Farrell-Maupin and Oden *Res Math Sci(2017)4:14*
DOI 10.1186/s40687-017-0104-2

Research in the Mathematical Sciences

RESEARCH Open Access

Adaptive selection and validation of models of complex systems in the presence of uncertainty

Kathryn Farrell-Maupin[1] and J. T. Oden[2*]

*Correspondence: oden@ices.utexas.edu
[2]Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, USA
Full list of author information is available at the end of the article

Abstract

This paper describes versions of OPAL, the Occam-Plausibility Algorithm (Farrell et al. in J Comput Phys 295:189–208, 2015) in which the use of Bayesian model plausibilities is replaced with information-theoretic methods, such as the Akaike information criterion and the Bayesian information criterion. Applications to complex systems of coarse-grained molecular models approximating atomistic models of polyethylene materials are described. All of these model selection methods take into account uncertainties in the model, the observational data, the model parameters, and the predicted quantities of interest. A comparison of the models chosen by Bayesian model selection criteria and those chosen by the information-theoretic criteria is given.

1 Background

One of the principal sources of uncertainty in computer predictions of physical events is the selection of the mathematical model used as a basis for the prediction. For any model selected, there remains the critical question of whether the model can be judged to be valid for the purpose of predicting key quantities of interest. In [11], we presented the Occam-Plausibility ALgorithm (OPAL) as a systematic adaptive procedure for selecting and validating models among a set of possible mathematical models of complex physical phenomena, and specifically in [11], among possible coarse-grained models of atomistic systems. The qualifier "Occam," of course, refers to Occam's razor, in reference to an attempt to select the "simplest" valid model among a set of models. The notions of model simplicity and validity must be made specific to give meaning to such procedures, and are discussed in more detail later in this paper.

An appeal to Occam's razor is not at all new in the history of model selection. In 1970, Box and Jenkins [7] suggested that the *principle of parsimony* should lead to a model with the smallest number of parameters that adequately represent the observational data. The information-theoretic approaches to model selection embodied in Akaike-type criteria and its generalizations do, indeed, lead to measures explicitly dependent on the number of parameters in each model among a set of candidate models. There is a large literature in statistics referencing Occam's razor as a principle for model selection (again, the "prin-

Springer Open

ciple of parsimony of explanations"). So-called Occam factors as a measure of the relative value of one model over another are discussed in, for example, Jaynes [15], Loredo [23], Wolpert [27,28], and elsewhere as a form of a Bayes factor weighted by maximum likelihoods of the respective models. Our own version, embedded in OPAL, provides for not only a parsimonious approach, but, importantly, an approach for addressing model inadequacy and determining model validity relative to estimates of the accuracy with which the model predicts representations of quantities of interest. OPAL is based on Bayesian model plausibilities, but also involves partitioning models into "Occam Categories," which are derived from a measure of simplicity based on the number of parameters in a model.

In the art and science of model validation, much depends upon how one determines that a model "adequately represents" observational data. Such a determination requires a notion of adequacy, i.e., a measure of accuracy with which a model can predict specific data, and a tolerance that must be met in order to deem a model sufficiently accurate. A famous quote, also attributed to Box [6], is "all models are wrong, but some are useful." Validation processes aim to judge which models are useful in predicting specific events in physical systems.

We remark that many studies have been performed on methods of model selection in statistics and biological literature, good examples being the work of Posada and Buckley [25] on model selection and averaging in phylogenetics; the work of Gelman, Hwang, and Vehtari comparing Akaike information criterion (AIC), deviance information criterion (DIC), and Watanabe–Akaike information criterion (WAIC) approximations of cross-validation [12]; the book of Burnham and Anderson [8] on ecological models; and the book of Konishi and Kitagawa [22] on information-theoretic methods in statistical modeling. These studies do not address the fundamental issue that the best model in a set of models, no matter how "goodness" of the model is measured, may be completely unacceptable for the predictive purpose at hand; i.e., the best model may be invalid. The approaches described in the present work address both relative model quality and validity.

In the present paper, we examine and compare alternative forms of OPAL in which different methods of model selection are employed. In particular, we depart from a fully Bayesian approach and explore the frequentist, information-theoretic methods embodied in the Akaike information criterion (AIC). Introduced in the 1970s [2–4], AIC has been studied and used predominantly in areas of ecological and biological sciences, as discussed, for example, in [8]. Variations and extensions have also been introduced to confront computational complications that may arise in cases of limited observational data; see, e.g., [14,22,26]. We give further details on AIC, Bayesian plausibilities, and other methods and compare validation results and predictions using each approach. We continue to focus on the difficult problem of selection and validation of coarse-grained models of atomistic systems, as it exhibits all of the challenges of model validation and quantification of uncertainties in predictions.

Following this Introduction, we review a number of basic concepts that are fundamental to predictive science and, particularly, model validation. We review general methods of model selection in Sect. 3 and describe OPAL in Sect. 4. Applications to coarse graining of models is taken up in Sect. 5 and conclusions are collected in a final section.

2 Preliminaries

To establish the setting for this exposition, we review briefly some basic concepts and notations relevant to predictive modeling following [24]. It is understood in the present work that the term *model* refers to a mathematical model, a collection of mathematical constructions put forth as mathematical abstractions of systems, particularly physical or engineered systems, but social and economic systems could conceivably be considered as well. It is useful to regard a model as characterizing an abstract mathematical problem, such as: Given $\boldsymbol{\theta} \in \Theta \subset \mathcal{R}^k$ and S, find $u(\boldsymbol{\theta}, S) \in V$ such that

$$A(\boldsymbol{\theta}, S; u(\boldsymbol{\theta}, S)) = 0, \tag{1}$$

where A is a collection of mathematical operators, $\boldsymbol{\theta}$ is a vector of model parameters taken from a parameter space Θ (assumed finite dimensional here), S is the scenario in which the model is implemented, and $u(\boldsymbol{\theta}, S)$ is the solution for given $\boldsymbol{\theta}$ and S belonging to a space V of trial functions. For any S, (1) is referred to as *the forward problem*, as the model extrapolates information forward into a solution $u(\boldsymbol{\theta}, S)$ from which predictions of features of the system (or specific events) are derived.

The concept of a scenario is important in the science and technology of model validation. Mathematically, a scenario is viewed as a set of parameter-independent features of the model that can generally be specified exactly, such as the domain of the solution $u(\boldsymbol{\theta}, S)$ or certain boundary and initial data, the idea being that the same model can be used in several different scenarios. The term scenario is used to refer to both the actual physical environment, in which experimental data are collected, and the computational environment, in which the reality to be predicted by the model resides.

In theory, the mathematical model is selected to solve the forward problem in the full prediction scenario S_{p}, and the solution $u(\boldsymbol{\theta}, S)$ is then used to compute specific quantities of interest (QoIs). The QoIs are specific realities targeted in the prediction process. Mathematically, they are usually characterized as functionals Q on the space V of functions in which the solution $u(\boldsymbol{\theta}, S_{\mathrm{p}})$ resides:

$$Q : V \to \mathbb{R}, \quad Q(u(\boldsymbol{\theta}, S_{\mathrm{p}})) = \tilde{Q}(\boldsymbol{\theta}). \tag{2}$$

Physically (and philosophically), the QoIs are not observables, as model predictions may be of events in systems that do not (yet) exist, such as engineering designs. In statistics, they are sometimes identified as "out-of-sample" data (e.g. [12]), extrapolations outside realm of measurable observations.

In processes of calibration and validation of models, other scenarios are considered. At a primitive level, calibration scenarios S_{c} are considered that involve unit tests on components of a model. They are designed to update prior information on model parameters by matching model predictions with experimental calibration data $\mathbf{y}_{\mathrm{c}}$. Multiple calibration scenarios may be considered, each involving different model parameters that could be subsets of those appearing in the prediction scenario. The fundamental issue of model validation is the process of assessing the validity of the model in question as a means to predict the QoI with acceptable accuracy. This involves the design of validation experiments on subsystem models in validation scenarios S_{v}, designed to compare model predictions with validation observable data $\mathbf{y}_{\mathrm{v}}$. The challenges of validation are to design experiments that deliver observational data adequately representing the QoI, to test the validity of hypotheses made in developing the model that may not be fully trusted, to choose an

appropriate metric to measure the accuracy with which the model predicts QoI-informed data $\mathbf{y}_c$, and to select a tolerance γ_{tol}, of the error that the modeler is willing to accept in order to declare the model "valid" (or not invalid). Once a model is determined to be valid (through this subjective process), the model parameters of the valid model (or their statistical representation by appropriate probability density functions) are introduced into the model implemented in the full prediction scenario, the forward problem is solved, and the QoI is evaluated.

Several remarks should be made at this point. Firstly, as noted earlier, and famously noted by Box [6], all models of physical reality are imperfect. The goal of predictive computational science is to determine whether model predictions are "close enough" to reality to use in predictions of events to contribute to scientific knowledge or as a basis for important decisions. Next, the validity of a model depends upon the QoI to be predicted; a model "valid" for one QoI may be invalid for another. It is emphasized that the notion of a valid model can be highly subjective, involving the design of an experiment to mimic a non-observable QoI, the choice of a metric, and the choice of a tolerance for acceptability. Furthermore, predictions are made in the presence of many uncertainties in model parameters, in data $(\mathbf{y}_c, \mathbf{y}_v)$, and in selecting the model itself. Validation methods may or may not address these uncertainties. The QoI is generally a random number or variable. An important challenge is to quantify the uncertainty in the prediction. In addition, a computational model is generally derived from the mathematical model to render it to a form that can be processed by a computer. The discretization of the model, of course, introduces additional errors in the prediction. While not addressed in the current study, this subject is taken up in earlier work [1]. Finally, not all of the parameters of a model necessarily influence the QoI for a particular prediction scenario S_p. Effective methods of measuring the sensitivity of predictions to choices of model parameters and estimates of parameter sensitivity can lead to elimination of models that do not significantly inform the QoI, resulting in a substantial reduction in the complexity of the model selection process.

3 Model selection

The development of a rigorous basis for selecting the "best" model among a set of possible models has been a goal of some modelers, particularly in statistics, for decades. Various measures to assess the quality of one model relative to another have foundations in Bayesian arguments, such as Bayes' factors and Occam factors, as well as information-theoretic arguments derived from frequentist statistics and maximum likelihood approaches. All of the methods of interest involve calculations designed to assess how well model predictions using parametric models agree with observational data or how closely the parametric model can approximate a probability distribution representing the "truth," i.e., the true reality.

In the Bayesian setting, the idea of model evidence and *posterior model plausibilities* is extremely powerful. The notion of posterior plausibilities is mentioned in the 1981 paper of Chow [9], who attributes the idea to Jeffreys' treatments of probability theory [16]. It was certainly known to Schwarz [26], who developed easily implemented approximations to model evidence that lead to the Bayesian information criterion (BIC) in analogy to information-theoretic approaches. More recently, the use of such Bayesian probability approaches for model selection were advocated by Beck and Yuen [5], Hawkins-Daarud et al. [13], and Farrell et al. [10,11]; see also [24].

On the information-theoretic side, the work of Akaike [2,4] leading to the *Akaike information criteria* or its various generalizations [14] are perhaps best known in the domain of frequentist statistics. An account of information-based model selection criteria, including extensions to "generalized information criteria" (GIC), is given in the book of Konishi and Kitagawa [22].

We explore the case in which we have many models from which to choose, and the goal is to identify the model with the most potential for predicting target quantities of interest. Consider the set $\mathcal{M}$ of m parametric model classes,

$$\mathcal{M} = \{\mathcal{P}_1(\boldsymbol{\theta}_1), \mathcal{P}_2(\boldsymbol{\theta}_2), \ldots, \mathcal{P}_m(\boldsymbol{\theta}_m)\}, \tag{3}$$

each with its own parameter space Θ_i from which the parameter vector $\boldsymbol{\theta}_i$ may be chosen. Each model $\mathcal{P}_i(\boldsymbol{\theta}_i)$ is presumed to be developed on the basis of physical and empirical laws described by mathematical constructions. Assume also that we have acquired a set of observational data $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\} \in \mathcal{Y}_S$ where $\mathcal{Y}_S$ is a space of observational data accessible in scenario S.

3.1 Bayesian posterior plausibilities

Suppose that for each model, a prior probability density $\pi(\boldsymbol{\theta}_i|\mathcal{P}_i, \mathcal{M})$ is specified. We simply rewrite Bayes' rule, acknowledging that we have additional conditional information; namely, that the model $\mathcal{P}_j(\boldsymbol{\theta}_j)$ is known to belong to the set $\mathcal{M}$:

$$\pi(\boldsymbol{\theta}_j|\mathbf{y}, \mathcal{P}_j, \mathcal{M}) = \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_j, \mathcal{P}_j, \mathcal{M})}{\pi(\mathbf{y}|\mathcal{P}_j, \mathcal{M})} \cdot \pi(\boldsymbol{\theta}_j|\mathcal{P}_j, \mathcal{M}), \quad 1 \le j \le m. \tag{4}$$

The key term is the *model evidence*, the denominator on the right-hand side of (4). It is the marginalization of the numerator with respect to the parameters:

$$\pi(\mathbf{y}|\mathcal{P}_j, \mathcal{M}) = \int_{\Theta_j} \pi(\mathbf{y}|\boldsymbol{\theta}_j, \mathcal{P}_j, \mathcal{M}) \pi(\boldsymbol{\theta}_j|\mathcal{P}_j, \mathcal{M}) \, \mathrm{d}\boldsymbol{\theta}_j, \quad 1 \le j \le m. \tag{5}$$

The model evidence can be interpreted as a new likelihood function for a discrete version of Bayes' rule over the set $\mathcal{M}$ of models. Its posterior, for each model $\mathcal{P}_j$, denoted ρ_j, is the *posterior model plausibility*,

$$\rho_j = \rho_j(\mathcal{P}_j|\mathbf{y}, \mathcal{M}) = \frac{\pi(\mathbf{y}|\mathcal{P}_j, \mathcal{M})\pi(\mathcal{P}_j|\mathcal{M})}{\pi(\mathbf{y}|\mathcal{M})}, \quad 1 \le j \le m. \tag{6}$$

Choosing the denominator to normalize the set of discrete plausibilities, we have

$$\sum_{j=1}^{n} \rho_j = 1. \tag{7}$$

The model (or models) in $\mathcal{M}$ with the largest plausibility (or plausibilities) ρ_k such that $\rho_k \ge \rho_j$, $1 \le j \le m$, is deemed the most plausible (the "best") model in the set $\mathcal{M}$ for given data $\mathbf{y}$. The prior $\pi(\mathcal{P}_j|\mathcal{M})$ may be set equal to $1/m$ if, initially, all m models are regarded as equally plausible. Otherwise, prior experience may be called upon to weigh one model over another.

3.2 The Akaike information criterion

As in the Bayesian setting, each model has its own likelihood distribution $\pi_j(\mathbf{y}|\boldsymbol{\theta}_j)$. We drop the dependence on $\mathcal{M}$ and replace the dependence on $\mathcal{P}_j$ with the subscript π_j for the moment to simplify notation. Note that each likelihood captures the probability that the model $\mathcal{P}_j$ is able to reproduce the observed data $\mathbf{y}$.

It is customary and convenient in mathematical statistics, particularly in frequentist statistics, to assume the existence of the "truth" or full reality embodied in a probability density f. The expected value with respect to the truth f of the log-likelihood has the following important property: Denote by $\boldsymbol{\theta}^*$ the vector such that

$$\begin{aligned}\boldsymbol{\theta}^* &= \operatorname*{argmax}_{\Theta} \mathbb{E}_f[\log \pi(\mathbf{y}|\boldsymbol{\theta})] \\ &= \operatorname*{argmax}_{\Theta} \int_{\mathcal{Y}_S} f(\mathbf{y}) \log \pi(\mathbf{y}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{y} \\ &= \operatorname*{argmin}_{\Theta} \left[- \int_{\mathcal{Y}_S} f(\mathbf{y}) \log \pi(\mathbf{y}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{y} \right] \\ &= \operatorname*{argmin}_{\Theta} D_{\mathrm{KL}}(f \| \pi(\cdot|\boldsymbol{\theta})), \end{aligned} \tag{8}$$

where D_{KL} is the Kullback–Leibler divergence, also called the the *relative entropy* or the *information loss* between the truth f and the model, represented by the likelihood $\pi(\mathbf{y}|\boldsymbol{\theta})$,

$$D_{\mathrm{KL}}(f \| \pi(\cdot|\boldsymbol{\theta})) = \int_{\mathcal{Y}_S} f(\mathbf{y}) \log \frac{f(\mathbf{y})}{\pi(\mathbf{y}|\boldsymbol{\theta})}\, \mathrm{d}\mathbf{y}. \tag{9}$$

That is, the parameter vector $\boldsymbol{\theta}^*$ that maximizes the log-likelihood with respect to f for given data $\mathbf{y}$ minimizes information lost in approximating the truth with the model.

For a set $\mathcal{M}$ of m models, each with likelihood $\pi_j(\mathbf{y}|\boldsymbol{\theta}_j)$, the averaged D_{KL} over the truth leads to the following measure of information loss for each model:

$$A_j = \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y}} \log \pi_j \left(\mathbf{x} | \hat{\boldsymbol{\theta}}_j(\mathbf{y}) \right), \quad 1 \le j \le m, \tag{10}$$

$\mathbf{x}$ and $\mathbf{y}$ being independent samples taken from the same distribution for scenario S and $\hat{\boldsymbol{\theta}}_j(\mathbf{y})$ is the *maximum likelihood estimate* (MLE) for model $\mathcal{P}_j$, found using data $\mathbf{y}$. As in (8), the expectations $\mathbb{E}_{\mathbf{x}}$ and $\mathbb{E}_{\mathbf{y}}$ are taken with respect to the truth f.

Obviously, since the truth f is not known, the measures A_j cannot be evaluated. However, under certain assumptions on smoothness of the likelihood function and asymptotic behavior of the likelihood as the number n of samples becomes larger, the quantities A_j can be approximated by the Akaike information criterion,

$$\mathrm{AIC}_j = -2 \log \pi_j(\mathbf{y}|\hat{\boldsymbol{\theta}}_j) + 2K_j, \tag{11}$$

where K_j is the number of parameters of model $\mathcal{P}_j$. The model with a lower AIC indicates a model with a lower D_{KL}-distance to reality and is, therefore, a "better" model. Therefore, in general, we seek model $\mathcal{P}_j \in \mathcal{M}$ such that

$$\mathrm{AIC}_j \le \mathrm{AIC}_i, \quad \forall i = 1, 2, \ldots, m. \tag{12}$$

A derivation of (11) is given in [8].

An accepted alternative, the Bayesian information criterion (occasionally referred to as SIC, the Schwarz information criterion), is given by [26],

$$\mathrm{BIC}_j = -2 \log \pi(\mathbf{y}|\hat{\boldsymbol{\theta}}) + K_j \log(n). \tag{13}$$

Also, a "second-order" *AIC* for small sample sizes has been proposed by Hurvich and Tsai [14] which replaced AIC_j of [7] by

$$\mathrm{AIC}_{cj}(\mathbf{y}) = \mathrm{AIC}_j + \frac{2K_j(1+K_j)}{n - K_j + 1}, \tag{14}$$

n being the sample size. Burnham and Andersen [8] comment that "unless the sample size is large with respect to the number of estimated parameters, use of AIC_{cj} is recommended" over AIC_j.

The individual values of AIC_j, BIC_j, or $\mathrm{AIC}c_j$ are generally not interpretable; however, the *relative* values are informative. To this end, and recalling that the minimum AIC_j (or BIC_j or $\mathrm{AIC}c_j$) value indicates to the "best" model, we can calculate the AIC differences,

$$\Delta_j = \mathrm{AIC}_j - \mathrm{AIC}_{\min}. \tag{15}$$

The larger Δ_j is, the less plausible it is that $\mathcal{P}_j$ is the best model for minimizing the information lost in using model $\mathcal{P}_j$ over the "truth" model. Recalling that the evidence $\pi(\mathbf{y}|\mathcal{P}_j, \mathcal{M})$ of (5) can be regarded as the likelihood of the model $\mathcal{P}_j$ given the data $\mathbf{y}$, an information-theoretic version of this likelihood can be put forth that is related to the AIC differences of the form,

$$\pi(\mathbf{y}|\mathcal{P}_j, \mathcal{M}) \propto \exp\left(-\frac{1}{2}\Delta_j\right). \tag{16}$$

Then the Akaike version of model plausibilities is, in analogy with (6), the so-called Akaike weights,

$$w_j = \frac{\exp(-\Delta_j/2)}{\sum_i \exp(-\Delta_i/2)}. \tag{17}$$

The Akaike weights w_i are akin to Bayesian model plausibilities and can be used as a model selection criterion, and the model with weight w_i closest to unity is deemed the best.

4 OPAL

As previously stated, OPAL is an algorithm designed to systematically select the simplest valid model. In the version advocated in [11], the simplicity was defined by the number of parameters such that the simplest model has the fewest parameters among a set of models, and validity was established by passing a validation criterion. It is clear from (11), (13), and (14) that this measure of simplicity is consistent with similar measures of model quality found in frequentist or information theory on model selection criteria, but it could be replaced by other notions of model complexity, if appropriate.

Briefly stated, OPAL consists of the following steps:

1. A set $\mathcal{M}$ of parametric models,

$$\mathcal{M} = \left\{\mathcal{P}_1(\boldsymbol{\theta}_1), \mathcal{P}_2(\boldsymbol{\theta}_2), \ldots, \mathcal{P}_m(\boldsymbol{\theta}_m)\right\} \tag{18}$$

   is identified, each with parameters $\boldsymbol{\theta}_i$ belonging to an appropriate parameter space Θ_i, $1 \le i \le m$.
2. A parameter sensitivity analysis is performed to assess the sensitivity of a model output function $Y(\boldsymbol{\theta})$ on perturbations in model parameters. Those models with parameters not appreciably affecting the output are eliminated, yielding a reduced set $\tilde{\mathcal{M}}$ of models,

$$\tilde{\mathcal{M}} = \left\{\tilde{\mathcal{P}}_1(\tilde{\boldsymbol{\theta}}_1), \tilde{\mathcal{P}}_2(\tilde{\boldsymbol{\theta}}_2), \ldots, \tilde{\mathcal{P}}_l(\tilde{\boldsymbol{\theta}}_l)\right\}, \quad l \le m. \tag{19}$$

3. The models surviving Step 2 are partitioned into "Occam Categories" according to their complexity. Those with the fewest parameters, for example, are put in Category 1, those with the next highest number of parameters in Category 2, and so forth.
4. Models in the set $\mathcal{M}^*$ of Category 1 are calibrated in calibration experiments involving calibration data $\mathbf{y}_c$, yielding a calibrated set of Category 1 models,

$$\mathcal{M}^* = \left\{\mathcal{P}_1^*(\boldsymbol{\theta}_1^*), \mathcal{P}_2^*(\boldsymbol{\theta}_2^*), \ldots, \mathcal{P}_k^*(\boldsymbol{\theta}_k^*)\right\}. \tag{20}$$

5. The posterior Bayesian plausibilities ρ_i of all models in $\mathcal{M}^*$ are computed. Recall that these plausibilities depend explicitly (see, e.g., (6)) and implicitly (via the calibration process (4)) on the calibration data $\mathbf{y}_c$. Only the most plausible models with $\rho_i \geq \rho_j,\ 1 \leq j \leq m$ are retained.
6. An experimental validation scenario is constructed yielding validation observational data $\mathbf{y}_v$, and the most plausible model in Category 1 is used to compute a prediction of the observables $\mathbf{y}_v$; if the difference between the observables and the prediction, measured in an appropriate metric or pseudo-metric, is within a preset tolerance γ_{tol}, the model is deemed "valid." If not, one returns to Step 3 and repeats the process for the next category of models until a valid model is found. If no models of any category are deemed valid, one returns to Step 1 and enlarges the set $\mathcal{M}$ of possible model classes and then proceeds with the steps listed above.
7. Upon identifying a valid model, the forward problem is solved in the prediction scenario and the original QoI is computed, completing the prediction process.

All of these steps are designed to cope with uncertainties in the parameters, the observational data, and the target QoIs, all generally characterized by probability densities. The output $Y(\boldsymbol{\theta})$, when feasible, may be taken to be the QoI available in the full prediction scenario of the model. In regard to eliminating parameters with small sensitivity indices with respect to the output function $Y(\boldsymbol{\theta})$, it should be noted that the elimination of a parameter should be done only if 0 is in the domain of the parameter itself. Otherwise, another nominal value of the parameter must be chosen for the reduced model(s). It should be understood that among the set $\mathcal{M}$ of model classes, there may be many better models than that selected by OPAL, i.e. models in a higher Occam category, which produce predictions closer to the validation observations by some appropriate metric. OPAL is designed to uncover the simplest model (as measured by the number of parameters, for example) that satisfies a predefined validation criterion.

Step 4 in the OPAL algorithm is often the most computationally intensive, and it may be meaningful to consider other simpler methods of model selection when feasible. One goal of this study is to explore, through numerical experiments, the results of model validation when simpler methods, such as the AIC and BIC, are used instead of plausibilities for complex multi-parameter problems. Both the AIC and the BIC are derived using several simplifying approximations that involve truncation error and use of asymptotic estimates, and are not regarded to deliver model selections as accurate as plausibility measures.

5 Application to the selection of coarse-grained models of atomistic systems

One of the most complex challenges in model selection and validation occurs in the construction of coarse-grained (CG) models of atomistic systems—a standard approach in molecular dynamics simulations of chemical and biological systems. CG models are created by aggregating atoms together into representative groups. Interactions between these new groups are generally unknown and must be defined in terms of force potentials to characterize the mathematical representation of each CG model of the molecular system, the parameters of which should be determined following theories, ideas, and processes discussed earlier.

Traditionally, many interatomic potentials are represented by four types of interactions: bonded, angular, torsional (dihedral), and van der Waals. It is quite common to represent

bonds and angles with a harmonic potential and the van der Waals interactions with a Lennard-Jones potential. In the OPLS functional form [17,18], torsional interactions are given a cosine expansion. The OPLS potential energy is therefore given by,

$$\begin{aligned} V(\mathbf{r}) = & \sum_{\text{bonds}} \frac{1}{2} k_{r,i}(r_i - r_{0,i})^2 + \sum_{\text{angles}} \frac{1}{2} k_{\theta,i}(\theta_i - \theta_{0,i})^2 \\ & + \sum_{\text{dihedrals}} \sum_{n=1}^{4} \frac{V_{n,i}}{2} \left[1 + (-1)^{n-1} \cos(n\phi_i)\right] \\ & + \sum_{\text{non-bonded}} 4\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right). \end{aligned} \tag{21}$$

In this equation, $\mathbf{r}$ is the configuration of the atoms (the vector of atomic coordinate vectors), the parameters $k_{r,i}$ and $k_{\theta,i}$ are spring stiffness parameters for the bonds and angles, respectively, $r_{0,i}$ and $\theta_{0,i}$ are the equilibrium bond and angle values, and r_i and θ_i are the bond and angle distances at the given configuration $\mathbf{r}$. The constants $V_{n,i}$ are the dihedral coefficients, and ϕ_i are the torsional angles. For any pair of non-bonded atoms, the Lennard-Jones parameters are ϵ_{ij} and σ_{ij}, and r_{ij} is the distance between them. An electrostatic interaction in the form of a Coulomb potential can also be added; however, in this example, the CG particles are assumed to be charge neutral.

In both the atomistic and coarse-grained systems, the potential energy drives the computational simulation. During these implementations, configurations are sampled and the corresponding potential energy is calculated. The probability density approximated using these samples will be used to the compute the validation metrics discussed later in this section.

Following [11], we consider as a representative example the problem of computing the potential energy of a polyethylene cube. To initialize OPAL, a set $\mathcal{M}$ of possible parametric models is identified. First, the AA-to-CG map is defined, shown in Fig. 1. Each CG particle, shown in red, is defined to represent two carbon atoms and their attached hydrogen atoms. In this example, the set $\mathcal{M}$ is created by tabulating the possible combinations of interactions shown in (21), as shown in Table 1. Further details regarding these types of interactions can be found in [10,11]. The interactions and parameters of each model are the same for those used in the Bayesian setting; therefore, the sensitivity analysis detailed and completed in [11] may be used here. As discussed in previous work, the potential energy in this particular application is insensitive to dihedral parameters. In some cases,

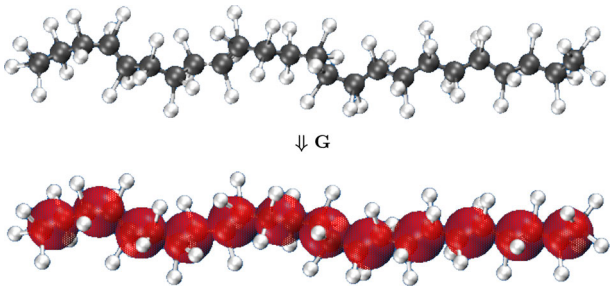

Fig. 1 Representation of the AA-to-CG map G for polyethylene. Carbon atoms are shown in *gray*, hydrogen atoms are shown in *white*, and the CG particles are idealized by the *red spheres*

Table 1 Possible CG models are created by including various combinations of interactions

Model	Bonds	Angles	Dihedrals	LJ 12-6	LJ 9-6	Param	Cat.
$\mathcal{P}_1$	✓					3	1
$\mathcal{P}_2$		✓				3	
$\mathcal{P}_3$				✓		3	
$\mathcal{P}_4$					✓	3	
$\mathcal{P}_5$	✓	✓				5	2
$\mathcal{P}_6$	✓			✓		5	
$\mathcal{P}_7$	✓				✓	5	
$\mathcal{P}_8$		✓		✓		5	
$\mathcal{P}_9$		✓			✓	5	
$\mathcal{P}_{10}$			✓			5	
$\mathcal{P}_{11}$	✓	✓		✓		7	3
$\mathcal{P}_{12}$	✓	✓			✓	7	
$\mathcal{P}_{13}$	✓		✓			7	
$\mathcal{P}_{14}$		✓	✓			7	
$\mathcal{P}_{15}$			✓	✓		7	
$\mathcal{P}_{16}$			✓		✓	7	
$\mathcal{P}_{17}$	✓	✓	✓			9	4
$\mathcal{P}_{18}$	✓		✓	✓		9	
$\mathcal{P}_{19}$	✓		✓		✓	9	
$\mathcal{P}_{20}$		✓	✓	✓		9	
$\mathcal{P}_{21}$		✓	✓		✓	9	
$\mathcal{P}_{22}$	✓	✓	✓	✓		11	5
$\mathcal{P}_{23}$	✓	✓	✓		✓	11	

The Occam categories are determined by counting the number of parameters in the model

insensitive parameters may be set to nominal, constant values. In the present example, the models are nested; thus, the models containing dihedral interactions are redundant and may therefore be eliminated from consideration.

The remaining models are collected into the set $\tilde{\mathcal{M}}$ such that $\tilde{\mathcal{M}} = \{\tilde{\mathcal{P}}_1, \tilde{\mathcal{P}}_2, \ldots, \tilde{\mathcal{P}}_{11}\} = \{\mathcal{P}_1, \ldots, \mathcal{P}_9, \mathcal{P}_{11}, \mathcal{P}_{12}\}$. From Table 1, it can be seen that the lowest category contains those models which depend upon only three parameters. Thus, $\mathcal{M}^* = \{\mathcal{P}_1^*, \mathcal{P}_2^*, \mathcal{P}_3^*, \mathcal{P}_4^*\} = \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4\}$. The MLE for each of these models is determined via a quasi-Newton optimization scheme in which the starting point is the mean value of the parameters determined in an analysis of a simplified AA scenario. Specifically, these mean values are those used in the maximum entropy prior distributions in the Bayesian implementation. See appendix of [11] for complete details.

The calibration scenario consists of a single chain of polyethylene ($C_{80}H_{162}$), simulated in a canonical ensemble. The calibration data $\mathbf{y}_c$ is a vector of observed potential energy values. Once the MLE for each model in Category 1 is obtained, the AIC and Akaike weight for each model may be calculated according to (11), (15) and (17). In the present application, we find that

$$w_1 \approx 1, \quad w_2 \approx 0, \quad w_3 \approx 0, \quad w_4 \approx 0. \tag{22}$$

For this set of models, it is clear that $\mathcal{P}_1^* = \mathcal{P}_1$, which contains only bonded interactions, is considered the "best" model. This result agrees with the best model choice determined using Bayesian plausibilities for model selection [11].

Separate validation scenarios in which parameters are updated again are not typical in deterministic model development. However, for the purpose of following OPAL, we construct a validation test for the AIC-best model. We consider two chains of $C_{80}H_{162}$ simulated in a canonical ensemble and the data $\mathbf{y}_v$ is a set of potential energies. Using the calibration MLE as the starting point for the validation likelihood maximization, we update the MLE.

As validation metrics, we calculate the D_{KL} between the AA and CG distributions produced in the validation scenario, as well as the normalized Euclidean distance between the AA and CG ensemble averages. That is, if u_{AA} and u_{CG} are the set of samples of the potential energy produced by the AA and CG models, respectively, in a given validation scenario,

$$\gamma_1 = |Q_{\mathrm{AA}} - Q_{\mathrm{CG}}| = |\langle u_{\mathrm{AA}}\rangle - \langle u_{\mathrm{CG}}\rangle|, \tag{23}$$

and

$$\gamma_2 = D_{\mathrm{KL}}(u_{\mathrm{AA}}\|u_{\mathrm{CG}}) = \int \pi(u_{\mathrm{AA}}) \log \frac{\pi(u_{\mathrm{AA}})}{\pi(u_{\mathrm{CG}})}\, \mathrm{d}\omega. \tag{24}$$

We take as validation tolerances,

$$\gamma_{1,\mathrm{tol}} = 0.1 Q_{\mathrm{AA}}, \quad \gamma_{2,\mathrm{tol}} = 0.15\sigma_{\mathrm{AA}}^2 \mathcal{O}(Q_{\mathrm{AA}}), \tag{25}$$

for the normalized Euclidean distance and D_{KL} metrics, respectively. Notationally, Q_{AA} is the ensemble average of the observable, $\mathcal{O}(Q_{\mathrm{AA}})$ is its order of magnitude, and σ_{AA}^2 is its variance.

For model $\mathcal{P}_1^*$, the MLE parameters are updated using data from the validation scenario. These parameters are used in a forward implementation of polyethylene. Then, using (23) and (24),

$$\gamma_1 = 1.8371 \times 10^{-4} Q_{\mathrm{AA}} \leq \gamma_{1,\mathrm{tol}}, \quad \gamma_2 = 0.0061\sigma_{\mathrm{AA}}^2 \mathcal{O}(Q_{AA}) \leq \gamma_{2,\mathrm{tol}}, \tag{26}$$

rendering $\mathcal{P}_1^*$ "valid" (not invalid). In a second validation scenario consisting of four chains of $C_{80}H_{162}$, the parameters, without another update, are used to produce a second u_{AA} and u_{CG}. Using the same validation tolerances $\gamma_{1,\mathrm{tol}}$ and $\gamma_{2,\mathrm{tol}}$, both models are again deemed "not invalid" since

$$\gamma_1 = 0.0064 Q_{\mathrm{AA}}, \quad \gamma_2 = 0.0788\sigma_{\mathrm{AA}}^2 \mathcal{O}(Q_{\mathrm{AA}}). \tag{27}$$

Since this model has passed two validation tests, we have confidence to say that it may be used to predict the QoI in our prediction scenario.

Although $\mathcal{P}_1^*$ has been vetted as a "valid" model, we shall move through another iteration of OPAL for the purposes of illustration. Tightening the validation criteria so that

$$\gamma_{2,\mathrm{tol}} = 0.06\sigma_{\mathrm{AA}}^2 \mathcal{O}(Q_{\mathrm{AA}}) \tag{28}$$

renders the Category 1 model $\mathcal{P}_1^*$ invalid. Following the OPAL algorithm, we move to the subset of models in Category 2, in which $\mathcal{M}^* = \{\mathcal{P}_1^*, \mathcal{P}_2^*, \mathcal{P}_3^*, \mathcal{P}_4^*, \mathcal{P}_5^*\} = \{\mathcal{P}_5, \mathcal{P}_6, \mathcal{P}_7, \mathcal{P}_8, \mathcal{P}_9\}$. For each model, the MLE parameters and Akaike weights are calculated for each model,

$$w_1 \approx 1, \quad w_2 \approx 0, \quad w_3 \approx 0, \quad w_4 \approx 0, \quad w_5 \approx 0. \tag{29}$$

Clearly, $\mathcal{P}_1^*$, which accounts for the potential energy in bonds and angles, is the AIC-best model. *Note that this differs from the Bayesian implementation*, in which $\mathcal{P}_2^*$, consisting of bonds and Lennard-Jones 12-6 interactions, is chosen to be the most plausible.

In the validation scenario comprised of two chains of polyethylene, the MLE is updated, and we calculate

$$\gamma_1 = 3.2970 \times 10^{-7}, \quad \gamma_2 = 0.0070\sigma_{\mathrm{AA}}^2 \mathcal{O}(\mu_{\mathrm{AA}}), \tag{30}$$

using (23) and (24), respectively. Moving into the second validation scenario yields

$$\gamma_1 = 0.0063, \quad \gamma_2 = 0.0618\sigma_{\mathrm{AA}}^2 \mathcal{O}(\mu_{\mathrm{AA}}), \tag{31}$$

rendering the Category 2 model *invalid*, and necessitating another iteration through OPAL to Category 3. Note that when OPAL was implemented with Bayesian plausibilities as a model selection criterion, the Category 2 model was found to be *not invalid*. Recalling that γ_1 measures only the difference in mean, while γ_2 takes into account the entire distribution of the potential energy, this result implies that the Bayesian approach better accounts for the various uncertainties present in this application.

Continuing another iteration of OPAL, we select the Category 3 models such that $\mathcal{M}^* = \{\mathcal{P}_1^*, \mathcal{P}_2^*\} = \{\mathcal{P}_{11}, \mathcal{P}_{12}\}$. Both Category 3 models take into account bond, angle, and Lennard-Jones interactions; they differ only in the representation of the Lennard-Jones interaction. Calculating the Akaike weights,

$$w_1 \approx 0, \quad w_2 \approx 1, \tag{32}$$

making the model with 9-6 Lennard-Jones interactions "better" than that with 12-6 Lennard-Jones interactions. In the first validation scenario,

$$\gamma_1 = 6.5408 \times 10^{-8}, \quad \gamma_2 = 1.0518 \times 10^{-4}\sigma_{\mathrm{AA}}^2 \mathcal{O}(\mu_{\mathrm{AA}}), \tag{33}$$

and in the second validation scenario,

$$\gamma_1 = 0.0174, \quad \gamma_2 = 0.0259\sigma_{\mathrm{AA}}^2 \mathcal{O}(\mu_{\mathrm{AA}}), \tag{34}$$

making the Category 3 model $\mathcal{P}_2^*$ not invalid for use in the prediction scenario. The potential energy distributions for all the "AIC-best" models in each of the three Occam Categories in the validation scenarios are given in Fig. 2.

A summary of the Bayesian implementation of OPAL results presented in [11] and the frequentist, AIC-based version of OPAL presented here is given in Table 2. Recall that the parameters were updated in the first validation scenario in both the deterministic and

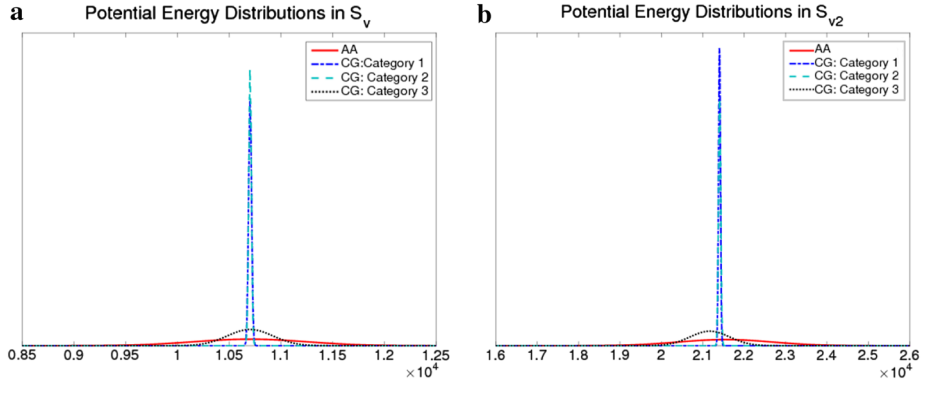

Fig. 2 Potential energy distributions produced by the models chosen using the AIC in the first validation scenario (**a**), and the second validation scenario (**b**)

Table 2 Possible CG models are created by including various combinations of interactions

Method	Model	S_{v1}		S_{v2}	
		γ_1	γ_2	γ_1	γ_2
Level 1					
Bayes	$\mathcal{P}_1^*$	0.0118	0.0622	0.0181	0.0826
AIC	$\mathcal{P}_1^*$	1.8371×10^{-4}	0.0061	0.0064	0.0788
Level 2					
Bayes	$\mathcal{P}_2^*$	0.0115	0.0440	0.0178	0.0587
AIC	$\mathcal{P}_1^*$	3.2970×10^{-7}	0.0070	0.0063	0.0618
Level 3					
Bayes	–	–	–	–	–
AIC	$\mathcal{P}_1^*$	6.5408×10^{-8}	1.0518×10^{-4}	0.0174	0.0259

The Occam categories are determined by counting the number of parameters in the model

Bayesian processes. The MLE calibration is much closer to the data than the parameter distributions produced by Bayesian calibration, as can be seen by a comparison of γ_1 and γ_2 for S_{v1}. Although most of the validation metrics computed in the first validation scenario for the MLE models are lower than those computed for the corresponding Bayesian models, there is a larger jump in these values as the complexity of the scenario increases (e.g., to the second validation scenario). Consider, for example, the validation metric values produced in Level 2. The relative change in γ_1 from S_{v1} to S_{v2} for the Bayesian plausibility is about 55%, while the change in AIC is about 20,000%. For γ_2, this relative change is $<1\%$ for plausibility and nearly 8% for AIC. This may imply that the Bayesian models are more robust for extrapolation to more complex scenarios.

It should be noted that these results depend on the data $\mathbf{y}$ that is used to calibrate the parameters, if Bayes' rule or maximum likelihood estimation is used. Theoretically, as the amount of data increases, the Bayesian posterior $\pi(\boldsymbol{\theta}|\mathbf{d})$ and the MLE $\boldsymbol{\theta}^*$ of the truly best model converge to the true distribution or true value of the parameters, respectively [19–21]. It can be argued that, similarly, as the amount of data increases, Bayesian plausibilities

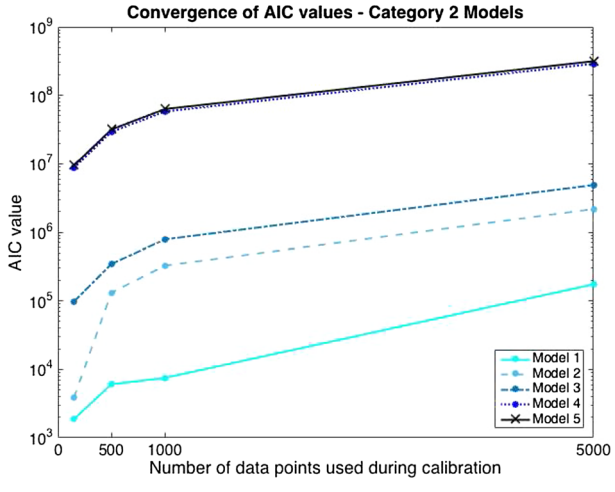

Fig. 3 Evolution of AIC values for Category 2 models as the number of data points increases. In each case, Model 1 is the "best" model. Furthermore, Model 4 and Model 5 differ only in the representation of the Lennard-Jones interaction, which may explain the similarity in their AIC values

and AIC values will converge to indicate the model that will best represent reality. Figure 3 provides plots of the dependence of the AIC values on the available data for models.

6 Concluding comments

On the basis of the sample calculations described in this work on the problem of validating coarse-grained models of atomistic systems, the Akaike and Bayesian criteria for model selection provide an efficient alternative to the more rigorous methods of Bayesian plausibility. Examples of implementations of the OPAL algorithm in model selection and validation suggest that the information-theoretic AIC selection procedures, as expected, seem to provide acceptable criteria for model selection. But in at least one case considered here, the best model selected by AIC differed from that pointed to by Bayesian plausibilities. From a practical point of view, even if the chosen model selection criteria fail to select the best model among a set of models proposed for a prediction, and if this model is invalid, this fact will be caught during the validation phase of OPAL. The better computational efficiency of the AIC methods in comparison with Bayesian plausibilities could make feasible new approaches to model selection and validation in the presence of uncertainties.

Author details
[1] Sandia National Laboratories, Albuquerque, NM, USA, [2] Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, USA.

Acknowledgements
The authors gratefully acknowledge support of their work on predictive science by the US Department of Energy Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under Award Number DE-5D0009286.

Received: 30 June 2016 Accepted: 15 March 2017
Published online: 01 August 2017

References

1. Ainsworth, M., Oden, J.: A Posteriori Error Estimation in Finite Element Analysis. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, Hoboken (2011)
2. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)
3. Akaike, H.: Canonical correlation analysis of time series and the use of an information criterion. Comput. Methods Model. Nonlin. Syst. **126**, 27 (1977)
4. Akaike, H.: On entropy maximization principle. Appl. Stat. (1977). http://ci.nii.ac.jp/naid/10006297543/
5. Beck, J.L., Yuen, K.-V.: Model selection using response measurements: Bayesian probabilistic approach. J. Eng. Mech. **130**(2), 192–203 (2004)
6. Box, G.E.P.: Science and statistics. J. Am. Stat. Assoc. **71**(356), 791–799 (1976)
7. Box, G.E.P., Jenkins, G.M.: Time Series Analysis: Forecasting and Control. Holden-Day Series in Time Series Analysis. Holden-Day, San Francisco (1970)
8. Burnham, K., Anderson, D.: Model Selection and Inference: A Practical Information-Theoretic Approach. Springer, New York (2013)
9. Chow, G.C.: A comparison of the information and posterior probability criteria for model selection. J. Econom. **16**(1), 21–33 (1981)
10. Farrell, K., Oden, J.T.: Calibration and validation of coarse-grained models of atomic systems: application to semiconductor manufacturing. Comput. Mech. **54**(1), 3–19 (2014)
11. Farrell, K., Oden, J.T., Faghihi, D.: A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems. J. Comput. Phys. **295**, 189–208 (2015)
12. Gelman, A., Hwang, J., Vehtari, A.: Understanding predictive information criteria for Bayesian models. Stat. Comput. **24**(6), 997–1016 (2014)
13. Hawkins-Daarud, A., Prudhomme, S., van der Zee, K.G., Oden, J.T.: Bayesian calibration, validation, and uncertainty quantification of diffuse interface models of tumor growth. J. Math. Biol. **67**(6–7), 1457–1485 (2013)
14. Hurvich, C.M., Tsai, C.-L.: Regression and time series model selection in small samples. Biometrika **76**(2), 297–307 (1989)
15. Jaynes, E.T.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge (2003)
16. Jeffreys, H.: The Theory of Probability. OUP, Oxford (1998)
17. Jorgensen, W.L., Maxwell, D.S., Tirado-Rives, J.: Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. **118**(45), 11225–11236 (1996)
18. Jorgensen, W.L., Tirado-Rives, J.: The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. J. Am. Chem. Soc. **110**(6), 1657–1666 (1988)

19. Kleijn, B.J.K.: Bayesian asymptotics under misspecification. Ph.D. thesis, Free University Amsterdam (2004)
20. Kleijn, B.J.K., van der Vaart, A.: The asymptotics of misspecified Bayesian statistics. In: Mikosch, T., Janzura, M. (eds.) Proceedings of the 24th European Meeting of Statisticians (2002)
21. Kleijn, B.J.K., van der Vaart, A.: The Bernstein-von-Mises theorem under misspecification. Electron. J. Stat. **6**, 354–381 (2012)
22. Konishi, S., Kitagawa, G.: Information Criteria and Statistical Modeling. Springer Series in Statistics. Springer, New York (2008)
23. Loredo, T.J.: From Laplace to supernova SN 1987A: Bayesian inference in astrophysics. In: Fougère, P.F. (eds.) Maximum Entropy and Bayesian Methods, pp. 81–142. Kluwer Academic/Springer, Dordrecht (1990)
24. Oden, J.T., Babuska, I., Faghihi, D.: Predictive computational science: computer predictions in the presence of uncertainties. In: Stein, E., de Borst, R., Hughes, T.J.R. (eds.) Encyclopedia of Computational Mechanics. Wiley (2017) (**to appear**)
25. Posada, D., Buckley, T.R.: Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. **53**(5), 793–808 (2004)
26. Schwarz, G., et al.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
27. Wolpert, D.H.: The relationship between Occam's razor and convergent guessing. Complex Syst. **4**, 319–368 (1990)
28. Wolpert, D.H.: A rigorous investigation of evidence and Occam factors in Bayesian reasoning. In: The Sante Fe Institute, 1660 Old Pecos Trail, Suite A, Sante Fe, NM (1992)