Karabatsos (2006) proposed a Polya tree model. Both of the latter approaches, while presented only for predicting one-dimensional data, are improvements to the non-informative Dirichlet process as there is a positive probability for also other points than the observed ones.

5.2.2. Mixed reference and posterior predictive

Gutiérrez-Peña and Walker (2001) proposed to use a mixture of the BMA model and the Dirichlet process model. In their approach the resulting expected utility estimate is a weighted average of the posterior predictive expected utility estimate and the reference predictive estimate. The weighting is selected by the user. Gutiérrez-Peña and Walker propose how the weighting can be chosen based on prior predictive distributions, given that proper priors have been specified.

5.2.3. Self predictive

In the reference predictive approach the models are compared with respect to a common reference model M_*. In the self predictive approach this common yardstick is not available, and the candidate models are assessed based on their own predictive properties as described in Section 3.2.

For example, Bernardo and Bermúdez (1985) describe a self predictive approach. The single prediction expected utility with the logarithmic utility function for any candidate model M_k is given as

$$\bar{u} \approx \bar{u}_k(M_k) = \frac{1}{n}\sum_{i=1}^{n}\int p(\tilde{y}|\dot{x}_i, D, M_k)\log p(\tilde{y}|\dot{x}_i, D, M_k)d\tilde{y}, \tag{109}$$

which can be seen to be equivalent to the negative differential entropy of the predictive distribution. Maximizing this utility with respect to models is equivalent to finding a model with minimum predictive entropy. The corresponding (unconditional) squared error loss function results can be found in Section 3.1.2.

Similarly as in the reference predictive approach the cross-validation predictive densities could be used in order to take into account the out-of-sample performance at locations $\tilde{x}$ not necessarily included in D.

Using the model-conditioned predictive distribution $p(\tilde{y}|\tilde{x}, D, M_k)$ as a belief model for the future data results in a smaller variance for the expected utility estimate for the said model. The dependence on the adequacy of the model makes the self predictive approach highly sensitive to model assumptions, which emphasizes the importance of model criticism. For example, if candidate models include a model with a zero variance predictive distribution, the self predictive approach will choose that model.

Gneiting, Balabdaoui and Raftery (2007) discuss measuring the *sharpness* (concentration) of the predictive distribution and recommend checking the *calibration* (statistical consistency) between the predictive distribution and the observations, a form of model criticism. Model criticism and checking based on the posterior predictions and the data may detect some problems, but severely overfitted models may seem to be posterior calibrated and provide very sharp predictions. For example, in classification with flexible models, it is possible that a model can give class probabilities very close to one with seemingly good calibration.

In the predictive entropy approach by Corander and Marttinen (2006), the posterior expectation of conditional predictive entropy is taken for each model

$$\bar{u} \approx \iint p(\tilde{y}|\tilde{x}, \theta, M_k) \log p(\tilde{y}|\tilde{x}, \theta, M_k) p(\theta|D, M_k) d\tilde{y} d\theta. \tag{110}$$

Corander and Marttinen (2006) consider only unconditional models (without explanatory variables x) and use the Gibbs utility form to make the calculations easier.

5.2.4. Mixed self and posterior predictive

There is no inherent safeguard against poor fit in the self predictive approach: the self predictive expected utility estimate is maximized with as narrow a predictive distribution as possible. Several authors have proposed methods in which the closeness to observations is included into the utility function to satisfy an informal goal of obtaining predictions close to the actual observations.

L-criterion The L-criterion (Ibrahim and Laud, 1994; Laud and Ibrahim, 1995) is based on assuming a designed x, a replicated experiment and a squared error loss function

$$L^2(M_k) = \mathbb{E}_{\tilde{y}} \left[(\tilde{y} - y)^T (\tilde{y} - y) | \dot{x}_i, D, M_k \right] \tag{111}$$

$$= \sum_{i=1}^{n} \mathbb{E}_{\tilde{y}} \left[(y_i - \tilde{y})^2 | \dot{x}_i, D, M_k \right]. \tag{112}$$

The difference from the posterior predictive squared error, Eq. (83), is the expectation of the squared error taken over $\tilde{y}$ instead of using expectation of $\tilde{y}$ as the prediction. Given the loss function the L-criterion is defined as

$$\text{L-crit}(M_k) = \sqrt{L^2(M_k)}, \tag{113}$$

which is scaled to have the same units as the outcome variable. Eq. (112) can be decomposed as

$$\sum_{i=1}^{n} \left(y_i - \mathbb{E}[\tilde{y}|\dot{x}_i, D, M_k]\right)^2 + \sum_{i=1}^{n} \text{Var}[\tilde{y}|\dot{x}_i, D, M_k], \tag{114}$$

where we see that there is no difference between simultaneous and single prediction due to the form of the loss function. In the L-criterion literature the focus is in model selection; how to use the selected model for prediction is not discussed. If the expectation $\mathbb{E}_{\tilde{y}|\dot{x}_i, D, M_k}[\tilde{y}]$ is used for prediction the L-criterion corresponds to an expected utility estimate with a form of squared distance between the mean of the predictive distribution and the observation (the posterior predictive part) plus the predictive variance (the self predictive part). In model comparison, L-criterion penalizes the more complex model asymptotically with penalty halfway between the posterior predictive approach and the cross-validation predictive approach (Ibrahim and Laud, 1994), which is natural as the L-criterion is a sum of posterior and self predictive estimates.

Ibrahim and Chen (1997) presented two multivariate version of L-criteria,

$$\text{L-crit}_m(M_k) = |R_m|^{1/2p} \tag{115}$$

$$\text{L-crit-J}_m(M_k) = (\text{tr}(R_m))^{1/2} \tag{116}$$

where $|\cdot|$ denotes determinant, p is dimension of y, and

$$R_m = \sum_{i=1}^{n} \left\{ \left(\mathbb{E}[\tilde{y}_i|\dot{x}_i, D, M_k] - y_i\right) \left(\mathbb{E}[\tilde{y}_i|\dot{x}_i, D, M_k] - y_i\right)^T + \text{Cov}[\tilde{y}_i|\dot{x}_i, D, M_k] \right\}. \tag{117}$$

Meyer and Laud (2002) presented how to estimate the L-criterion for generalized linear models.

Posterior predictive criterion Gelfand and Ghosh (1998) discuss single prediction involving a choice of the optimal point prediction for the future observation. For a given model M the optimal point prediction follows from maximizing the expected utility

$$\hat{a}_i = \arg\max_{a_i} \left\{ \mathbb{E}\left[u(\tilde{y}_i, a_i) + k u(y_i, a_i) | D, M_k\right] \right\}, \tag{118}$$

where k is a parameter controlling the relative importance of the two terms in the utility function; $u(\tilde{y}_i, a_i)$ measures the utility for predicting the future observations modeled with the predictive distribution of model M_k, and $u(y_i, a_i)$

gives the utility of predicting the observed data. The resulting optimal point prediction is designed to be close to both the future and observed data. Determining the optimal point prediction is an integral and non-trivial part of the approach.

Although the utility function can be of any shape, Gelfand and Ghosh (1998) discussed the squared error loss function in detail, as it gives analytic results. The single prediction expected loss is given by

$$D_k(M) = \sum_{i=1}^{n} \min_{a_i} \mathbb{E}\left[u(\tilde{y}_i, a_i) + ku(y_i, a_i)|D, M\right] \tag{119}$$

$$= \frac{k}{k+1} \sum_{i=1}^{n} \left(y_i - \mathbb{E}[\tilde{y}|\dot{x}_i, D, M]\right)^2 + \sum_{i=1}^{n} \mathrm{Var}[\tilde{y}_i|\dot{x}_i, D, M], \tag{120}$$

where the optimal point prediction is $a_i = \frac{1}{k+1}\,\mathbb{E}[\tilde{y}|\dot{x}_i, D, M] + \frac{k}{k+1} y_i$. In the limit $k \to \infty$, $D_k(M)$ becomes the L-criterion (Section 5.2.4) with a rather unconventional prediction $a_i = \tilde{y}_i$. When $k = 0$, the criterion reduces to the sum of marginal predictive variances with a point prediction $a_i = \mathbb{E}[\tilde{y}|\dot{x}_i, D, M]$, which is same as the self-predictive approach (Section 5.2.3).

Sinha, Chen and Ghosh (1999) suggest using $k = 1$ giving an equal weight for the terms in Eq (118). The optimal point prediction at any observed $\dot{x}_i$ is $a_i = \frac{1}{2}\,\mathbb{E}[\tilde{y}|\dot{x}_i, D, M] + \frac{1}{2} y_i$, which is halfway between the posterior mean and the observation. If, contrary to the derivation of the criterion, the posterior expectation $\mathbb{E}[\tilde{y}|\dot{x}_i, D, M]$ is used as a point prediction, the expected loss estimate becomes the posterior predictive estimate divided by two plus the self predictive estimate.

Ibrahim, Chen and Sinha (2001) adopt the squared error criterion by Gelfand and Ghosh (1998), calling it the L-measure. They write $\nu = k/(k+1)$ and show that in covariate selection for linear models with orthogonal covariates and a conjugate prior the true model achieves on average the smallest L-measure when compared to other candidate models if $\frac{1}{2+\sigma_0^2} < \nu < \nu \frac{1+\sigma_0^2}{2+\sigma_0^2}$. Ibrahim, Chen and Sinha (2001) used value $\nu = 1/2$ in their experiments.

Gelfand and Ghosh (1998) also discuss the logarithmic utility function and exponential family models with the location parameter as the point prediction. Ibrahim, Chen and Sinha (2001) used the L-measure with generalized linear models following Meyer and Laud (2002).

Chen, Dey and Ibrahim (2004) proposed a weighted quadratic loss L-measure for generalized linear models and especially for the categorical data models

$$L = \sum_{i=1}^{n} \int \frac{L_i}{\mathrm{Var}[\tilde{y}_i|w_i\dot{x}_i, \theta, D, M]} p(\theta|D, M) d\theta, \tag{121}$$

where the L_i-measure for each i is divided by the conditional predictive variance, with an additional ad hoc weighting w_i for $\dot{x}_i$, and the expectation is taken over the posterior of θ. When the predictive variances are not equal, normalizing by the predictive variance, makes the loss to better approximate the logarithmic

loss. Chen, Dey and Ibrahim (2004) do not give any explanation for the need of introducing weighting w_i, but report that their loss measure is robust in the range $0.3 \leq w_i \leq 0.6$ and $0.4 \leq \nu \leq 0.6$.

5.2.5. Mixed self and cross-validation predictive

Mitchell and Beauchamp (1988) proposed a predictive error based on a squared error loss function and cross-validation

$$\mathrm{PE} = \sqrt{(1/n)\sum_{i=1}^{n} \mathbb{E}[y_i - \tilde{y}_i]^2}, \tag{122}$$

where the expectation is over the CV predictive density $p(\tilde{y}_i|x_i, D_{(\setminus i)}, M_k)$. PE is a cross-validation version of the L-criterion (section 5.2.4) and L_q-criterion by Marriott, Spencer and Pettitt (2001) is equal to cross-validation version of the L^2-criterion. The decomposed form of PE^2 and L_q

$$\mathrm{PE}^2 = \frac{1}{n}\sum_{i=1}^{n} \left\{ \left(\mathbb{E}[\tilde{y}_i|D_{(\setminus i)}] - y_i\right)^2 + \mathrm{Var}[\tilde{y}_i|D_{(\setminus i)}] \right\} \tag{123}$$

is composed of the squared distance between the mean of the leave-one-out predictive distribution and the observation (cross-validation predictive estimate) plus the leave-one-out predictive variance (cross-validation self predictive estimate) (compare to Eq. (114)). A similar mixed self and cross-validation predictive estimate could be formed for the logarithmic utility function.

5.2.6. Mixed reference and self predictive

Young (1987a) proposed for covariate selection a predictive criterion based on a weighted sum of the reference predictive and the self predictive squared error

$$\mathbb{E}\left[(\tilde{y} - \mathbb{E}[\tilde{y}|D, M_k])^2|D, M_*\right] + w\,\mathrm{Var}[\tilde{y}|D, M_k], \tag{124}$$

where M_* is the full model consisting of all the explanatory variables. Young proposed to analyse the sensitivity of selection with regard to w, but the final choice is based on an ad hoc choice of which part of the criterion is decided to be more important.

5.3. $\mathcal{M}$-closed/completed treatment for $\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}$

In the full $\mathcal{M}$-closed/completed approach, the explanatory variables are explicitly modeled. For example, a parametric prior model $p(x|\varphi, M_*)p(\varphi|M_*)$ independent of the other model parameters and the outcome variable y leads, via

a standard Bayesian treatment, to the predictive distribution for an unknown future explanatory variable

$$p(\tilde{x}|x_{(1:n)}, M_*) = \int p(\tilde{x}|\varphi, M_*)p(\varphi|x_{(1:n)}, M_*)d\varphi. \tag{125}$$

The predictive distribution can be now used as a reference model $p(\tilde{x}|D, M_*)$ for the future explanatory variables. For example, for the logarithmic utility function the expected utility is

$$\bar{u}_*(M_k) = \iint \log p(\tilde{y}|\tilde{x}, D, M_k)p(\tilde{y}|\tilde{x}, D, M_*)p(\tilde{x}|D, M_*)d\tilde{y}d\tilde{x}. \tag{126}$$

If the model $p(\tilde{x}|D, M_*)$ is an accurate description of the future observations, the variance of the expected utility estimate is smaller than in the sample re-use approaches. However, in a typical case the modeling of explanatory variables is difficult; this is the usual reason for using conditional models in the first place.

In the context of variable selection in linear regression, Lindley (1968) forms models $p(\tilde{x}|\varphi, M_*)$ and $p(\tilde{y}|\tilde{x}, \theta, D, M_*)$ and the corresponding predictive distributions $p(\tilde{x}|M_*)$ and $p(\tilde{y}|\tilde{x}, D, M_*)$. The actual belief model M_* is the encompassing model with all the covariates included. Lindley conditions the inference on all observed covariates in D and makes the selection of which covariates will be measured in the future, so that

$$p(\tilde{y}|\tilde{x}_k, D, M_*) = \int p(\tilde{x}|\tilde{x}_k, M_*)p(\tilde{y}|\tilde{x}, D, M_*)d\tilde{x}, \tag{127}$$

where $p(\tilde{x}|\tilde{x}_k, M_*)$ is obtained from the joint prior distribution $p(\tilde{x}|M_*)$. Lindley used a Gaussian model for $p(\tilde{x}|M_*)$, a Gaussian linear model with known noise variance σ^2 for $p(\tilde{y}|\tilde{x}, D, M_*)$, non-informative priors and a squared error loss

$$\mathbb{E}_{\tilde{y},\tilde{x}}\left[(\tilde{y} - \mathbb{E}[\tilde{y}|\tilde{x}_k, D, M_*])^2 \,|D, M_*\right]. \tag{128}$$

With these choices the predictive distribution of the submodel simplifies to a regular submodel conditioned on the selected covariates, that is

$$p(\tilde{y}|\tilde{x}_k, D, M_*) = p(\tilde{y}|\tilde{x}_k, D_k, M_k). \tag{129}$$

In this case, Lindley's approach is a reference predictive approach with a parametric model for $p(\tilde{x}|M_*)$ and $p(\tilde{y}|\tilde{x}, M_*)$. Due to canceling terms the specific form of $p(\tilde{x}|M_*)$ is not important when computing the expected difference to the reference model as knowing the covariance of the posterior predictive distribution for $\tilde{x}$ is sufficient.

Lindley's approach has been extended to multivariate y and a non-conjugate prior by Brown, Fearn and Vannucci (1999) and to BMA reference and candidate models with conjugate prior by Brown, Vannucci and Fearn (2002), and used for multivariate y and a conjugate prior by Vannucci, Brown and Fearn (2003). Brown, Vannucci and Fearn (2002) and Barbieri and Berger (2004) presented

also a version with a BMA reference model corresponding to the BMA reference predictive approach by San Martini and Spezzaferri (1984), except that Brown, Vannucci and Fearn and Barbieri and Berger use squared error and integrate analytically over the predictive distribution of $\tilde{x}$ (again for model selection the determination of the covariance of the posterior predictive distribution for $\tilde{x}$ is sufficient). If the covariates are orthogonal, the best single model according to this criterion is the median probability model, that is, the model including the variables having the marginal posterior probability greater than or equal to 1/2 (Barbieri and Berger, 2004).

Fearn, Brown and Besbeas (2002) use a generative mixture model and one of their proposed approaches uses simulated replicate observations from the joint $p(\tilde{x}, \tilde{y}|M_*) = p(\tilde{y}|M_*)p(\tilde{x}|\tilde{y}, M_*)$, where $p(\tilde{x}|\tilde{y}, M_*)$ is a multivariate normal.

5.4. Projection methods

In projection methods, the predictive properties of a reference model are projected onto a candidate model. The definition of the decision problem and the structure of the candidate model together determine which properties of the reference model are reflected in the resulting projection.

A major difference between the reference predictive approaches and the projection approaches is that a full prior specification is required only for the actual belief model M_*. Values for (at least some) of the unknown parameters of the candidate models are determined by the optimal projection. For example, in variable selection, information (or equivalently, uncertainty) about parameters related to explanatory variables included in the actual belief model but not in the candidate model may be projected onto parameter estimates related to the explanatory variables included in the candidate model. Thus, relevant predictive properties of the actual belief model may be partially conserved in the candidate models, even if the candidate model specification cannot take into account these aspects (for example, explanatory variables not included in the candidate model).

Predictive parametric point estimation Leamer (1979), commenting the use of information criteria for covariate selection in linear regression, presented a Bayesian predictive point estimation (Section 3.4.1) solution where the submodel parameters are obtained by minimizing the expected KL-divergence loss function (following the information criteria ideas) from the actual belief model to a submodel, with the expectation taken over the posterior distribution of the unknown parameters of the actual belief model. Leamer illustrated the idea with a linear subspace projection (dimensionality reduction) of multivariate Gaussian with an unknown mean and a diagonal covariance. The result for the linear model parameters is the usual least squares result.

Tran, Nott and Leng (2011) proposed a Lasso-type (Tibshirani, 1996) predictive point estimation approach for variable selection in generalized linear models. The predictive Lasso of Tran, Nott and Leng is formulated directly by the predictive distributions as in Section 3.3.2.

Predictive posterior approximation Lacoste-Julien, Huszár and Ghahramani (2011) present the general idea of predictive posterior approximation (posterior projection $q(\theta_k)$ in Section 3.4.1), which they refer to as loss-calibrated posterior approximation. As examples Lacoste-Julien, Huszár and Ghahramani consider Gaussian process regression with a Gaussian likelihood and a squared error loss function and Gaussian process classification with a probit likelihood and asymmetric binary loss. For a practical implementation the authors propose a suboptimal expectation maximization algorithm minimizing a variational lower bound. The proposed loss-calibrated expectation maximization algorithm, however, requires several approximations and according to authors, is closer to the variational type KL-divergence $d_{\mathrm{KL}}\{q(\theta), p(\theta|D, M_*)\}$, which could explain why the experimental results are inferior to predictions from expectation propagation posterior approximation.

The predictive posterior approximation is different from many parametric posterior approximation approaches, such as Laplace approximation (Tierney and Kadane, 1986), variational bound (Jordan et al., 1999; Jaakkola, 2001) and expectation propagation (EP) (Minka, 2001), in the sense that the predictive posterior approximation is determined directly through the predictive properties of the actual belief model, and not by matching the posterior distributions of model parameters. As a side note, although EP directly matches posterior distributions, it has been shown experimentally to give good predictions (Nickisch and Rasmussen, 2008; Vanhatalo, Pietiläinen and Vehtari, 2010; Jylänki, Vanhatalo and Vehtari, 2011).

Parametric projections Goutis and Robert (1998) and Dupuis and Robert (1997, 2003) presented an alternative projection framework for submodel selection, which could also be generalized to non-nested models.

The full model M_* is parametrized by $\theta \in \Theta$ and the parameter space of a submodel M_k is a restricted subspace of Θ so that $\theta_k \in \Theta_k \subset \Theta$. For example, Dupuis and Robert (1997, 2003) propose the restriction $\beta_j = 0$ for a subset of the generalized linear model parameters. The parameter projection is defined so that a discrepancy measure d achieves the infimum

$$d\{p(\tilde{y}|\tilde{x}, \theta, M_*), p(\tilde{y}|\tilde{x}, \theta_k^{\perp}, M_k)\} = \inf_{\theta_k \in \Theta_k} d\{p(\tilde{y}|\tilde{x}, \theta, M_*), p(\tilde{y}|\tilde{x}, \theta_k, M_k)\} \quad (130)$$

and the projected model is $p(\tilde{y}|\tilde{x}, \theta_k^{\perp}, M_k)$, where $\theta_k^{\perp}$ denotes the projected parameters. Goutis and Robert (1998) and Dupuis and Robert (1997, 2003) use the Kullback-Leibler divergence d_{KL} (from the encompassing model to a submodel), as it has an information theoretic justification of measuring the amount of information lost by using the simpler model. Additional benefit of KL-divergence is quick computation for generalized linear models, that is, projection equations having a form of likelihood equations associated with generalized linear models.

Unlike Goutis and Robert (1998), Dupuis and Robert (1997, 2003) stated explicitly that the divergence was between the joint models $p(\tilde{y}, \tilde{x}|\theta, \varphi, M)$ and $p(\tilde{y}, \tilde{x}|\theta_k, \varphi_k, M_k)$. Assuming $p(x|\varphi, M_*) = p(x|\varphi_k, M_k)$ they showed that the

variable selection procedure depended only on the expectation of the divergence between the conditional models $\mathbb{E}_{\tilde{x}}\, d\{p(\tilde{y}|\tilde{x},\theta,M_*), p(\tilde{y}|\tilde{x},\theta_k^\perp,M_k)\}$. They approximated the expectation over $\tilde{x}$ in an $\mathcal{M}$-open way as

$$\frac{1}{n}\sum_{i=1}^{n} d\{p(\tilde{y}|\dot{x}_i,\theta,M_*), p(\tilde{y}|\dot{x}_i,\theta_k^\perp,M_k)\}. \tag{131}$$

Goutis and Robert (1998) and Dupuis and Robert (1997, 2003) determined the criterion for variable selection as the posterior expectation of Eq. (131),

$$\frac{1}{n}\sum_{i=1}^{n}\int d\{p(\tilde{y}|\dot{x}_i,\theta,M_*), p(\tilde{y}|\dot{x}_i,\theta_k^\perp,M_k)\}p(\theta|D,M_*)d\theta. \tag{132}$$

The posterior expectation over θ is typically approximated by MCMC. The projection can be made for each MCMC sample separately, as the projection from θ to $\theta_k^\perp$ separates in a pointwise fashion.

The above projection approach is based on the posterior expectation of the KL-divergence between the models, which corresponds to using Gibbs utility instead of logarithmic utility in prediction calibrated posterior approximation $q(\theta_k)$ presented in Section 3.3.2). The different order of integration and logarithm in Gibbs loss makes the projection computations easier.

Goutis and Robert (1998) and Dupuis and Robert (1997, 2003) focused on model selection and do not consider projected predictive distributions. Projected predictive distributions can be computed using the projected posterior distribution of $\theta^\perp$, which is done, for example, by Nott and Leng (2010). Goutis and Robert (1998) and Dupuis and Robert (1997, 2003) used submodels with equality constraints while Nott and Leng (2010) proposed inequality constraints, for example, of the form $\sum_{j=1}^{p}|\beta_j| \leq \lambda$, which produces a lasso-type procedure with the value of a continuous parameter λ under selection.

The Gibbs loss projection approach is equivalent to a Bayesian hypothesis testing by Bernardo (1999) (and related to earlier methods Bernardo and Bayarri, 1985; Bayarri, 1987; Gutiérrez-Peña, 1992; Rueda, 1992). Bernardo considers only unconditional models $p(y|\theta)$, predicts replicate data of the same size, and the use of reference priors (Berger and Bernardo, 1992) is an essential part of the procedure.

Bernardo and Rueda (2002) switched to use a symmetric divergence they call *intrinsic divergence*, that is the minimum of directed KL-divergences

$$\min\left\{d_{\mathrm{KL}}\left\{p(y|\theta,M_*), p(y|\theta_k,M_k)\right\}, d_{\mathrm{KL}}\left\{p(y|\theta_k,M_k), p(y|\theta,M_*)\right\}\right\}, \tag{133}$$

to be able to handle cases where $p(y|\theta,M_*)$ and $p(y|\theta_k,M_k)$ may have different supports. Although allowing for a more general automatic procedure, this change makes the criterion honor less the predictive properties.

Bernardo and Juárez (2003) use symmetric KL-divergence for intrinsic estimation. In this case, point estimates for some continuous parameters are chosen so that the criterion is minimized while other parameters are projected.

If a point estimate for all the parameters of the reference model were used and no parameters would be projected (and given the same divergence) then logarithmic utility and Gibbs utility would produce the same results. If undirected KL-divergence from the reference model were used instead of the symmetric KL-divergence, the approach by Bernardo and Juárez (2003) would be equivalent to selecting a point estimate minimizing the reference predictive criterion, that is

$$\iint p(\tilde{y}|\theta, M_*) \log \frac{p(\tilde{y}|\theta, M_*)}{p(\tilde{y}|\hat{\theta}, M_k)} d\tilde{y} p(\theta|D, M_*) d\theta \tag{134}$$

$$= C - \int p(\tilde{y}|D, M_*) \log p(\tilde{y}|\hat{\theta}, M_k) d\tilde{y}, \tag{135}$$

where $\hat{\theta}$ is the point estimate minimizing the criterion (compare to Eq (132)). A natural extension of the point estimates are credible region estimates using the highest utility regions. Bernardo (2005b) proposed to use the symmetric KL-divergence as the utility in this case, too. The directed KL-divergence could be used in similar way.

5.5. Information criteria

Information criteria are commonly used for selecting Bayesian models. Many information criteria are directly related to assessing the predictive performance of the candidate models. The logarithmic utility function or the deviance loss function for n independent replicate observations are often used; for consistency we continue to write the formulas with the logarithmic utility function. For a better understanding we also review the influential frequentist information criteria.

Regardless on the background theory, information criteria are typically of form

$$\bar{u}_{\mathrm{IC}}(M_k|D) = \frac{1}{n} \sum_{i=1}^{n} u(M_k, \hat{a}_k, y_i) + \text{bias correction}. \tag{136}$$

where the first part is the training utility, Eq. (82), and the bias correction is computed either using the model M_k itself or a reference model M_*. From the training utility part it can be seen that the variance of the information criteria under repeated sampling of training data is similar to the full $\mathcal{M}$-open approaches (Section 5.1). Bias corrections in information criteria are derived considering $p(y|M_k)$. For conditional models $p(y|x, M_k)$ information criteria typically consist of a sum of terms conditional on x_i. Thus they do not evaluate the out-of-sample performance outside the observed $x_{(1:n)}$. This is sufficient for a fixed x, but for random x and deterministic $\tilde{x}$ outside the observed values the estimates may be optimistic.

AIC Akaike (1974) developed the information criterion AIC to be an information theoretic generalization of the expected prediction error for models $p(y|\hat{\theta}_k, M_k)$, where $\hat{\theta}_k$ is the maximum likelihood estimate of the parameters. AIC estimates the expected log score given maximum likelihood estimate $\hat{\theta}$ (e.g. Burnham and Anderson, 2002, p. 364)

$$\text{AIC} \approx E_{D_n}\left[\int p_t(\tilde{y}) \log p(\tilde{y}|\hat{\theta}_k(D_n), M_k) d\tilde{y}\right], \tag{137}$$

where the expectation is taken over all the possible training sets D_n from $p_t(\cdot)$ (see Section 3.5.2). Equation (137) can be written in the form

$$\text{AIC} \approx E_{\hat{\theta}_k|p_t}\left[\int p_t(\tilde{y}) \log p(\tilde{y}|\hat{\theta}_k, M_k) d\tilde{y}\right], \tag{138}$$

where the expectation with regard to $\hat{\theta}_k$ is over the sampling distribution of the estimator $\hat{\theta}_k$. For conditional models the following form is used

$$\text{AIC} \approx E_{\hat{\theta}_k|p_t}\left[\frac{1}{n}\sum_{1=1}^{n}\int p_t(\tilde{y}_i|x_i) \log p(\tilde{y}_i|x_i, \hat{\theta}_k, M_k) d\tilde{y}_i\right]. \tag{139}$$

Under the assumption that the true data-generating model $p_t(y)$ can be approximated well by the pseudo-true model $p(y|\theta_0, M_k)$ one can write a Taylor series expansion for $\hat{\theta}_k$ to get the following AIC estimate (see e.g. Burnham and Anderson, 2002, for a clearly presented derivation)

$$\text{AIC} = \frac{1}{n}\sum_{i=1}^{n} \log p(\dot{y}_i|\dot{x}_i, \hat{\theta}_k(D), M_k) - \frac{p}{n}, \tag{140}$$

where p is the number of parameters in the model. The first term on the right hand side is the maximum likelihood predictive estimate (cf. posterior predictive estimate) evaluated at the observed data. The estimate is optimistic due to using the data twice, and p/n is an asymptotic estimate of this optimism.

Small sample corrections to the AIC have been discussed for example by Hurvich and Tsai (1989, 1991) and Burnham and Anderson (1998). A "Bayesian extension" of AIC by Akaike (1979) adds a prior on the model space and an optional averaging of models, but no integration over the parameter space.

TIC, RIC, NIC Takeuchi (1976) (since the original paper is in Japanese, see also, e.g., Shibata (1989); Burnham and Anderson (1998)) provided a more general derivation giving a better estimate for the expected prediction error if the candidate models are not particularly close to the true model $p_t(\tilde{y})$. The TIC criterion is defined as

$$\text{TIC} = \frac{1}{n}\sum_{i=1}^{n} \log p(\dot{y}_i|\dot{x}_i, \hat{\theta}_k(D), M_k) - \frac{1}{n}\,\text{tr}\left[\hat{J}(\hat{\theta})\hat{I}^{-1}(\hat{\theta})\right], \tag{141}$$

where $\hat{J}(\hat{\theta}) = \text{Cov}[l'_{\hat{\theta}}]$, $\hat{I}(\hat{\theta}) = \mathbb{E}[l''_{\hat{\theta}}]$, and $l'_{\hat{\theta}}$ and $l''_{\hat{\theta}}$ are the first and second derivatives of the likelihood with respect to θ evaluated at $\hat{\theta}$. The trace term $\text{tr}\big[\hat{J}(\hat{\theta})\hat{I}^{-1}(\hat{\theta})\big] \leq p$ was called by Moody (1992) the *effective number of parameters*. Shibata (1989) proposed another criterion called regularized information criterion (RIC), which extended TIC to penalized likelihoods. Later, Murata, Yoshizawa and Amari (1994) proposed yet another criterion called network information criterion (NIC) which extends TIC to arbitrary differentiable utility functions and penalized or regularized likelihoods (i.e., maximum a posteriori approach in Bayesian terms). TIC/RIC/NIC approximations are seldom used in practice as the variance of the estimate is increased due to stability problems and computational difficulties in estimating the unknown $p \times p$ matrices J and I. Often a simpler AIC approximation with a smaller variance gives a better estimate (Shibata, 1989; Burnham and Anderson, 2002).When the assumptions of AIC hold, AIC and TIC are asymptotically equal.

Stone (1977) considered the asymptotic behavior of cross-validation with maximum-likelihood plug-in estimate. Using a first order Taylor approximation, Stone heuristically showed that the LOO-CV (with maximum likelihood estimate $\hat{\theta}$) is asymptotically equivalent with TIC. See also Shibata (1989) for a derivation of the asymptotic equivalence of TIC and cross-validation.

Sawa (1978), Chow (1981) & Young (1987b) Theoretically expectations in Eq. (137) are over $p_t(\cdot)$, but the unknown true data distribution $p_t(\cdot)$ is typically approximated by a *pseudo-true model* $p(\cdot|\theta_0, M_k)$ and θ_0 is approximated with $\hat{\theta}_k(D)$. Thus $p_t(\tilde{y})$ is partially approximated by reusing $y_{(1:n)}$ and partially by $p(\tilde{y}|\hat{\theta}_k, M_k)$. TIC corrects to some extent the problem arising from assuming that the pseudo-true model $p(\tilde{y}|\theta_0, M_k)$ is close to an unknown $p_0(\tilde{y})$.

Sawa (1978) proposes a different modification of AIC, in which some properties of the true $p_0(y)$ are known, and the pseudo-true distribution $p(y|\theta_0, M_*)$ is obtained by projecting the true data-generating distribution into a restricted class of models. Sawa considers only a case where $p(y|\theta_0, M_*)$ is essentially assumed to be a multivariate Gaussian with a diagonal covariance matrix and the candidate models are Gaussian linear regression models. Sawa proposes a criterion (unfortunately named as B information criterion (BIC), causing confusion with the Bayesian information criterion (BIC)), which requires estimates for the unknown parameters of both $p(y|\theta_0, M_*)$ and the candidate models $p(y|\theta, M_k)$. When the candidate models are reasonably close to $p(y|\theta, M_*)$, Sawa's criterion reduces to AIC. Similarly as in AIC, Sawa's criterion is evaluated re-using the observed data D. Sawa also presented a pseudo-Bayesian criterion where a prior is placed on the pseudo-true parameters, and optimal estimators for the parameters of every candidate model are found by maximizing the utility.

Chow (1981) developed Sawa's approach further into a TIC type estimate (calling it *an information criterion* as Akaike (1974) had called his criterion). Chow also used a reference model $p(\tilde{y}|\hat{\theta}, M_*)$ (which he called the most general model) with a plug-in estimate $\hat{\theta}$.

Young (1987b) separated parameter estimation and model selection by removing the dependency on the parameter estimates through introducing priors to both the reference model parameters and the candidate model parameters. Young, using a multivariate Gaussian as the reference model $p(\tilde{y}|\theta_*, M_*)$ in a linear regression setting, proposed a criterion obtained by taking posterior expectations of the utility $\int p(\tilde{y}|\theta_*, M_*) \log p(\tilde{y}|\theta_k, M_k) d\tilde{y}$ over both θ_* and θ_k.

The use of plug-in estimates in Sawa (1978) and Chow (1981) for the parameters of the reference model $p(\tilde{y}|\hat{\theta}, M_*)$ was criticized by Leamer (1979), who pointed out that the problem should be formulated by placing a prior on the parameters of the reference model, and the parameters of the candidate models should be selected by maximizing the expected utility.

AIC/TIC/RIC/NIC criteria (and DIC/BPIC/WAIC criteria below) can be said to estimate the bias correction in a self-referential way while the criteria proposed by Sawa (1978), Chow (1981) and Young (1987b) can be said to estimate the bias correction in a reference way.

DIC Deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) estimates the plug-in generalization utility using plug-in predictive distribution (original DIC is multiplied by $-2n$)

$$\int p(\theta_k|D, M_k) \int p(\tilde{y}|\theta_k, M_k) \log p(\tilde{y}|\bar{\theta}_k, M_k) d\tilde{y} d\theta_k. \tag{142}$$

DIC has a generic form

$$\mathrm{DIC} = \frac{1}{n} \sum_{i=1}^{n} \log p(\dot{y}_i|\dot{x}_i, \hat{\theta}_k, M_k) - \frac{p_{\mathrm{eff}}}{n}, \tag{143}$$

where the effective number of parameters p_{eff} can be estimated in two ways, which both can be derived from the properties of the distribution of the deviance (Gelman et al., 2003; Raftery et al., 2007). For regular models the distribution of the deviance approaches shifted χ^2_ν, where the degrees of freedom ν can be interpreted as the effective number of parameters. The shifted χ^2_ν has following properties: $\mathbb{E}(\cdot) = \mathrm{offset} + \nu$ and $\mathrm{Var}(\cdot) = 2\nu$. Using the mean approach, p_{eff} can be estimated as

$$p_{\mathrm{eff}} \approx 2 \sum_{i=1}^{n} \left[\log p(\dot{y}_i|E_{\theta_k|D,M_k}[\theta], M_k) - E_{\theta_k|D,M_k} \log p(\dot{y}_i|\theta_k, M_k) \right], \tag{144}$$

where the plug-in deviance is used to estimate the offset. This estimate is not generally invariant to reparametrization because in general plug-in estimates are not invariant to reparametrization. Using the variance approach p_{eff} can be estimated as

$$p_{\mathrm{eff}} \approx \mathrm{Var}_{\theta_k|D,M_k}[\log p(\dot{y}_i|\theta_k, M_k)]. \tag{145}$$

This form avoids the use of plug-in estimate in estimation of p_{eff}, but this variance estimate can be unstable due to the long tail of the χ^2_ν distribution (Raftery et al., 2007; Carlin and Spiegelhalter, 2007).

DIC type bias correction b can be used, also, for other utilities (Vehtari, 2002)

$$b \approx \frac{2}{n}\sum_{i=1}^{n}\left[u(M_k, E_{\theta_k|D,M_k}[\theta], \dot{y}_i) - E_{\theta_k|D,M_k} u(M_k, \theta_k, \dot{y}_i)\right]. \tag{146}$$

Use of plug-in estimates and the assumption that deviance is χ^2_ν distributed in AIC/TIC/RIC/NIC/BPIC is problematic for irregular models such as mixture models. Richardson (2002) in her comment on DIC (Spiegelhalter et al., 2002), proposed without a formal justification a DIC version using the predictive distributions,

$$p_{\text{eff}} \approx 2\sum_{i=1}^{n}\left[\log E_{\theta_k|D,M_k}[p(\dot{y}_i|\theta, M_k)] - E_{\theta_k|D,M_k}\log p(\dot{y}_i|\theta_k, M_k)\right], \tag{147}$$

which avoids the use of plug-in estimate. This version with a name DIC_3 was tested in a mixture-modeling context by Celeux et al. (2006) along with seven other ad-hoc versions of DIC, without proper formal justification for any of them. Celeux et al. (2006) noted that the predictive version was stable, but based on the experiments did not think it was the best one. Celeux et al. (2006) did not compare directly the bias or the variance of the different proposals, but tested them in a model selection with Galaxy data and one simulated data, and thus it is not easy to interpret the results with regard to ability to estimate the actual predictive performance as discussed by Cawley and Talbot (2010).

BPIC The Bayesian predictive information criterion (BPIC) by Ando (2007) estimates the expected Gibbs utility (Section 3.4.2)

$$\bar{u}_t^G(M_k) = \int p_t(\tilde{y}) \int p(\theta_k|D, M_k)\log p(\tilde{y}|\theta_k, M_k)d\theta_k d\tilde{y}, \tag{148}$$

in which the parameters θ_k are integrated over the posterior distribution instead of using a plug-in estimate as in DIC. Ando proposes an estimate based on replacing the expectation over the unknown future data distribution by an average over observed data. Using a similar derivation as in TIC, Ando proposes to estimate the resulting bias by

$$\begin{aligned} n\hat{b} = &\sum_{i=1}^{n}\int p(\theta|D, M_k)\log\left[p(y_i|\theta, M_k)p(\theta|M_k)\right]d\theta \\ &- \sum_{i=1}^{n}\log\left[p(y_i|\hat{\theta}, M_k)p(\hat{\theta}|M_k)\right] + \operatorname{tr}\left\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\right\} + \frac{p}{2}, \end{aligned} \tag{149}$$

where $\hat{\theta}$ is the MAP estimate. The BPIC criterion (divided by $-2n$) is given by

$$\text{BPIC}(M_k) = \frac{1}{n}\sum_{i=1}^{n}\int p(\theta_k|D, M_k)\log p(y_i|\theta_k, M_k)d\theta_k - \hat{b}. \tag{150}$$

The BPIC criterion has the same problems as TIC/NIC due to instability and computational difficulties in estimating $\hat{I}$ and $\hat{J}$.

WAIC Watanabe (2009, 2010b,c,a) presented a criterion which he called *widely applicable information criterion* (WAIC) and gave a formal proof of its properties as an estimate for the expected utility

$$\bar{u}_t(M_k) = \int p_t(\tilde{y}) \log p(\tilde{y}|D, M_k) d\tilde{y} \tag{151}$$

for both regular and singular models. A criterion of a similar form was independently proposed by Richardson (2002) as a version of DIC (see DIC section above). Other information criteria are based on Fisher's asymptotic theory assuming a regular model for which the likelihood or the posterior converges to a single point and MLE, MAP, and plug-in estimates are asymptotically equivalent. With singular models the true model is projected onto a set of parameters consisting of more than one point, the Fisher information matrix is not positive definite, plug-in estimates are not representative of the posterior and the distribution of the deviance does not converge to a χ^2_ν distribution. Watanabe uses singular learning theory (Watanabe, 2009) to derive more general results which also hold for singular models. Watanabe uses loss functions but for consistency in this review we use the equivalent utility functions.

Watanabe shows that the Bayesian generalization utility can be estimated by a criterion

$$\mathrm{WAIC}_G = \bar{u}_{\mathrm{train}} - 2(\bar{u}_{\mathrm{train}} - \bar{u}^G_{\mathrm{train}}), \tag{152}$$

where $\bar{u}_{\mathrm{train}}$ is Bayes training utility and $\bar{u}^G_{\mathrm{train}}$ is Gibbs training utility defined in Figure 11. The estimate is asymptotically equal to the true logarithmic utility in single prediction in both regular and singular statistical models and the error in a finite case is $o(1/n)$. Watanabe (2010a) shows also that the WAIC estimate is asymptotically equal to the Bayesian cross-validation estimate (section 5.1.3) and the Gibbs generalization utility can be estimated as

$$\mathrm{WAIC}_{GG} = \bar{u}^G_{\mathrm{train}} - 2(\bar{u}_{\mathrm{train}} - \bar{u}^G_{\mathrm{train}}) \tag{153}$$

and the error of this estimate is again $o(1/n)$.

WAIC can also be given as a functional variance form

$$\mathrm{WAIC}_V = \bar{u}_{\mathrm{train}} - V/n, \tag{154}$$

where the functional variance

$$V = \sum_{i=1}^{n} \Bigg\{ \mathbb{E}_{\theta|D,M_k} \left[\left(\log p(\dot{y}_i|\dot{x}_i, \theta, M_k) \right)^2 \right] - \left(\mathbb{E}_{\theta|D,M_k} \left[\log p(\dot{y}_i|\dot{x}_i, \theta, M_k) \right] \right)^2 \Bigg\}, \tag{155}$$

describes the fluctuation of the posterior distribution. WAIC_G and WAIC_V are asymptotically equal, but the series expansion of WAIC_V has closer resemblance to the series expansion of the logarithmic leave-one-out utility. Watanabe

(2010b) derives the error of the asymptotic theory for WAIC as

$$|\bar{u}_{\text{train}} - \bar{u}^G_{\text{train}} - V/(2n)| = |\text{WAIC}_V - \text{WAIC}_G|/2. \tag{156}$$

To better see the connection between DIC and WAIC, DIC can be written as

$$\text{DIC} = \bar{u}^P_{\text{train}} - 2(\bar{u}^P_{\text{train}} - \bar{u}^G_{\text{train}}), \tag{157}$$

where $\bar{u}^P_{\text{train}}$ is the plug-in training utility defined in Figure 11. The similarity to WAIC_G is obvious. However, the plug-in estimate $\bar{\theta}$ is sensible only for regular models. Moreover, the plug-in predictions differ from the posterior predictive predictions. Watanabe (2010a) shows that in a regular and realizable case a modification of DIC,

$$\text{DIC}* = \bar{u}_{\text{train}} - 2(\bar{u}^P_{\text{train}} - \bar{u}^G_{\text{train}}), \tag{158}$$

is asymptotically equivalent to WAIC. The asymptotical equivalence does not hold in the unrealizable and singular cases.

In a similar fashion to NIC the DIC/WAIC approach can be formulated for arbitrary differentiable utility functions with appropriate properties. As an illustrative example consider WAIC with a squared error loss function,

$$\text{WAIC}_S = \bar{s}_{\text{train}} - 2(\bar{s}_{\text{train}} - \bar{s}^G_{\text{train}}), \tag{159}$$

where

$$\bar{s}_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n} (\dot{y}_i - \mathbb{E}[\tilde{y}|\dot{x}_i, D, M_k])^2 \tag{160}$$

$$\bar{s}^G_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta_k}\left[(\dot{y}_i - \mathbb{E}[\tilde{y}|\dot{x}_i, \theta_k, M_k])^2 \middle| D, M_k\right]. \tag{161}$$

Rearranging the terms gives

$$\text{WAIC}_S = \bar{s}_{\text{train}} + \frac{1}{n}\sum_{i=1}^{n} \text{Var}_{\theta_k}\left[\mathbb{E}[\tilde{y}|\dot{x}_i, \theta_k, M_k]|D, M_k\right], \tag{162}$$

where the bias correction is the posterior variance of the mean prediction. This can be compared to the variance of the predictive distribution (Equation (15)) which combines the expected variance of the observations and the variance from the posterior uncertainty.

Ando and Tsay (2010): simultaneous prediction All the previous information criteria are formulated to estimate the single prediction utility (possibly scaled with n). Ando and Tsay (2010) propose a criterion for estimating the simultaneous generalization utility

$$\bar{u}_t(M_k) = \int p_t(\tilde{y}_{(1:n)}) \log p(\tilde{y}_{(1:n)}|D, M_k) d\tilde{y}_{(1:n)}, \tag{163}$$

with a TIC type criterion they refer to as the *predictive likelihood score*

$$\mathrm{PL}(M_k) = \frac{1}{n}\log p(y_{(1:n)}|D, M_k) - \frac{1}{2n}\,\mathrm{tr}\left[\hat{J}^{-1}\{\hat{\theta}\}\hat{I}\{\hat{\theta}\}\right]. \tag{164}$$

This criterion possesses the same problems as TIC/NIC due to instability and computational difficulties in estimation of $\hat{I}$ and $\hat{J}$. Ando and Tsay (2010) propose also a simplification

$$\mathrm{PL2}(M_k) = \frac{1}{n}\log p(y_{(1:n)}|D, M_k) - \frac{p}{2n} \tag{165}$$

which resembles AIC.

Other information criteria are asymptotically equivalent to different forms of leave-one-out cross-validation, but it is not clear what relationship the criterion of Ando and Tsay (2010) has to cross-validation, since it is not possible to leave n data points out in the cross-validation approach.

5.6. Prior predictive distribution and Bayes factor

A fairly common approach to model selection in Bayesian statistics is based on maximizing the joint prior predictive distribution

$$p(D|M) = \int p(D|\theta, M)p(\theta|M)d\theta \tag{166}$$

with respect to the model M. The quantity $p(D|M)$ is called the marginal likelihood or the evidence of a model M.

If one makes the explicit assumption that one of the candidate models is true, the optimal model choice under the the zero-one utility/loss (see section 3.4.1) is the model with the highest posterior probability $p(M|D)$. In a pairwise comparison the *posterior odds* for two models M_0 and M_1 can be written as

$$\frac{p(M_0|D)}{p(M_1|D)} = \frac{p(D|M_0)}{p(D|M_1)}\frac{p(M_0)}{p(M_1)}, \tag{167}$$

that is, as a product of the prior odds and the *Bayes factor*

$$B_{01} = \frac{p(D|M_0)}{p(D|M_1)}. \tag{168}$$

Given equal prior probabilities for the models, the posterior probabilities $p(M|D)$ are proportional to the marginal likelihoods $p(D|M)$, and the comparison of models in a pairwise fashion can be based on the Bayes factor (see, e.g., Kass and Raftery, 1995, for review)).

A predictive perspective can be taken by writing the marginal likelihood by the chain rule

$$p(D|M_k) = p(y_1|M_k)p(y_2|y_1, M_k), \ldots, p(y_n|y_{(1:n-1)}, M_k), \tag{169}$$

which shows that the joint prior predictive approach corresponds to making the first prediction with the prior and then sequentially predicting the remaining data by updating the predictions one data point at time. With the logarithmic utility function the utility of the joint prior prediction corresponds to the average utility of posterior predictive distributions with the number of data points used for training ranging from 0 to $n-1$. In other words, the utility of prior predictive approach can be considered to measure the expected predictive performance of a model. As the learning curve is usually steeper with a small number of data points, the average is less than the logarithmic utility resulting from a posterior predictive distribution with $n/2$ training data points. As the first terms in the chain rule are conditioned on none or very few data points, the prior predictive approach can be sensitive to prior definitions (see discussion, e.g., in O'Hagan and Forster, 2004, Ch 7.17). Furthermore, the first term of the series shows that the marginal likelihood does not exist if the prior is improper, even if the said improper prior leads to a proper posterior. Especially with vague priors and flexible models the few first terms dominate the expectation unless n is very large. Only if $n \gg p_{\text{eff}}$, terms conditioned on a large number of data points start to dominate and the result gets closer to the expected predictive utility.

In a general case, computation the marginal likelihood is far from trivial. A large number of different methods have been proposed, such as the Laplace approximation, harmonic mean estimator, annealed importance sampling, nested sampling, path sampling, bridge sampling, parallel tempering with thermodynamic integration and reversible jump MCMC (last one just for Bayes factor or posterior odds) reviewed, for example, in Kass and Raftery (1995); Han and Carlin (2000); Chen, Shao and Ibrahim (2000); Robert and Wraith (2009); Marin and Robert (2010); Friel and Wyse (2012).

There are several modifications of Bayes factors reducing the effect of improper or vague priors by using part of the data (e.g. O'Hagan, 1995; Berger and Pericchi, 1996). These methods can be considered as hold-out predictive approaches with small amounts of training data and joint prediction mentioned in Section 5.1.2.

From the expected predictive performance viewpoint the marginal likelihood $p(D|M)$ can be used as an indicative estimate of the predictive performance. For example, in the $\mathcal{M}$-closed Bayesian model averaging approaches a negligible posterior probability of a model lets us ignore the model also in model selection. Sensitivity to the prior definition affects also which models have a high posterior probability in the model averaging. However, it is possible that all the models with a considerable posterior probability under any sensible prior have similar predictions, so that the predictions are not so sensitive after integrating over the model space.

5.7. Model comparison approaches

Comparison of the expected utilities has two aspects: 1) what is the uncertainty related to the comparison and 2) what is a practically significant difference between the utility values.

5.7.1. Uncertainty related to the comparison

Sensitivity with regard to $\mathcal{M}$-open sampling error In $\mathcal{M}$-open type approaches the sampling error of sample re-use can be substantial, and needs to be taken into account. As discussed in Section 4.5 the related uncertainty in paired comparison can be estimated, for example, by computing the variance or using Bayesian bootstrap. These uncertainties tell whether in pairwise comparison we can be confident that there is a difference between the expected utilities (but they don't tell whether the difference is practically significant). In model selection, the smallest model not statistically worse than the best model could be chosen.

Sensitivity with regard to replication A group of methods answer the question "what is the sensitivity of the statistic" in a situation in which we have seen some other realization of the data and the significance of the expected utility difference is calibrated using the sensitivity estimate.

Ibrahim and Laud (1994) and Laud and Ibrahim (1995) proposed a *calibration number* for the L-criterion by estimating the standard deviation given possible datasets seen, by replicating datasets of size n from the marginal prior predictive distribution of the model giving the best L-criterion value. If the marginal prior predictive distribution is improper in linear regression, Laud and Ibrahim (1995) fix the error variance in the model to the MAP value. Ibrahim, Chen and Sinha (2001) extended the calibration approach by plotting the whole calibration distribution. Meyer and Laud (2002) presented a calibration of L-criterion for generalized linear models. Vlachos and Gelfand (2003) estimate calibration by replicating datasets of size n from the posterior predictive distribution of the model M. In case of pairwise model selection, they replicate datasets from both models and plot the two different summary statistic distributions obtained. They present their method only for pairwise model comparison, but a similar posterior replicate calibration approach could be used when estimating the expected utility for a single model.

Based on replicate dataset (bootstrap) methods by Ibrahim, Chen and Sinha (2001) and Vlachos and Gelfand (2003), it is also possible to compute the probability for one model being better than the other, but this probability is related to hypothetical data we might have seen and not to future data. In addition Vlachos and Gelfand (2003) use both models for replication, giving two different probabilities.

These methods use a model for the data in a parametric bootstrap way (e.g. Efron and Tibshirani, 1993). They are also related to prior predictive checking (Box, 1980) and posterior predictive checking (Gelman et al., 1995; Gelman, Meng and Stern, 1996). Non-parametric bootstrap by sample re-use could also be used, although it would require smaller training sets and thereby increase the bias of the estimates.

TABLE 1
Significance of posterior odds

1 : 1 to 3 : 1	Barely worth mentioning
3 : 1 to 10 : 1	Substantial
10 : 1 to 30 : 1	Strong
30 : 1 to 100 : 1	Very strong
$\geq$ 100 : 1	Decisive

5.7.2. Model comparison based on calibration of criteria

Several calibration approaches have been proposed for characterizing the practically significant difference in expected utilities.

Calibration scale for pairwise comparisons in terms of posterior odds, Eq. (167), was proposed by Jeffreys (1961). Jeffreys' calibration scale, presented in Table 1, relates the posterior odds to words describing significance. Similar scales with small variations have been proposed, some of which follow from presenting rounded values after taking a logarithm of the odds.

Bayes factor compares the prior joint predictive densities for n observations. Comparing the product of n single prediction densities from the cross-validation has a similar form as the Bayes factor and has been called quasi or pseudo Bayes factor (Geisser and Eddy, 1979; Gelfand, Dey and Chang, 1992), and a similar scale as for Bayes factor might be used. For information criteria there are heuristic scales for differences in log-scale (Burnham and Anderson, 2002). As information criteria and cross-validation are asymptotically equivalent, the same scales could be used for both. For DIC the proposed scale[4] is close to Jeffreys' scale. As cross-validation and information criteria have variance $o(1/n)$, it may be helpful to compare expected log-score times n, which has variance $o(1)$ and thus calibration with respect to variance does not depend on n.

Calibration of Kullback-Leibler divergence refers to introducing a yardstick in a form of Kullback-Leibler divergence between known statistical models. For example, in McCulloch (1989) the comparison between the reference model to a candidate model is calibrated by equating the KL-divergence between them to the KL-divergence from Bin(0.5) to Bin(θ), where θ is determined by setting the KL-divergences equal. In other words, comparing the reference model and the candidate model is set to be equivalent to comparing a binomial model with the true parameter value 0.5 to a Binomial model with parameter value θ. If using θ is deemed acceptable, the candidate model can be considered to be a reasonable proxy for the reference model. Use of other one-parameter exponential family models, such as Poisson and unit-variance Gaussian, was proposed by Goutis and Robert (1998). A calibration based on a Gaussian model with known σ was presented in Bernardo (1999), with an additional discussion on the correspondence between the threshold value and type I error probabilities. Bernardo and Rueda (2002) also discuss calibration for hypothesis testing.

[4] http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml

In the context of variable selection, Dupuis and Robert (1997, 2003) proposed to use the null and the full model to scale the KL-divergence by forming an explanatory power scale from 0% to 100%. The scale can be understood as an expression of the proportion of the information, when including all the covariates, which is ignored when using a smaller subset of covariates. Vehtari and Lampinen (2004) used this calibration approach to stop basic forward search in explanatory variable selection when 99% predictive power was obtained.

6. Conclusion

This qualitative review is an approach to presenting different proposed methods in a unified theoretical framework and notation, and hopefully it provides a useful map in the wild jungle of "Bayesian predictive criteria".

Surprisingly, although many comparisons between model selection approaches have been made, a systematic analysis of typical bias and variance properties of different Bayesian methods for finite sample size appears still to be lacking. The considerable variation in published results of performance rankings may be explained to some extent by the selection induced and learning curve bias. Thus, we have not aimed at an explicit ranking of the methods, as we believe that many model selection method comparisons in the literature may be misleading.

If the goal is to estimate the predictive performance of a Bayesian model then one should compute the expected utility, Eq. (5), which by construction is estimated by Bayesian cross-validation and WAIC. Both methods are known to be unbiased and asymptotically true estimates of the generalization utility. Reference and self-predictive approaches give an expected utility estimate with reduced variance, but have an unknown bias dependent on the reference model.

Even though CV and WAIC are asymptotically unbiased for a single model, overfitting to the observed data takes place during the model selection process, which in turn induces selection bias. The selection induced bias can be negligible if there are only a small number of models to be compared. The reference predictive and projection methods avoid using the data several times in model selection, which should result in reduced selection induced bias.

Although a qualitative comparison can provide advice, we think that there is still a strong need for a quantitative analysis comparing the performance of different methods.

Acknowledgements

The authors would like to thank Elja Arjas, Andrew Gelman, Tommi Mononen, Mari Myllymäki, and the entire BECS-Bayes group for the valuable discussions and comments on the manuscript. This work was supported financially by the Academy of Finland (grant 218248).

References

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

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki, eds.) 267–281. Academiai Kiado, Budapest. Reprinted in Kotz, S. and Johnson, N. L., editors, (1992). *Breakthroughs in Statistics Volume I: Foundations and Basic Theory*, pp. 610–624. Springer-Verlag. MR0483125

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **AC-19** 716–723. MR0423716

Akaike, H. (1979). A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting. *Biometrika* **66** 237–242. MR0548189

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika* **94** 443–458. MR2380571

Ando, T. and Tsay, R. (2010). Predictive likelihood for Bayesian model selection and averaging. *International Journal of Forecasting* **26** 744–763.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys* **4** 40–79. MR2602303

Barbieri, M. M. and Berger, J. O. (2004). Optimal Predictive Model Selection. *The Annals of Statistics* **32** 870–897. MR2065192

Bayarri, M. J. (1987). Comment to J. O. Berger and M. Delampady. *Statistical Science* **3** 342–344.

Bayarri, M. J. (2003). Which 'base' distribution for model criticism? In *Highly Structured Stochastic Systems* (P. J. Green, N. L. Hjort and S. Richardson, eds.) 445–453. Oxford University Press. MR2082403

Bayarri, M. J. and Berger, J. O. (1999). Quantifying Surprise in the Data and Model Verification. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger and A. P. Dawid, eds.) 53–82. Oxford University Press. MR1723493

Bayarri, M. J. and Berger, J. O. (2000). P Values for Composite Null Models. *Journal of the American Statistical Association* **95** 1127–1142. MR1804239

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer-Verlag. MR0804611

Berger, J. O. and Bernardo, J. M. (1992). On the Development of Reference Priors. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 35–60. Oxford University Press. MR1380269

Berger, J. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91** 109–122. MR1394065

Bernardo, J. M. (1979). Expected Information as Expected Utility. *Annals of Statistics* **7** 686–690. MR0527503

Bernardo, J. M. (1999). Nested Hypothesis Testing: The Bayesian Reference Criterion. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger and A. P. Dawid, eds.) 101–130. Oxford University Press. MR1723495

Bernardo, J. M. (2005a). Reference Analysis. In *Handbook of Statistics*, (D. Dey and C. R. Rao, eds.) **25** Elsevier 17–90. MR2490522

Bernardo, J. M. (2005b). Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test* **14**. 317–384. MR2211385

Bernardo, J. M. and Bayarri, M. J. (1985). Bayesian model criticism. In *Model choice: proceedings of the 4th Franco-Belgian meeting of statisticians* (J. P. Florens, M. Mouchart, J. P. Raoult and L. Simar, eds.). Facultés universitaires Saint-Louis, Bruxelles.

Bernardo, J. M. and Bermúdez, J. D. (1985). The Choice of Variables in Probabilistic Classification. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. deGroot, D. V. Lindley and A. F. M. Smith, eds.) 67–82. Elsevier Science Publishers. MR0862484

Bernardo, J. M. and Juárez, M. A. (2003). Intrinsic Estimation. In *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 456–476. Oxford University Press. MR2003181

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: a reference approach. *International Statistical Review* **70** 351–372.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons. MR1274699

Bhattacharya, S. and Haslett, J. (2007). Importance Re-sampling MCMC for Cross-Validation in Inverse Problems. *Bayesian Analysis* **2** 385–408. MR2312288

Birgé, L. and Massart, P. (2007). Minimal Penalties for Gaussian Model Selection. *Probability Theory and Related Fields* **138** 33–73. MR2288064

Bornn, L., Doucet, A. and Gottardo, R. (2010). An efficient computational approach for prior sensitivity analysis and cross-validation. *The Canadian Journal of Statistics* **38** 47–64. MR2676929

Box, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society. Series A (General)* **143** 383–430. MR0603745

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Chapman and Hall. MR0726392

Brown, P. J., Fearn, T. and Vannucci, M. (1999). The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika* **86** 635–648. MR1723783

Brown, P. J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **60** 627–641. MR1626005

Brown, P. J., Vannucci, M. and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64** 519–536. MR1924304

Burman, P. (1989). A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika* **76** 503–514. MR1040644

Burman, P., Chow, E. and Nolan, D. (1994). A Cross-Validatory Method for Dependent Data. *Biometrika* **81** 351–358. MR1294896

Burman, P. and Nolan, D. (1992). Data dependent estimation of prediction functions. *Journal of Time Series Analysis* **13** 189–207. MR1168164

Burnham, K. P. and Anderson, D. R. (1998). *Model selection and inference*. Springer.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer. MR1919620

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis* **69**. Chapman & Hall. MR1427749

Carlin, B. P. and Spiegelhalter, D. J. (2007). Discussion to 'Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity'. In *Bayesian Statistics 8* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 33–36. Oxford University Press. MR2452343

Cawley, G. C. and Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* **11** 2079–2107. MR2678023

Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006). Deviance Information Criteria for Missing Data Models. *Bayesian Analysis* **1** 651–674. MR2282197

Chakrabarti, A. and Ghosh, J. K. (2007). Some Aspects of Bayesian Model Selection for Prediction. In *Bayesian Statistics 8* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 51–90. Oxford University Press. MR2433189

Chen, M.-H., Dey, D. K. and Ibrahim, J. G. (2004). Bayesian criterion based model assessment for categorical data. **91** 45–63. http://biomet.oxfordjournals.org/content/91/1/45.abstract MR2050459

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. Q. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag. MR1742311

Chow, G. C. (1981). A comparison of the information and posterior probability criteria for model selection. *Journal of Econometrics* **16** 21–33.

Corander, J. and Marttinen, P. (2006). Bayesian Model Learning Based on Predictive Entropy. *Journal of Logic, Language, and Information* **15** 5–20. http://www.jstor.org/stable/40180417 MR2254565

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation* **10** 1895–1924.

Draper, D. and Fouskakis, D. (2000). A Case Study of Stochastic Optimization in Health Policy: Problem Formulation and Preliminary Results. *Journal of Global Optimization* **18** 399–416.

Dupuis, J. A. and Robert, C. P. (1997). Bayesian Variable Selection in Qualitative Models by Kullback-Leibler Projections Working papers, Centre de Recherche en Economie et Statistique.

Dupuis, J. A. and Robert, C. P. (2003). Variable selection in qualitative models via an entropic explanatory power. *Journal of Statistical Planning and Inference* **111** 77–94. MR1955873

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap* **57**. Chapman & Hall. MR1270903

Epifani, I., MacEachern, S. N. and Peruggia, M. (2008). Case-deletion importance sampling estimators: Central limit theorems and related results. *Electronic Journal of Statistics* **2** 774–806. MR2443196

Fearn, T., Brown, P. J. and Besbeas, P. (2002). A Bayesian decision theory approach to variable selection for discrimination. *Statistics and Computing* **12** 253–260. MR1933511

Fouskakis, D. and Draper, D. (2008). Comparing stochastic optimization methods for variable selection in binary outcome prediction with application to health policy. *Journal of the American Statistical Association* **103** 1367–1381. MR2655719

Fouskakis, D., Ntzoufras, I. and Draper, D. (2009). Population-based reversible-jump Markov chain Monte Carlo for Bayesian variable selection and evaluation under cost limit restrictions. *Journal of the Royal Statistical Society, Series C: Applied Statistics* **58** 383–403. MR2750012

Friel, N. and Wyse, J. (2012). Estimating the evidence – a review. *Statistica Neerlandica*. Early view online. DOI: 10.1111/j.1467-9574.2011.00515.x. MR2955421

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing* **21** 137–146. MR2774847

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association* **70** 320–328.

Geisser, S. and Eddy, W. F. (1979). A Predictive Approach to Model Selection. *Journal of the American Statistical Association* **74** 153–160. MR0529531

Gelfand, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.) 145–162. Chapman & Hall. MR1397969

Gelfand, A. E. (2003). Some comments on model criticism. In *Highly Structured Stochastic Systems* (P. J. Green, N. L. Hjort and S. Richardson, eds.) 449–453. Oxford University Press. MR2082403

Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model Determination using Predictive Distributions with Implementation via Sampling-Based Methods (with discussion). In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 147–167. Oxford University Press. MR1380275

Gelfand, A. E. and Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society. Series B (Methodological)* **56** 501–514. MR1278223

Gelfand, A. E. and Ghosh, S. K. (1998). Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika* **85** 1–11. MR1627258

Gelman, A., Meng, X.-L. and Stern, H. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies (with discussion). *Statistica Sinica* **6** 733–807. MR1422404

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. R. (1995). *Bayesian Data Analysis*. Chapman & Hall. MR1385925

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. R. (2003). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall. MR2027492

George, E. I. and McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association* **88** 881–889.

Geweke, J. (1989). Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica* **57** 1317–1339. MR1035115

Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association* **106** 746–762. MR2847988

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **69** 243–268. MR2325275

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of American Statistical Association* **102** 359–378. MR2345548

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

Goutis, C. and Robert, C. P. (1998). Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. *Biometrika* **85** 29–37. MR1627250

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics* **32** 1367–1433. MR2089128

Gutiérrez-Peña, E. (1992). Expected logarithmic divergence for exponential families. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 669–674. Oxford University Press. MR1380301

Gutiérrez-Peña, E. (1997). A Bayesian Predictive Semiparametric Approach to Variable Selection and Model Comparison in Regression. In *Bulletin of the International Statistical Institute, Tome LVII. (Proceedings of the 51st Session of the ISI, Invited Papers, Book 1.)* 17–29.

Gutiérrez-Peña, E. and Walker, S. G. (2001). A Bayesian predictive approach to model selection. *Journal of Statistical Planning and Inference* **93** 259–276. MR1822401

Gutiérrez-Peña, E. and Walker, S. G. (2005). Statistical decision problems and Bayesian nonparametric methods. *International Statistical Review* **73** 309–330.

Guttman, I. (1967). The Use of the Concept of a Future Observation in Goodness-of-Fit Problems. *Journal of the Royal Statistical Society. Series B (Methodological)* **29** 83–100. MR0216699

Han, C. and Carlin, B. P. (2000). MCMC methods for computing Bayes factors: A comparative review Research Report No. 2000-001, Division of Biostatistics, University of Minnesota.

Held, L., Schrödle, B. and Rue, H. (2010). Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA. In *Statistical Modelling and Regression Structures* (T. Kneib and G. Tutz, eds.) 91–110. Springer. MR2664630

Hoeting, J., Madigan, D., Raftery, A. and Volinsky, C. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science* **14** 382–401. MR1765176

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika* **76** 297–307. MR1016020

Hurvich, C. M. and Tsai, C.-L. (1991). Bias of the Corrected AIC Criterion for Underfitted Regression and time Series Models. *Biometrika* **78** 499–509. MR1130918

Ibrahim, J. G. and Chen, M.-H. (1997). Predictive Variable Selection for the Multivariate Linear Model. *Biometrics* **53** 465–478. http://www.jstor.org/stable/2533950

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). Criterion-based methods for Bayesian model assessment. *Statistica Sinica* **11** 419–443. MR1844533

Ibrahim, J. G. and Laud, P. W. (1994). A Predictive Approach to the Analysis of Designed Experiments. *Journal of the American Statistical Association* **89** 309–319. MR1266302

Jaakkola, T. S. (2001). Tutorial on variational approximation methods. In *Advanced Mean Field Methods* (M. Opper and D. Saad, eds.) 129–160. The MIT Press. MR1863214

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press (1st edition 1939). MR0187257

Jonathan, P., Krzanowski, W. J. and McCarthy, W. V. (2000). On the use of cross-validation to assess performance in multivariate prediction. *Statistics and Computing* **10** 209–229.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning* **37** 183–233.

Jylänki, P., Vanhatalo, J. and Vehtari, A. (2011). Gaussian Process Regression with a Student-t Likelihood. *Journal of Machine Learning Research* **12** 3227–3257. MR2877599

Karabatsos, G. (2006). Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology* **50**. MR2215142

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90** 773–795.

Key, J. T., Pericchi, L. R. and Smith, A. F. M. (1999). Bayesian Model Choice: What and Why? In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger and A. P. Dawid, eds.) 343–370. Oxford University Press. MR1723504

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics* **22** 79–86. MR0039968

Lacoste-Julien, S., Huszár, F. and Ghahramani, Z. (2011). Approximate inference for the loss-calibrated Bayesian. *Journal of Machine Learning Research: Workshop and Conference Proceedings* **15** 416–424. AISTATS 2011 special issue.

Laud, P. and Ibrahim, J. (1995). Predictive model selection. *Journal of the Royal Statistical Society. Series B (Methodological)* **57** 247–262. MR1325389

Leamer, E. E. (1979). Information Criteria for Choice of Regression Models: A Comment. *Econometrica* **47** 507–510. MR0525791

Leung, D. H.-Y. (2005). Cross-validation in nonparametric regression with outliers. *Annals of Statistics* **33** 2291–2310. MR2211087

Lindley, D. V. (1968). The Choice of Variables in Multiple Regression. *Journal of the Royal Statistical Society. Series B (Methodological)* **30** 31–66. MR0231492

Lo, A. Y. (1987). A Large Sample Study of the Bayesian Bootstrap. *Annals of Statistics* **15** 360–375. MR0885742

MacKay, D. J. C. (1992). A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation* **4** 448–472.

Marin, J.-M. and Robert, C. P. (2010). Importance sampling methods for Bayesian discrimination between embedded models. In *Frontiers of Statistical Decision Making and Bayesian Analysis* (M. H. Chen, D. K. Dey, P. Müller, D. Sun and K. Ye, eds.) 14, 513–553. Springer.

Marriott, J. M., Spencer, N. M. and Pettitt, A. N. (2001). A Bayesian Approach to Selecting Covariates for Prediction. *Scandinavian Journal of Statistics* **28** 87–97. MR1844350

Marshall, E. C. and Spiegelhalter, D. J. (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine* **22** 1649–1660.

San Martini, A. and Spezzaferri, F. (1984). A Predictive Model Selection Criterion. *Journal of the Royal Statistical Society. Series B (Methodological)* **46** 296–303. MR0781890

Mason, D. M. and Newton, M. A. (1992). A Rank Statistics Approach to the Consistency of a General Bootstrap. *Annals of Statistics* **20** 1611–1624. MR1186268

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* **37**, Second ed. Chapman & Hall. MR0727836

McCulloch, R. E. (1989). Local Model Influence. *Journal of the American Statistical Association* **84** 473–478. http://www.jstor.org/stable/2289932

Meng, X.-L. (1994). Posterior Predictive p-Values. *Annals of Statistics* **22** 1142–1160. MR1311969

Meyer, M. C. and Laud, P. W. (2002). Predictive Variable Selection in Generalized Linear Models. *Journal of the American Statistical Association* **97** 859–871. MR1941415

Minka, T. (2001). A Family of Algorithms for Approximate Bayesian Inference PhD thesis, Massachusetts Institute of Technology. MR2717007

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression (with discussion). *Journal of the American Statistical Association* **83**. MR0997578

Miyamoto, J. M. (1999). Quality-Adjusted Life Years (QALY) Utility Models under Expected Utility and Rank Dependent Utility Assumptions. *Journal of Mathematical Psychology* **43** 201–237. MR1689346

Moody, J. E. (1992). The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems. In *Advances in Neural Information Processing Systems 4* (J. E. Moody, S. J. Hanson and R. P. Lippmann, eds.) 847–854. Morgan Kaufmann Publishers.

Murata, N., Yoshizawa, S. and Amari, S.-I. (1994). Network Information Criterion—Determining the number of hidden units for an Artificial Neural Network model. *IEEE Transactions on Neural Networks* **5** 865–872.

Nadeau, C. and Bengio, S. (2000). Inference for the Generalization Error. In *Advances in Neural Information Processing Systems 12* (S. A. Solla, T. K. Leen and K.-R. Müller, eds.) 307–313. MIT Press.

Neal, R. M. (1998). Assessing Relevance Determination Methods Using DELVE. In *Neural Networks and Machine Learning* (C. M. Bishop, ed.) 97–129. Springer-Verlag.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)* **56** 3–48. MR1257793

Nickisch, H. and Rasmussen, C. E. (2008). Approximations for Binary Gaussian Process Classification. *Journal of Machine Learning Research* **9** 2035–2078. MR2452620

Nott, D. J. and Leng, C. (2010). Bayesian projection approaches to variable selection in generalized linear models. *Computational Statistics & Data Analysis* **54** 3227–3241. http://dx.doi.org/10.1016/j.csda.2010.01.036 MR2727748

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

O'Hagan, A. (2003). HSSS model criticism. In *Highly Structured Stochastic Systems* (P. J. Green, N. L. Hjort and S. Richardson, eds.) 423–444. Oxford University Press. MR2082418

O'Hagan, A. and Forster, J. (2004). *Bayesian Inference*, 2nd ed. *Kendalls's Advanced Theory of Statistics* **2B**. Arnold.

Opper, M. and Winther, O. (2000). Gaussian Processes for Classification: Mean-Field Algorithms. *Neural Computation* **12** 2655–2684.

Orr, M. J. L. (1996). Introduction to Radial Basis Function Networks [online] Technical Report, Centre for Cognitive Science, University of Edinburgh. April 1996. Available at http://www.anc.ed.ac.uk/~mjo/papers/intro.ps.gz.

Peruggia, M. (1997). On the Variability of Case-Deletion Importance Sampling Weights in the Bayesian Linear Model. *Journal of the American Statistical Association* **92** 199–207. MR1436108

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics (Oxford, England)* **9** 523–39.

Raftery, A. E. and Zheng, Y. (2003). Discussion: Performance Of Bayesian Model Averaging. *Journal of American Statistical Association* **98** 931–938.

Raftery, A. E., Newton, M. A., Satagopan, J. M. and Krivitsky, P. (2007). Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity (with discussion). In *Bayesian Statistics 8* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 1–45. Oxford University Press. MR2433201

Raiffa, H. and Schlaifer, R. (2000). *Applied Statistical Decision Theory*. John Wiley & Sons. MR1789469

Rasmussen, C. E. and Ghahramani, Z. (2003). Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems 15* (S. Becker, S. Thrun and K. Obermayer, eds.) 489–496. MIT Press, Cambridge, MA. MR2003529

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press. MR2514435

Rasmussen, C. E., Neal, R. M., Hinton, G. E., van Camp, D., Revow, M., Ghahramani, Z., Kustra, R. and Tibshirani, R. (1996). The DELVE Manual [online]. Version 1.1. Available at `ftp://ftp.cs.utoronto.ca/pub/neuron/delve/doc/manual.ps.gz`.

Rencher, A. C. and Pun, F. C. (1980). Inflation of R^2 in Best Subset Regression. *Technometrics* **22** 49–53.

Reunanen, J. (2003). Overfitting in Making Comparisons Between Variable Selection Methods. *Journal of Machine Learning Research* **3** 1371–1382.

Richardson, S. (2002). Discussion to 'Bayesian measures of model complexity and fit' by Spiegelhalter et al. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64** 626–627.

Robert, C. P. (1996). Intrinsic losses. *Theory and decision* **40** 191–214. MR1385186

Robert, C. P. (2001). *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation*, 2nd ed. Springer. MR1835885

Robert, C. P. and Wraith, D. (2009). Computational methods for Bayesian model choice. In *The 29th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. AIP Proceedings* **1193** 251–262.

Robins, J. M., van Der Vaart, A. and Ventura, V. (2000). Asymptotic Distribution of P Values in Composite Null Models. *Journal of the American Statistical Association* **95** 1143–1156. MR1804240

Rubin, D. B. (1981). The Bayesian Bootstrap. *Annals of Statistics* **9** 130–134. MR0600538

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *Annals of Statistics* **12** 1151–1172. MR0760681

Rueda, R. (1992). A Bayesian alternative to parametric hypothesis testing. *Test* **1** 61–67. http://www.springerlink.com/content/37501636313583g2/ MR1266129

Sawa, T. (1978). Information Criteria for Discriminating Among Alternative Regression Models. *Econometrica* **46** 1273–1291. MR0513693

Shao, J. (1993). Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association* **88** 486–494. MR1224373

Shen, X., Huang, H.-C. and Ye, J. (2004). Inference after Model Selection. *Journal of the American Statistical Association* **99** 751–762. http://www.jstor.org/stable/27590445 MR2090908

Shibata, R. (1989). Statistical aspects of model selection. In *From data to model* (J. C. Willems, ed.) 215–240. Springer-Verlag.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* **90** 227–244. MR1795598

Sinha, D., Chen, M.-H. and Ghosh, S. K. (1999). Bayesian Analysis and Model Selection for Interval-Censored Survival Data. *Biometrics* **55** 585–590. MR1705161

Skare, Ø., Bølviken, E. and Holden, L. (2003). Improved sampling-importance resampling and reduced bias importance sampling. *Scandivanian Journal of Statistics* **30** 719–737. MR2155479

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64** 583–639. MR1979380

Stern, H. S. and Cressie, N. (2000). Posterior predictive model checks for disease mapping models. *Statistics in Medicine* **19** 2377–2397.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)* **36** 111–147. MR0356377

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B (Methodological)* **39** 44–47. MR0501454

Sugiyama, M., and Müller, K.-R. (2005). Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions* **23** 249–279. MR2255627

Sugiyama, M., Krauledat, M. and Müller, K.-R. (2007). Covariate Shift Adaptation by Importance Weighted Cross Validation. *Journal of Machine Learning Research* **8** 985–1005.

Sundararajan, S. and Keerthi, S. S. (2001). Predictive Approaches for Choosing Hyperparameters in Gaussian Processes. *Neural Computation* **13** 1103–1118.

Takeuchi, K. (1976). Distribution of Informational Statistics and a Criterion of Model Fitting (in Japanese). *Suri-Kagaku (Mathematic Sciences)* **153** 12–18.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267–288. MR1379242

Tibshirani, R. J. and Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *Annals of Applied Statistics* **3** 822–829. MR2750683

Tierney, L. and Kadane, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association* **81** 82–86. MR0830567

Tran, M.-N., Nott, D. J. and Leng, C. (2011). The predictive Lasso. *Statistics and Computing* 1–16. MR2950086

Trottini, M. and Spezzaferri, F. (2002). A generalized predictive criterion for model selection. *The Canadian Journal of Statistics* **30** 79–96. MR1907678

Vanhatalo, J., Pietiläinen, V. and Vehtari, A. (2010). Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine* **29** 1580–1607. MR2758849

Vannucci, M., Brown, P. J. and Fearn, T. (2003). A Decision theoretical approach to wavelet regression on curves with a high number of regressors. *Journal of Statistical Planning and Inference* **112** 195–212. MR1961730

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7** 91. http://www.ncbi.nlm.nih.gov/pubmed/16504092

Vehtari, A. (2002). Discussion of "Bayesian measures of model complexity and fit" by Spiegelhalter et al. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64** 620.

Vehtari, A. and Lampinen, J. (2002). Bayesian Model Assessment and Comparison Using Cross-Validation Predictive Densities. *Neural Computation* **14** 2439–2468.

Vehtari, A. and Lampinen, J. (2004). Model Selection via Predictive Explanatory Power Technical Report No. B38, Helsinki University of Technology, Laboratory of Computational Engineering.

Vlachos, P. K. and Gelfand, A. E. (2003). On the Calibration of Bayesian Model Choice Criteria. *Journal of Statistical Planning and Inference* **111** 223–234. MR1955883

Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory.* Cambridge University Press. MR2554932

Watanabe, S. (2010a). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research* **11** 3571–3594. MR2756194

Watanabe, S. (2010b). Equations of states in singular statistical estimation. *Neural Networks* **23** 20–34.

Watanabe, S. (2010c). A limit theorem in singular regression problem. *Advanced Studies of Pure Mathematics* **57** 473–492. MR2648274

Weng, C.-S. (1989). On a Second-Order Asymptotic Property of the Bayesian Bootstrap Mean. *Annals of Statistics* **17** 705–710. MR0994261

Yang, Y. (2005). Can the Strengths of AIC and BIC Be Shared? A Conflict between Model Indentification and Regression Estimation. *Biometrika* **92** 937–950. MR2234196

Yang, Y. (2007). Consistency of Cross Validation for Comparing Regression Procedures. *The Annals of Statistics* **35** 2450–2473. MR2382654

Young, A. S. (1987a). On a Bayesian criterion for choosing predictive submodels in linear regression. *Metrika* **34** 325–339. MR0916417

Young, A. S. (1987b). On the information criterion for selecting regressors. *Metrika* **34** 185–194. MR0899029

Zhu, L. and Carlin, B. P. (2000). Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine* **19** 2265–2278.