The Bayesian Validation Metric: A Framework for Probabilistic Model Calibration and Validation

by

Tony Tohme

Submitted to the Center for Computational Science and Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Computation for Design and Optimization

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author . . . .
Center for Computational Science and Engineering
May 20, 2020

Certified by. . . . .
Kamal Youcef-Toumi
Professor of Mechanical Engineering
Thesis Supervisor

Accepted by . . . .
Youssef Marzouk
Associate Professor of Aeronautics and Astronautics
Co-Director, Center for Computational Science and Engineering

The Bayesian Validation Metric: A Framework for Probabilistic Model Calibration and Validation

by

Tony Tohme

Submitted to the Center for Computational Science and Engineering
on May 20, 2020, in partial fulfillment of the
requirements for the degree of
Master of Science in Computation for Design and Optimization

Abstract

In model development, model calibration and validation play complementary roles toward learning reliable models. In this thesis, we propose and develop the "Bayesian Validation Metric" (BVM) as a general model validation and testing tool. We show that the BVM can represent all the standard validation metrics – square error, reliability, probability of agreement, frequentist, area, probability density comparison, statistical hypothesis testing, and Bayesian model testing – as special cases while improving, generalizing and further quantifying their uncertainties. In addition, the BVM assists users and analysts in designing and selecting their models by allowing them to specify their own validation conditions and requirements. Further, we expand the BVM framework to a general calibration and validation framework by inverting the validation mathematics into a method for generalized Bayesian regression and model learning. We perform Bayesian regression based on a user's definition of model-data agreement. This allows for model selection on any type of data distribution, unlike Bayesian and standard regression techniques, that "fail" in some cases. We show that our tool is capable of representing and combining Bayesian regression, standard regression, and likelihood-based calibration techniques in a single framework while being able to generalize aspects of these methods. This tool also offers new insights into the interpretation of the predictive envelopes in Bayesian regression, standard regression, and likelihood-based methods while giving the analyst more control over these envelopes.

Thesis Supervisor: Kamal Youcef-Toumi
Title: Professor of Mechanical Engineering

This thesis is dedicated to my family

Acknowledgments

This research was performed collaboratively with Dr. Kevin Vanslette under the supervision of Professor Kamal Youcef-Toumi, and it was generously supported by the Center for Complex Engineering Systems (CCES) at King Abdulaziz City for Science and Technology (KACST) and the Massachusetts Institute of Technology (MIT).

I would like to express my sincere gratitude and deep appreciation to Kamal for believing in me and inspiring me. His guidance and support have been invaluable throughout this research.

Much credit for the work in this thesis goes to my fantastic friend, collaborator, and mentor Kevin. His motivation and encouragement have been the driving force behind this research. I honestly owe him tremendously.

I am highly indebted to Associate Provost Philip Khoury whose inspiration and support made my journey at MIT superb and fruitful.

I was very fortunate to take a course with Professor Gilbert Strang. I am genuinely blessed to have worked with him.

Many thanks to Professor Youssef Marzouk, Professor Nicolas Hadjiconstantinou, and Kate Nelson for being behind the success of the Center for Computational Science and Engineering (CCSE).

I am profoundly grateful to Tariq, Kathy, and my friends for being there for me and celebrating with me every success.

Finally, words will never be enough to express my gratitude and love to my parents, Mike and Nicole, my girlfriend Maria, my uncles, Georges and Ziad, my aunts, Simone, Mona and Arlette, my granduncle Emile, and my grandparents, Salma, Yvette and Pierre, for their endless support, encouragement, sacrifices and love; this thesis is dedicated to them. Lastly, yet most importantly, I would like to thank God who has given me the insight, strength, and perseverance to complete this work.

A Note on the Content

This thesis contains material which I authored or co-authored [58, 62].

Chapter 2 of this thesis is based on the following previous publication:

[62] Kevin Vanslette, Tony Tohme, and Kamal Youcef-Toumi. A general model validation and testing tool. *Reliability Engineering & System Safety*, 195, March 2020.

Chapter 3 of this thesis is based on the following manuscript:

[58] Tony Tohme, Kevin Vanslette, and Kamal Youcef-Toumi. Generalized bayesian regression and model learning. *arXiv preprint arXiv:1911.11715*, 2019.

Chapter 1 and Chapter 4 contain material from both papers [58, 62].

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction and Overview

Engineering systems are often represented and described by computational models in order to make predictions about the behavior of the system. Often, models possess parameters that cannot be directly measured, and instead they are inferred based on experimental data of relevant inputs and outputs, in a process known as model calibration [22, 36, 60, 65]. Model calibration is the process of estimating and adjusting model parameters to obtain a model representation of the system (or data) of interest while satisfying a specific criterion (objective function). Once the parameters are inferred, the designed model is tested with respect to real data-generating process, in a task known as model validation [39, 55, 60].

In model development, model calibration comes in the stage prior to the validation stage [52], and it usually consists of estimating model parameters given a set of observed input-output data. Then, validation is performed on a different independent data set called the validation data set. In what follows, we will use the terms "model calibration", "model learning" and "regression" interchangeably.

We are interested in studying the calibration and validation of multivariate computational models that represent uncertain situations and observations (or data). It is understood that complete certainty is a special case of uncertainty as both may be represented with probability distributions.

The uncertainty in a model or data set may originate from stochasticity, model parameter and input data uncertainty, measurement uncertainty, or other possible aleatoric or epistemic sources of uncertainty. Each of the following data modeling schemes may include quantifiable amounts of uncertainty (or certainty) that we would like to calibrate and then validate on the basis of a set of calibration and validation data: neural networks and AI models, machine learning models, Gaussian Process Regression models [18], polynomial chaos and other surrogate models [1, 7, 15, 21], spatial and time series stochastic models, physics based models (usually solutions to differential equations), engineering based models (which are sufficiently abstracted physics based models), Monte Carlo simulation models [16, 35], and more. Model output uncertainties may be quantified through uncertainty propagation techniques (that may or may not include verification, calibration, and validation) [1, 7, 15, 18, 25, 31, 41, 49, 50, 52, 56].

1.2 Related Work

Model calibration techniques have been widely studied in the literature. Least squares (or standard regression) [64], likelihood-based [9, 43], and Bayesian regression methods [23, 24, 32, 40, 63] are often used for model parameter estimation. Nonprobabilistic methods, such as parametric model regression, nonparametric neural networks, and support vector machines (SVM) [4] are able to tackle these types of problems efficiently. In Bayesian probability theory [29, 30, 53], Bayesian model testing and maximum likelihood methods provide probabilistic features (i.e. mean, covariance, distribution) for the parameters we aim to estimate, based on prior knowledge (i.e. prior distribution) and the uncertainty of the data. Bayesian model testing, which uses Bayesian parameter regression, was shown to be successful for signal detection, light sensor characterization [20], exoplanet detection [46], extra-solar planet detection [45], laser peening process [40], time series [59], astronomical data analyses [10], and cosmology and particle physics [11].

Model validation techniques have also been extensively studied in the literature. There exist several validation metrics. Each metric is designed to compare features of a model-data pair to quantify validation: square error compares the difference in the data and model values in a point to point or interval fashion [63], the reliability metric [47] and the probability of agreement [57] compare continuous model outputs and data expectation values (the model reliability metric was extended past expectation values in [51]), the frequentist validation metric [38] and statistical hypothesis testing compare data and model test statistics, the area metric compares the cumulative distribution of the model to the estimated cumulative distribution of the data [12, 26, 27, 49, 65, 67], probability density function (pdf) comparison metrics (such as the KL Divergence) measure and represent "closeness" between pdfs, and Bayesian model testing compares the posterior probability that each model would correctly output the observed data [13, 14, 31, 45, 50, 53, 66]. A detailed review of the majority of these metrics may be found in [22, 28, 34, 36] and the references therein. In particular, [34] is an up to date review that considers many validation metrics in the cases of data and model certainty, data uncertainty and model certainty, and data and model uncertainty.

To assist the comparison of the positive and negative aspects of the above validation metrics, reference [28] outlines six "desirable validation criteria" that a validation metric might have (they extend [12, 38]). One conclusion from [28] is that none of the available metrics simultaneously satisfy all six desirable validation criteria. We summarize the most important features of the desirable validation criteria with the following validation criterion:

1. A validation metric should be a statistically quantified quantitative measure (as opposed to a qualitative measure) of the *agreement* between (general) model predictions and data pairs, in the presence or absence of uncertainty.

The desire for objectivity, "that a metric will produce the same assessment for every analyst independent of their individual preferences" [28], is difficult to satisfy because there are no rules in place to guide a modeler toward selecting one validation metric over another. For this reason, the individual might simply choose a metric based on their preferences, or worse, be tempted to base their decision on which validation metric gives them the most favorable evaluation. Given individuals may choose different validation metrics for the same model-data pair, it is possible for individuals to impose accuracy requirements that are incompatible with one another and arrive at different conclusions regarding the validity of the same model-data pair. As the final goal is objectivity, when possible, a map between the accuracy requirements should be constructed such that the validation metrics yield consistent evaluations of the model-data validity when applicable.

Further, Liu et al. [28] suggest that there is no agreed upon unified model-data comparison function. Even including the results of this thesis, we expect this statement to hold as it is extremely difficult to guess the prior information about the utility of a model an analyst may be required to include in the validation of a model-data pair. For instance, given arbitrary data, "What features of the data are relevant to capture with a model?", "Of these features, are some more relevant than others?", and "What accuracy is required for the model to be valid?". *Agreement* and *validation* are ultimately human-made concepts designed for the purpose of expressing that "in general, not every feature or statistic between a model-data pair need to be equal to conserve the utility of the model". For some model-data pairs, all that may be required is that the model and data averages closely match within uncertainty, while for others, one may require that the model can accurately reproduce the probability distribution of a data set as a whole (as one would do to physically model a noisy measurement device). Given the wide variety of data and the large number of different inferences (and thus models and hypotheses) that one may be interested in drawing from a given data source (i.e. the context of the model-data pair), we do not expect any single set of comparison functions, statistics, or values to be equally relevant and maximally useful for all possible model-data contexts. This, however, does not stop us from quantifying the validity of a model-data pair given any arbitrary comparison function and with any arbitrary definition of agreement.

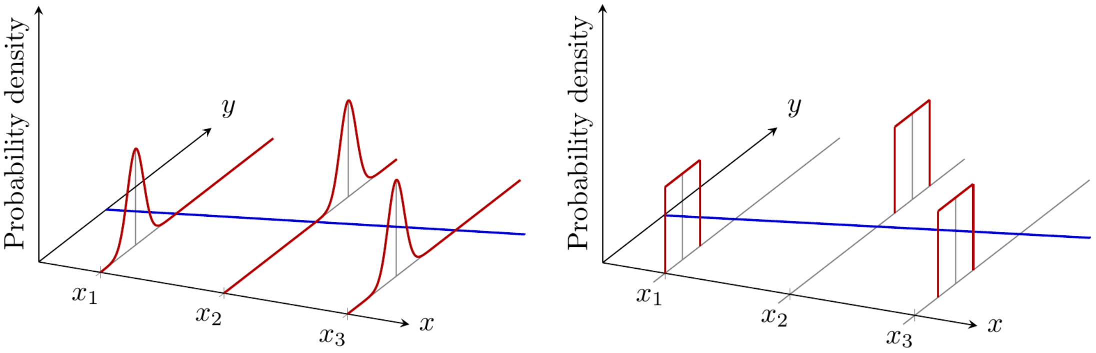
1.3 Objectives and Contributions

In this thesis, we propose and construct the “Bayesian Validation Metric” (BVM) as a general model validation and testing tool. We design the BVM to adhere to the desired validation criterion (1.) by using “four BVM inputs”: the model and data comparison values, the model output and data pdfs, the comparison value function, and the agreement function. The comparison value function is a function of model output and data comparison values that provides the desired quantitative comparison measure, e.g. square difference. Using the model output pdf and the data pdf, the value of the comparison value function is statistically quantified. In turn, the agreement function provides an accept/reject rule and effectively wraps the previous three BVM inputs together to give the BVM. From this, the BVM outputs “the probability the model and data agree”, where *agreement* is a user-defined Boolean function that meets, or does not meet, accuracy requirements between model and data comparison values. Thus, the BVM meets the desired validation criterion (1.) for arbitrary comparison value functions, arbitrary definitions of agreement, and in principle for arbitrary data types such as integers, vectors, tensors, strings, pictures, or others.

The BVM can be used to represent all of the aforementioned validation metrics as special cases. We find the conditions under which several of the current validation metrics are effectively equal to one another, which improves the objectivity of the current validation procedure. In brief we find that the frequentist metric (using natural definitions of agreement) is equal to the reliability metric and the probabilities from Bayesian model testing are equal to the probabilities of the improved model reliability metric [51] when one demands exact equality of the (uncertain) model-data comparison values. Because probability can represent both certain and uncertain situations, so can the BVM. Thus, these “special case” metrics can be generalized to quantify certain or uncertain cases, and even be combined into more complex validation requirements using the BVM framework. Thus, the BVM provides a standardized framework to improve, generalize, or further quantify these validation metrics.

By constructing the “BVM ratio”, we generalize the Bayesian model testing framework [53], in which one constructs the Bayes ratio to rank models according to the ratio of their posterior probabilities given the data. We show that these posterior probabilities are equal to a special case of the BVM under the definition of agreement that requires these uncertain model outputs and data to match exactly. Thus, nothing prevents us from extending the logic used in the Bayesian model testing framework to our framework and we construct the BVM ratio for the purpose of model selection under arbitrary definitions of agreement, i.e. for arbitrary validation scenarios.

Moving to Bayesian calibration techniques, we believe that the efficacy of parametric Bayesian regression, likelihood-based methods, and standard regression can be improved. Bayesian regression calculates the Bayesian evidence, which is the probability the model could have produced the observed, usually over noisy or uncertain, data. If this probability is nonzero, one can proceed to calculate posterior model parameter probabilities using Bayes' Theorem. In practice, there are models and parameters that may be of interest to the user that Bayesian regression fails to regress and produce posterior parameter distributions – Figure 1-1. For some of the instances that Bayesian regression fails to provide a solution, standard regression may actually succeed, but usually with some measure of expected error. How this error can translate into parameter and model uncertainty in the presence of certain or uncertain data is a problem that is largely omitted in the literature except for a few analytic cases.



(a) Bayesian regression works.

(b) Bayesian regression fails.

Figure 1-1: **Illustrative example of theoretical success and failure cases of Bayesian regression.** In blue is a deterministic linear model ($y = ax + b$) and in red are the data probability distributions that may come from epistemic and/or aleatoric uncertainty.

Figure 1-1a shows normally distributed data (infinite tails data distribution). In this case, parametric Bayesian regression finds a linear model that sits in low probability regions of the data. Figure 1-1b shows uniformly distributed data (truncated data distribution). In this case, Bayesian regression cannot find a linear model solution because no linear model can pass through each data distribution simultaneously – the model given the data is regarded as impossible. Standard regression methods can provide linear model solutions here despite the model lying in a zero probability region of the data. Although this solution may be considered "wrong" because it is not supported by the data, it successfully provides useful information to the modeler (an increasing trend). The fact that, for the same model, the solution given by the calibration method can differ from method to method supports the search for a framework for their joint representation so they may be compared more concretely.

In this thesis, we represent least squares, likelihood-based, and Bayesian regression (or calibration) methods by expanding the general validation framework (BVM) into a general calibration and validation framework. Our method uses the BVM to guide the regression of a model in a flexible way. Several of our examples use generalizations of the improved reliability metric and thus reliability is automatically regressed into our model solutions. Our method gives us better control over the predictive envelopes of the model under question, which can be used to improve model reliability and safety. By learning model parameters with the BVM, we are able to estimate and construct model parameters distributions for any type of data distribution (Gaussian, Uniform, Completely Certain), which addresses the concerns raised in Figure 1-1. This construction gives us additional insight into the meaning of the predictive envelopes of Bayesian regression methods.

We have found that a subset of our method shares mathematical features with Approximate Bayesian Computation (ABC) methods, which are also known as likelihood-free techniques [2, 33]. ABC methods are used strictly as an approximation method for nearly computationally intractable likelihoods in Bayesian regression. While our method gains this feature in some cases, our method's intention is not to approximate Bayesian regression, but instead to generalize it for the purpose of robust and flexible model calibration.

Our method is able to regress models over a multitude of different data distributions by using likelihoods that are modified by user's choice of a useful definition of agreement between the data and the model – leaning on the BVM formalism. This "choice" allows the user to program safety requirements into the model learning process if they desire. The nature of the BVM formalism forces one to express the model and data assumptions explicitly and thus our method leads to improved model transparency and safety. In our examples we show how such a procedure leads to a model that better represents the uncertain data at hand than Bayesian and standard regression techniques. This naturally improves the model's reliability and safety in the presence of uncertain data.

1.4 Thesis Outline

The remainder of the thesis is organized as follows. In Chapter 2, we derive and construct the BVM by following our validation criterion. Through some edge cases, we show that the BVM satisfies both the six desirable validation criteria from [28] as well as our validation criterion (1.). We also summarize the results derived in Appendix A, where we incorporate all of the above standard validation metrics as special cases of the BVM, draw relationships between several of the validation metrics, provide improvements and generalizations to these metrics as is suggested by the functional form of the BVM, and construct the BVM ratio. We end the chapter by representing three novel validation metrics using the BVM and comparing them to similar metrics from the literature. We then move to Chapter 3, where we employ the BVM framework to generalize Bayesian regression. This generalized Bayesian regression method, namely the BVM model learning technique, works for different types of data distributions and for arbitrary definitions of model-data agreement. We then present a simulation application using the BVM model learning technique on a nonlinear heuristic model, along with a compound Boolean agreement function example. Chapter 4 is left for conclusions and recommendations.

Chapter 2

The Bayesian Validation Metric

2.1 Introduction

In this chapter, we introduce and develop the Bayesian Validation Metric (BVM) which is *a general model validation and testing tool* [62]. We find the BVM to be capable of representing all of the well-known validation metrics as special cases, while improving, generalizing and further quantifying their uncertainties. The BVM has the capacity to allow users to invent and select models according to novel validation requirements. We formulate and test a few novel compound validation metrics that improve upon other validation metrics in the literature. In addition, we propose and construct the BVM Ratio for the purpose of quantifying model selection under user-specified definitions of model-data agreement in the presence or absence of uncertainty. This construction generalizes the Bayesian model testing framework.

2.2 Notation and Overview

For the remainder of the thesis, we will use the following notation and language. We will let $\hat{y}$ denote the output of a model, $\hat{Y} = \{\hat{y}_1, ..., \hat{y}_n\}$ a set of n model outputs, y a data point (or observed data), and $Y = \{y_1, ..., y_n\}$ a set of n data points. The proposition M essentially stands for "the model" or "coming from the model", i.e. $\hat{y} = M(x; \vec{\alpha})$, where x is the input and $\vec{\alpha} = (\alpha_1, \ldots, \alpha_m)$ represents a vector of m model parameters.

The proposition D stands for "the experiment" or "coming from the experiment". We let $\hat{z}$ and z represent the comparison quantities of interest, which pertain to the model and the data respectively. Further, we let the comparison quantities take general forms, such as multidimensional vectors, functions, or functionals (e.g. output values, expectation values, pdfs, ...), so we can represent any such pair of quantities we may wish to compare between the model and the data. When we refer to "the four BVM inputs" we mean: the comparison values $(\hat{z}, z)$, the model output and data pdf $\rho(\hat{z}, z|M, D)$, the comparison value function $f(\hat{z}, z)$, and the agreement function $B = B(f(\hat{z}, z))$. The (denoted) integrals may be integrals or sums depending on the nature of the variable being summed or integrated over, which is to be understood from the discrete or continuous context of the inference at hand. The dot " $\cdot$ " represents standard multiplication, which is mainly used to improve aesthetics. Finally, we let A denote the *agreement* between the model output and the observed data.

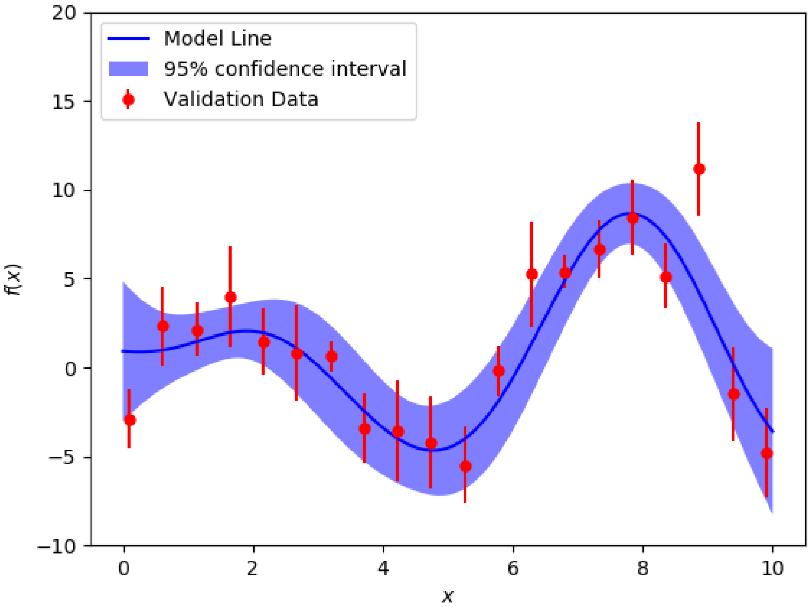


Figure 2-1: **A common model validation scenario.** The model line is trained on noisy data (not depicted in the figure) and is to be compared to a set of validation data. As both the model line and the data are uncertain in general, any quantitative measure (i.e. the comparison function) between these comparison values inherits this uncertainty. Thus, any accept/reject rule on the basis of these uncertain comparison function values is uncertain as well. A visual inspection of this graph seems to indicate, up to statistical fluctuation, that the comparison values of the model and data more or less (or probably) agree, but this intuitive measure has yet to be quantified. Graphic adapted from [44].

Performing uncertainty propagation through a model results in a model output probability (density) distribution $\rho(\hat{y}|M, D)$ that ultimately we would like to validate by comparing it to an uncertain validation data source $\rho(y|D)$, to see if they agree (as depicted in Figure 2-1). The immediate question is, however, "What values do we want to *compare* and what do we mean by *agree*?". Given the wide variety of data and the large number of different inferences (and thus models and hypotheses) that one may be interested in drawing from a given data source (i.e. the context of the model-data pair), we do not expect any single set of comparison functions to be equally relevant and maximally useful for all possible model-data contexts. In light of this, we instead quantify the validity of a model-data pair given any arbitrary comparison value function and according to any arbitrary definition of agreement.

2.3 Formulation of the BVM

In this section, we construct the Bayesian Validation Metric (BVM), we show its different representations, and we discuss its importance and statistical responsibility.

2.3.1 Derivation

Here we start constructing the Bayesian Validation Metric (BVM). To capture the concept of what we might mean by *agree*, we define $\hat{z}$ and z to agree, A, when the Boolean expression, B, is true. Both A and B are defined by the modeler and their prior knowledge of the context of the model-data pair. Naturally then, the agreement function $B = B\big(f(\hat{z}, z)\big) = B(\hat{z}, z)$ is some function or functional of a comparison value function $f(\hat{z}, z)$.

Given the values of $\hat{z}$ and z are known, i.e. certain, we quantify *agreement* using a probability distribution that assigns certainty,

$$p(A|\hat{z}, M, z, D) = \Theta\big(B(\hat{z}, z)\big). \tag{2.1}$$

The indicator function $\Theta\big(B\big)$ is defined to equal unity if B evaluates to "true" (i.e.

"agreeing") and equal to zero otherwise. Thus, in the completely certain case, we are certain as to whether the model and data comparison values agree or do not agree, *as defined by B and the deterministic evaluation of $f(\hat{z}, z)$.*[1] We will call $p(A|\hat{z}, M, z, D)$ the "agreement kernel".

Given that in general the comparison values are uncertain, and quantified by $\rho(\hat{z}, z|M, D)$, the probability the comparison values agree, *as defined by B and $f(\hat{z}, z)$*, is equal to,

$$
\begin{aligned}
p(A|M, D) &= \int_{\hat{z},z} p(A|\hat{z}, M, z, D) \cdot \rho(\hat{z}, z|M, D)\, d\hat{z}\, dz, &(2.2)\\
&\xrightarrow{ind.} \int_{\hat{z},z} \rho(\hat{z}|M, D) \cdot \Theta\big(B(\hat{z}, z)\big) \cdot \rho(z|D)\, d\hat{z}\, dz, &(2.3)
\end{aligned}
$$

which is a marginalization over the spaces of $(\hat{z}, z)$.[2] Equation (2.2) is the general form of the Bayesian Validation Metric (BVM). Because A is discrete, the BVM is a probability rather than a probability density and it therefore falls in the range $0 \leq p(A|M, D) \leq 1$. Equation (2.3) explicitly assumes that the uncertainty in the data is independent of the model, i.e. $\rho(z|M, \hat{z}, D) = \rho(z|D)$, that the data D does not take $\hat{z}$ or the model M (that it is currently being compared to) as inputs.[3] This is a relatively common scenario so it is stated explicitly. The BVM may be computed using any of the well-known computational integration methods.

[1] This binary yet probabilistic definition of agreement turns out to be completely satisfactory for our current purposes. As is briefly discussed later, the sharp boundaries of the indicator function can be smoothed out without employing fuzzy logic by allowing parameters in the Boolean function to themselves be uncertain and marginalized over.

[2] Recall that the propositions in the probability distributions $\rho(z|D)$ and $\rho(\hat{z}|M, D)$ are completely arbitrary (in some cases requiring propagation from $\rho(y|D)$ and $\rho(\hat{y}|M, D)$), they could be both continuous, discrete (with order), categorical variables (no well defined order, e.g. strings, pictures,...), or a mix.

[3] In a controls system this may not be the case, as the model may interact with the system of interest. In such a case this constraint may be lifted and one should use (2.2) instead. The joint probability $\rho(\hat{z}, z|M, D)$ can be used to account for the correlations between the model (the controller or reference) and the data (the measured response of the system being controlled) in a controls setting in principle.

2.3.2 An Identical Representation

In some cases, it is useful to work directly with the probability density $\rho(f|M,D)$, which quantifies the probability the comparison value function $f(\hat{z},z)$ takes the value f due to uncertainty in its inputs. This pdf is independent of any user-defined accuracy requirement. We will call this pdf the comparison value probability density, which is equal to,

$$\rho(f|M,D) = \int_{\hat{z},z} \delta(f - f(\hat{z},z)) \cdot \rho(\hat{z},z|M,D)\, d\hat{z}\, dz. \tag{2.4}$$

This is the net uncertainty propagated through the comparison value function $f(\hat{z},z)$ from the uncertain model and data comparison values. All of the expectation values that are associated with f may be generated from this pdf.

If one imposes an accuracy requirement with a Boolean expression $B = B(f)$ (i.e. defining agreement according to the value of f), the resulting accumulated probability is the BVM. That is, the BVM, i.e. Equation (2.2), may equally be expressed as,

$$p(A|M,D) = \int_f \rho(f|M,D) \cdot \Theta\big(B(f)\big)\, df, \tag{2.5}$$

which is proven through substitution and marginalization over f,

$$\begin{aligned} p(A|M,D) &= \int_f \Big(\int_{\hat{z},z} \delta(f - f(\hat{z},z)) \cdot \rho(\hat{z},z|M,D)\, d\hat{z}\, dz \Big) \cdot \Theta(B(f))\, df \\ &= \int_{\hat{z},z} \Theta(B(\hat{z},z)) \cdot \rho(\hat{z},z|M,D)\, d\hat{z}\, dz. \end{aligned} \tag{2.6}$$

2.3.3 Importance and Statistical Responsibility

The BVM allows the user to, in principle, quantify the probability the model and the data agree with one another under arbitrary comparison value functions and with arbitrary definitions of agreement. The BVM can therefore be used to fully quantify the probability of agreement between arbitrary model and data types using novel or existing comparison value functions and definitions of agreement. Thus, the problem of model-data validation may be reduced to the problem of finding the four BVM inputs in any model validation scenario.

When using the BVM framework, one should practice statistical responsibility by explicitly stating the definition of agreement that is implemented in the validation procedure. Although the flexibility of the BVM framework is a feature, different validation metrics often have different amounts of tolerance as what constitutes "agreement". Agreement according to one metric does *not* in general imply agreement according to another. Overly tolerant definitions of agreement have little resolution power and can only be used responsibly if a large degree of non-exactness between the model and data is permissible. In principle, the definition of agreement should be just as strict or stricter than it needs to be. By explicitly stating the definition of agreement alongside the BVM value, $p(A|M,D) \equiv p(A|M,D,B)$, one avoids statistical misrepresentation by not hiding their definition of agreement.

2.4 Meeting the Desirable Validation Criterion

First we will describe how the BVM, Equations (2.2) and (2.5), precisely match our validation criterion (1.). As can be seen by Equation (2.4), incorporated into the BVM is a statistically quantified quantitative measure that compares data and model outputs, $\rho(f|M,D)$. However, this pdf is in some sense lacking a context pertaining to the model-data pair. Not until an accept/reject rule is imparted on $\rho(f|M,D)$ does one define what is meant by *agreement* in the model-data context. Thus, the BVM only becomes the probability of *agreement* between the data and the model when the agreement function is also incorporated. The four BVM inputs are therefore adequate to satisfy (1.) as the BVM is a "statistically quantified quantitative measure (f) of agreement $p(A|M,D)$ between model predictions and data pairs $(\hat{z}, z)$, in the presence or absence of uncertainty $\rho(\hat{z}, z|M,D)$".

There are a few more BVM concepts worth discussing before moving forward. We will show that the BVM is capable of handling general multidimensional model-data comparisons and that there are no conceptual issues when agreement is exact, i.e. B is true iff $\hat{z} = z$, in the certain and uncertain cases. We will then make comments on the sense in which the BVM adheres to the full set of six desirable validation criteria given in [28] by discussing the criteria that are underrepresented in (1.).

2.4.1 Compound Booleans

Because Boolean operations between Boolean functions result in a Boolean function itself, the BVM is capable of handling multidimensional model-data comparisons. We will call a Boolean function with this property a "compound Boolean". A compound Boolean function results from *and*, $\wedge$, conjunctions and *or*, $\vee$, disjunctions between a set of Boolean functions, e.g.,

$$B(\{B_i\}) = \Big(B_1(\hat{z}, z) \vee B_2(\hat{z}, z)\Big) \wedge \Big(B_3(\hat{z}, z) \vee B_4(\hat{z}, z)\Big) \cdots \tag{2.7}$$

where each $B_i(\hat{z}, z) = B_i(f_i(\hat{z}, z))$ may use a different comparison function $f_i(\hat{z}, z)$. Compound Booleans using conjunctions quantify the validity of entire model functions (random fields and/or multidimensional vectors) by assessing agreement between each of the model-data comparison field points simultaneously, i.e over the comparison points 1 *and* points 2 *and* so on. The compound Booleans may be factored into their constituting Boolean functions using the standard product and sum rules of probability theory after being mapped to probabilities with the agreement kernel. One should be careful when defining an *and* Boolean; if one of the Booleans is false, then the entire Boolean is false. If this strict "all or nothing" validation requirement is not needed then other more flexible definitions of agreement may be instantiated instead (see the BVM examples in Section 2.6).

2.4.2 The BVM Under the Conditions of Exact Agreement

We can calculate the BVM under the conditions of exact agreement in the completely certain and uncertain cases. Because the BVM is a probability rather than a probability density, the agreement kernel falls in the range $[0, 1]$. Under the conditions of exact agreement, B is only true when $\hat{z} = z$, and the agreement kernel is $\Theta(B) = \Theta(\hat{z} = z) = \delta_{\hat{z},z}$, which is the Kronecker delta (i.e. it is 0 or 1) but with continuous labels. As it is uncommon to deal with Kronecker delta's having continuous labels under integration, we will show that the BVM gives reasonable results under the condition of exact agreement in the complete certainty as well as in the general uncertain case.

Complete certainty and exact agreement

Complete certainty is represented using Dirac delta pdf functions over the model and data comparison values. This gives the BVM,

$$\begin{aligned} p(A|M,D) &= \int_{\hat{z},z} \rho(\hat{z}|M,D)\cdot\Theta(\hat{z}=z)\cdot\rho(z|D)\,d\hat{z}\,dz \\ &= \int_{\hat{z},z} \delta(\hat{z}-\hat{z}')\cdot\delta_{\hat{z},z}\cdot\delta(z-z')\,d\hat{z}\,dz, \end{aligned} \tag{2.8}$$

where we are considering the model-data pair to agree iff the comparison values are exactly equal. Using the sifting property of the Dirac delta function, we find the reasonable result that,

$$p(A|M,D) = \delta_{\hat{z}',z'}, \tag{2.9}$$

which is equal to unity iff $\hat{z}'$ and z', the definite values of $\hat{z}$ and z, are equal.

Uncertainty and exact agreement

In the uncertain case under the condition of exact agreement, the BVM is

$$p(A|M,D) = \int_{\hat{z},z} \rho(\hat{z}|M,D)\cdot\delta_{\hat{z},z}\cdot\rho(z|D)\,d\hat{z}\,dz. \tag{2.10}$$

We will do the following trick to correctly interpret this integral. We will first let $B(\epsilon)$ be true if $z-\epsilon \leq \hat{z} \leq z+\epsilon$ and then take the limit as $\epsilon \to 0^+$ such that $\lim_{\epsilon\to 0^+} B(\epsilon) \to B$ when appropriate. With this Boolean expression, the BVM is,

$$\begin{aligned} p(A|M,D,\epsilon) &= \int_{\hat{z},z} \rho(\hat{z}|M,D)\cdot\Theta\Big(z-\epsilon\leq\hat{z}\leq z+\epsilon\Big)\cdot\rho(z|D)\,d\hat{z}\,dz \\ &= \int_{z} \rho(z|D)\bigg(\int_{z-\epsilon}^{z+\epsilon} \rho(\hat{z}|M,D)\,d\hat{z}\bigg)\,dz. \end{aligned} \tag{2.11}$$

In the limit $\epsilon \to 0^+$, the term $\int_{z-\epsilon}^{z+\epsilon} \rho(\hat{z}|M,D)d\hat{z} \to p(\hat{z}=z|M,D) = \rho(\hat{z}=z|M,D)d\hat{z}$ by the definition of probabilities. This gives,

$$
\begin{aligned}
p(A|M,D) &= \int_z \rho(z|D)\Big(p(\hat{z}=z|M,D)\Big)dz = \Big(\int_z \rho(\hat{z}=z|M,D)\cdot\rho(z|D)dz\Big)\,d\hat{z} \\
&\equiv \rho(\hat{z}\equiv z|M,D)\,d\hat{z} = p(\hat{z}\equiv z|M,D), \qquad (2.12)
\end{aligned}
$$

which is understood to be the sum of the model and the data probabilities that jointly output exactly the same values. We see that the BVM in this case is proportional to $d\hat{z}$, $p(A|M,D)\rightarrow\rho(\hat{z}\equiv z|M,D)d\hat{z}$, in the general case of exact agreement, and therefore the BVM goes to zero unless the pdf $\rho(\hat{z}\equiv z|M,D)\propto\delta(\ldots)$. Thus, we recover the standard logical result for probability densities $p(x)=\rho(x)dx\rightarrow 0$ unless it is offset by $\rho(x)\propto\delta(x)$. This result is easily generalized to the dependent case using $\rho(\hat{z},z|M,D)=\rho(z|M,D)\rho(\hat{z}|z,M,D)$. The result, Equation (2.12), is no more surprising than (2.9) in principle.

Due to the vast number of possibilities for continuous valued variables, having a pathological definition of exact agreement between continuous variables does not occur in practice. In a computational setting, $dx\rightarrow\Delta x$ becomes a finite difference and these infinitely improbable agreement conceptual issues are avoided. The Bayesian model testing framework avoids these issues by evaluating posterior odds ratios, in which case the measures, $d\hat{z}$, drop out.

2.4.3 Meeting Underrepresented Validation Criteria

Here we will discuss how the BVM also meets the validation criteria found in [28]. This is done by using the derived general and special cases of the BVM for each of the criteria which are underrepresented in (1.).

Perhaps the primary underrepresented criterion from [28] is their second. It states that "the criteria used for determining whether a model is acceptable or not should not be a part of the metric which is expected to provide a quantitative measurement only." We argue that the functional form of the BVM presented in Equation (2.5) clearly demonstrates this feature as it factors into $\rho(f|M,D)$ and $\Theta(B(f))$. The comparison function $f(\hat{z}, z)$ represents the "objective quantitative measure" from their first criterion that is separate from the accept/reject rule, which is our agreement function $B(f)$ – both of which require definition to ultimately evaluate the validity of a model. We see it as advantageous to quantify the probability the model is accepted or rejected through $B(f)$ due to the uncertainty in the value of f, which is the general case, and which gives the BVM as the result. As all of the validation metrics presented in [28] (and more) will be shown to be representable with the BVM, and thus placed on the same footing, we find our language of "comparison function" and "agreement function" to ultimately be more useful than a language that only considers comparison functions (without accept/reject rules) to be the validation metrics.

The third criteria in [28] is that ideally the metric should "degenerate to the value from a deterministic comparison between scalar values when uncertainty is absent". This is indeed the case as can be seen in Equation (2.1) or in Equations (2.2) and (2.5) by utilizing Dirac delta pdfs similar to their application in Equation (2.8).

The fifth desirable validation criteria in [28] states that artificially widening probability distributions should not lead to higher rates of validation. They find all but the frequentist metric to have this undesired feature; however, we later see that the frequentist metric may be considered a special case of the reliability metric (when reasonable accuracy requirements are imposed), meaning artificial widening can lead to higher rates of validation for more general instances of the frequentist metric.

Further, we argue that artificially introducing uncertainty for the express purpose of passing a validation test is indistinguishable from scientific misconduct. If there is objective reason to include more uncertainty into the analysis or if the circumstance for what constitutes validation has changed due to a change of context – and it happens to improve the rate of validation – so be it. This is a different context, model, or state of uncertainty than was originally proposed so different rates of acceptance should be expected. Reducing the uncertainty of either the data or the model (the inputs) through additional measurements or changing the model may later prove the model valid or invalid when it may have been initially accepted. Thus, to meet this validation criteria, we simply assume the user is not engaging in scientific misconduct.

Finally, due to the results of the *Compound Booleans* in Section 2.4.1, their sixth criterion is met. Because the BVM (2.2) can be used to assess single or multidimensional controllable settings (see footnote number 3) we can perform global function validity (in or out of a controls setting). As they note, “This last feature is critical from the viewpoint of engineering design”.

Thus, the BVM satisfies both our validation criterion and the six desirable validation criteria outlined in [28]. This was accomplished by representing model-data validation as an inference problem using the four BVM inputs.

2.5 Representing and Generalizing the Known Validation Metrics with the BVM

This section is a review of the material found in Appendix A. The following validation metrics will be represented with the BVM, which are then improved, generalized, and/or commented on: reliability/probability of agreement, improved reliability metric, frequentist, area metric, pdf comparison metrics, statistical hypothesis testing, and Bayesian model testing.

2.5.1 Representing the Known Validation Metrics

Table 2.1 shows the values of the four BVM inputs that result in the BVM representing the well-known validation metrics as special cases. The following notation is used for the comparison values $(\hat{z}, z)$. The brackets $\langle \ldots \rangle$ denote expectation values, μ’s denote averaged values, $\hat{y}, y$ denote single values, $\hat{Y}, Y$ denote multidimensional values, $F_{\hat{y}}, F_y$ denote cumulative distribution functions $\big(F_y = \int_{-\infty}^{\hat{y}} \rho(\hat{y}|M, D)\, dy\big)$, $S_{\hat{y}}, S_y$ denote test statistics, and $[-c_\alpha, c_\alpha]$ denotes the $1-\alpha$ confidence interval of the data. In the agreement function column, an element listed as $B(f)$ means the creators of the metric intentionally left the definition of agreement unspecified; however, it is natural to assume it is a function of the comparison function f.

Table 2.1 shows the specification of the four BVM inputs that give the other validation metrics as special cases. It also summarizes some of the similarities and difference between the known validation metrics. In particular, by looking at the validation metrics with the same type of comparison values, i.e. the reliability and frequentist or the improved reliability and Bayesian model testing, we can compare them directly. We see that if one lets the frequentist metric allow for more general input probability distributions and the use of a reasonable agreement function (i.e., $B(f)$ is true if $|f| < \epsilon$), then the frequentist metric is the reliability metric. Further, in Bayesian model testing, if the agreement function $B(f)$ is loosened to accept $f < \epsilon$, then the pdfs that appear in the Bayesian model testing framework are equal to the improved reliability metric. This information improves the objectivity of the current validation procedure because we now have a map between validation metrics that were originally thought to be different.

	Comp. Values		Probs.		Comp. Func.	Agree. Func.
BVM	$\hat{z}$	z	$\rho(\hat{z}\|M, D)$	$\rho(z\|D)$	$f(\hat{z}, z)$	$B(f)$
Reliability	$\langle\hat{y}\rangle$	μ_y	$\rho(\langle\hat{y}\rangle\|M, D)$	$\rho(\mu_y\|D)$	$\|\langle\hat{y}\rangle - \mu_y\|$	$f < \epsilon$
Imp. Reli.	$\hat{Y}$	Y	$\rho(\hat{Y}\|M, D)$	$\rho(Y\|D)$	$\|\hat{Y} - Y\|$	$f < \epsilon$
Frequentist	$\langle\hat{y}\rangle$	μ_y	$\delta(\langle\hat{y}\rangle - \langle\hat{y}\rangle')$	Stud. t	$\langle\hat{y}\rangle - \mu_y$	$B(f)$
Area	$F_{\hat{y}}$	F_y	$\delta(F_{\hat{y}} - F'_{\hat{y}})$	$\delta(F_y - F'_y)$	$\int_{\hat{y}} \|F_{\hat{y}} - F_{y=\hat{y}}\| d\hat{y}$	$B(f)$
Pdf Comp.	ρ_M	ρ_D	$\delta(\rho_M - \rho'_M)$	$\delta(\rho_D - \rho'_D)$	$G(\rho_D\|\|\rho_M)$	$B(f)$
Stat. Hyp.	$S_{\hat{y}}$	S_y	$\rho(S_{\hat{y}}\|M = D)$	$\rho(S_y\|D)$	$S_{\hat{y}}$	$S_{\hat{y}} \in [-c_\alpha, c_\alpha]$
Bayes Model	$\hat{Y}$	Y	$\rho(\hat{Y}\|M, D)$	$\rho(Y\|D)$	$\|\hat{Y} - Y\|$	$f = 0$

Table 2.1: Specification of the four BVM inputs that give the other validation metrics as special cases. The column headings are the four BVM input values: Comparison Values $(\hat{\hat{z}}, z)$, Probabilities $(\rho(\hat{z}|M, D), \rho(z|D))$, Comparison Function $f = f(\hat{z}, z)$, and the Boolean Agreement Function $B(f)$. The row headings read: Reliability, Improved Reliability, Frequentist, Area, Pdf Comparison Metrics, Statistical Hypothesis Testing, and Bayesian Model Testing. The denoted data probability for the average in the frequentist metric, Stud. t, is the Student's t-distribution.

Table 2.2 shows the resulting BVM using the specifications listed in Table 2.1. The value r is the standard notation for the reliability metric [47] and we use r_i for the improved reliability metric [51]. The BVM represents each of the known validation metrics as a probability of agreement between the model and the data from Equation (2.2). As no agreement function is specified directly for the frequentist and area metric, the problem is under constrained so the agreement functions are left as general functions over the comparison function $B(f)$. Thus, for any chosen agreement function, the BVM quantifies their probability of agreement. The remaining metrics all do specify (or indicate) an agreement function, and thus, have specified all of the information required to compute the BVM.

	BVM
BVM	$\int_f \rho(f\|M, D) \cdot \Theta(B(f))\, df$
Reliability	$\int_f \rho(f\|M, D) \cdot \Theta(f < \epsilon)\, df = r$
Imp. Reli.	$\int_f \rho(f\|M, D) \cdot \Theta(f < \epsilon)\, df = r_i$
Frequentist	$\int_{\mu_y} \rho(\mu_y\|D) \cdot \Theta(B(f(\langle \hat{y} \rangle', \mu_y)))\, d\mu_y$
Area	$\Theta(B(f(F'_{\hat{y}}, F'_y))$
Pdf Comp.	$\Theta(B(f(\rho'_M, \rho'_D))$
Stat. Hyp.	$\int_{S_{\hat{y}}} \rho(S_{\hat{y}}\|M = D) \cdot \Theta(S_{\hat{y}} \in [-c_\alpha, c_\alpha])\, dS_{\hat{y}} = 1 - \alpha$
Bayes Model	$\int_f \rho(f\|M, D) \cdot \Theta(f = 0)\, df = p(\hat{Y} \equiv Y\|M, D)$

Table 2.2: BVM representation of the other validation metrics as special cases using the comparison functions (the f's) specified in Table 2.1.

Statistical hypothesis testing is perhaps a bit out of place among the validation metrics. First, note that the comparison function for statistical hypothesis testing is not a function of both the data and the model. Further, note that the model pdf used for statistical hypothesis testing assumes the null hypothesis is true, which in our language is the assumption that $\rho(S_{\hat{y}}|M, D) = \rho(S_{\hat{y}}|M = D)$, i.e. that the pdf of the model is equal to the pdf of the data. This shows how statistical hypothesis testing is a bit out of place here among the validation metrics because here we are attempting to validate a model, usually with its own quantified pdf, rather than, perhaps irresponsibly, assuming it is equal to the data pdf before validating that to be the case. This causes standard statistical hypothesis pitfalls, such as type I (rejecting the null hypothesis when it is true) and type II errors (accepting the null hypothesis when it is false), to be carried over into the BVM, which is unwanted. Several comments are made in Appendix A.5 on this issue.

A perhaps surprising result is the proposed functional form of the BVM that represents Bayesian model testing $p(A|M, D) = p(\hat{Y} \equiv Y|M, D)$, which is the Bayesian evidence. This is the probability that the uncertain model and data output exactly the same values. Usually what is discussed when reviewing Bayesian model testing is the Bayes posterior odds ratio, i.e. the "Bayes Ratio",

$$R = \frac{p(M|Y)}{p(M'|Y)} \propto \frac{p(Y|M)}{p(Y|M')},$$

which tests one model M (i.e. for validation) against another model M'. However, in validation metric problems, we are first interested in considering the validation of a single model – the ratio is an extra bit of inference. In Appendix A.6, we show that the BVM result of $p(\hat{Y} \equiv Y|M, D)$ is exactly what we mean by $p(Y|M)$ in the numerator of the Bayes factor,[4] which effectively quantifies the validation of a single model against the data Y, all quantified under uncertainty.

2.5.2 Generalizing the Known Validation Metrics

The BVM offers several avenues to either generalize or improve many of the metrics. The types of generalizations the BVM offer pertain to generalizing the comparison values, comparison functions, definitions of agreement, and/or generalizing deterministic comparison values and metrics to the uncertain case. These generalizations are only useful if quantitative statements can be made on their behalf – in such a case, these generalizations are improvements. We will give a brief review of the improvements we found below, but the full discussion is located in Appendix A. By making generalizations or improvements to each of the known validation metrics as implied by the BVM, each metric can be made to satisfy our validation criterion as well as the six desirable validation criteria in [28], due to the results of Section 2.4.

[4] It should be noted that our notation for D differs from the notation typically used in Bayesian model testing. Their D is equal to our data Y, while our D refers to context "as having come from the data or experiment rather than the model".

Appendix A.1 uses the BVM to show that the reliability metric and the improved reliability metric can be generalized to compare values without a unique order, such as strings, in principle. This involves creating an agreement function over sets of values (such as synonymous sets of strings), rather than continuous intervals, that may be considered to "agree".

Appendix A.2 derives the frequentist validation metric and generalizes it to the case where both the model and data expectation values are uncertain. The frequentist metric assumes that the model outputs are known with certainty, which may or may not be true. If a model is stochastic, the model pdfs may be estimated with Monte Carlo or other uncertainty propagation methods that quantify the pdf directly.

Appendix A.3 shows that the area metric may be cast as a special case of the BVM. The area metric involves quantifying the difference between model and data cumulative distributions on a point to point basis; thus, the comparison values $(\hat{z}, z)$ are cumulative distributions themselves. The comparison values are assumed to be known with complete certainty, which in the case of cumulative distributions of data is often difficult to argue. Any quantifiable uncertainty in the cumulative distributions may be integrated over, which generalizes the area metric to situations when the model and/or the data cumulative distributions are uncertain. A drawback is that the BVM in these cases may be very computationally intensive and would likely need to be approximated using a random sampling or discretization scheme. A binned pdf metric is put forward to potentially reduce the computational complexity toward quantifying this generalized area validation metric. This applies similarly to the pdf comparison metrics in Appendix A.4.

In Appendix A.5, we invent an improved statistical hypothesis test using the BVM, called the "statistical power BVM", that takes into account both model and data pdfs. Because in principle we have a model output pdf $\rho(\hat{y}|M, D)$ in model validation problems, we can use it (in place of assuming the null hypothesis is true) to avoid both type I and type II errors.

In the statistical power BVM, the model and the data are defined to agree if both their test statistics lie within one another's confidence intervals (or "confidence sets" as explained in Appendix A.5). The statistical power BVM becomes the product of the statistical powers of the model and data, denoted $p(A|M, D) = \big(1-\beta_M(\alpha)\big)\cdot\big(1-\beta_D(\hat{\alpha})\big)$ in Equation (A.18). Further comments are made about how systematic error (defined as when a test statistic lies outside of its *own* confidence interval) may be removed.

It is concluded that the statistical power BVM has a relatively low resolving power compared to other BVMs. This is because large confidence intervals imply large tolerance intervals for acceptance. For this reason, statistical hypothesis testing should only be used for validation in situations where a high degree of nonexactness between model and data test statistics is permissible and the pdfs have very thin tails. This BVM does, however, have a greater resolution than the classical hypothesis test as was proved in Appendix A.5 and will be demonstrated in Section 2.6.

Appendix A.6 finds that Bayesian model testing has the highest possible resolving power because the model and the data are defined to agree only if their values are exactly equal. This is the reverse of what was concluded about statistical hypothesis testing.

Further in Appendix A.6, we argue that, analogous to the Bayesian model testing framework, nothing prevents us from constructing what we call the BVM factor. The BVM factor is,

$$K(B) = \frac{p(A|M, D, B)}{p(A|M', D, B)}, \tag{2.13}$$

which is a ratio of the BVMs of two models under arbitrary definitions of agreement B. Using Bayes' Theorem, $p(M|A, D, B) = p(A|M, D, B)p(M|D, B)/p(A|D, B)$, we may further construct the BVM ratio,

$$R(B) = \frac{p(M|A, D, B)}{p(M'|A, D, B)} = \frac{p(A|M, D, B)\, p(M|D, B)}{p(A|M', D, B)\, p(M'|D, B)} = K(B)\frac{p(M|D, B)}{p(M'|D, B)}, \tag{2.14}$$

for the purpose of comparative model selection under a general definition of agreement B. The ratio $p(M|D, B)/p(M'|D, B)$ is the ratio of prior probabilities of M and M'. Analogous to Bayesian model testing, if there is no reason to suspect that one model is a priori more probable than another, one may let $p(M|D, B)/p(M'|D, B) = 1$, and then $R(B) \to K(B)$ in value.

Thus, using the BVM ratio, we can perform *general model validation testing under arbitrary definitions of agreement and with any reasonable set of comparison functions*. The BVM ratio therefore generalizes the Bayesian model testing framework. This will be utilized in Section 2.6.

Finally we wanted to add a note about how one may mitigate the sharpness of the indicator function without using fuzzy logic while also allowing close models to be somewhat accepted. As we have seen, it is natural to use a threshold Boolean parameter ϵ to help define the boundary of agreement through $B(f \leq \epsilon)$. Such a BVM takes the form,

$$p(A|M, D, \epsilon) = \int_f \rho(f|M, D) \cdot \Theta(B(f \leq \epsilon))\, df, \tag{2.15}$$

where $\Theta(\ldots)$ instantaneously drops to zero for $f > \epsilon$. One may soften the boundary by allowing ϵ itself to be an uncertain quantity, which means one allows their definition of agreement to be somewhat uncertain (which can often be reasonably claimed). As an example, let this uncertainty be $\rho(\epsilon) = \lambda \exp(-\lambda(\epsilon - \epsilon'))$ for $\epsilon' > \epsilon$ and zero otherwise, where λ is positive. Marginalizing over ϵ then gives,

$$\begin{aligned} p(A|M, D) &= \int_\epsilon p(A|M, D, \epsilon) \cdot \rho(\epsilon)\, d\epsilon \\ &= \int_f \rho(f|M, D) \cdot \Big(\Theta\big(B(f \leq \epsilon')\big) + \Theta\big(B(f > \epsilon')\big) e^{-\lambda(f - \epsilon')}\Big)\, df, \end{aligned} \tag{2.16}$$

which allows some f's to be accepted outside the agreement region defined by $f \leq \epsilon'$, but with an exponentially decaying probability. Other potentially useful ϵ pdfs include, but are not limited to: negative slope linear, Gaussian, or decaying sigmoid functions. None of these ϵ type distributions were needed to obtain the results of the previous sections explicitly; however, these types of assumptions may have been part of the decision process made implicitly by a practitioner while performing model validation.

2.6 BVM Examples

In this section, we invent and quantify three novel validation metrics using the BVM to highlight the conceptual clarity, flexibility, and capacity of our framework.

2.6.1 The Statistical Power BVM

Here we consider the statistical power BVM proposed in Appendix A.5 and reviewed in the previous section. This metric defines agreement as occurring when both the model and data comparison values are within one another's confidence intervals, simultaneously. The BVM for this metric is the product of the statistical powers of the model and the data $p(A|M,D) = \big(1-\beta_M(\alpha)\big)\cdot\big(1-\beta_D(\hat{\alpha})\big)$, which is calculated in Equation (A.18). We contrast this with the standard statistical hypothesis test that, after assuming the model is correct, $M = D$, finds the probability that the model lies within the data's confidence interval equal to $1-\alpha$. In statistical hypothesis testing, one then proceeds to check the actual model output and speculates about type I and type II errors. As discussed in Appendix A.5, we do not assume $M = D$ before validation and therefore type I and type II errors are avoided. Rather, we let the statistical power BVM decide whether or not the model is valid. This provides a more informative validation procedure.

Figure 2-2 depicts a typical statistical hypothesis test scenario that is designed to check the validity of an uncertain model average prediction $\hat{\mu}$ (in blue) against an uncertain data average prediction μ (in red). The data's μ is t-distributed (the same distribution in each subfigure) according to $T(\overline{y}, n-1, \overline{s}) = T(0, 10, 1.75)$ where $(\overline{y}, n, \overline{s})$ are the sample mean, the number of collected data points, and the sample standard deviation, respectively. Each row depicts a normally distributed model centered at 0, but with increasing model variance per row.

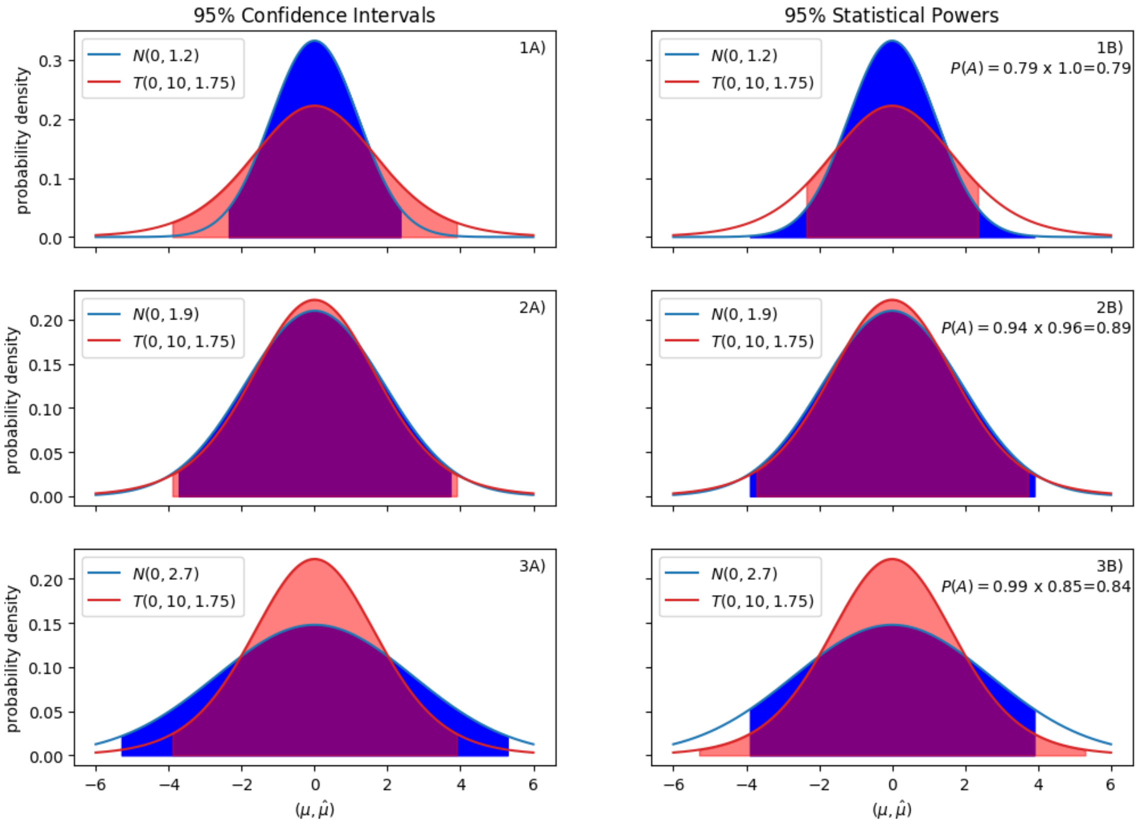


Figure 2-2: **Comparison between statistical hypothesis testing and statistical power BVM.** (a) The shaded regions in Column A depict the 95% confidence interval of each distribution, respectively. Because the data distribution is the same in each figure and because the statistical hypothesis test is independent of the proposed model due to assuming the hypothesis $M = D$, each model is equally valid by that test when it is clear that the model in row 2 is preferable. (b) The shaded regions in Column B depict the statistical power of the distributions – the 95% confidence intervals from each distribution is shaded (integrated) in the other's pdf. The statistical power BVM (denoted $P(A)$ in column B) is calculated for each model and indeed the model in row 2 is found to be preferable as it has the highest probability of agreement.

2.6.2 The $\big(\langle\epsilon\rangle, \beta_D\big)$ BVM

We invent a novel compound Boolean that defines agreement as when the model passes an average square error threshold of $\langle\epsilon\rangle$ *and* a check for probabilistic model representation. The latter is imposed by requiring that $95\% \pm 4\%$ of the uncertain data lies inside the model's $1 - \hat{\alpha} = 95\%$ confidence interval, i.e. $1 - \beta_D(\hat{\alpha}) \sim 95\%$. The $\pm 4\%$ tolerance was chosen such that overly uncertain models would be marked as "not agreeing" as they would be able to guarantee that 100% of the data lies within their excessively wide confidence intervals. We call this compound Boolean the $\big(\langle\epsilon\rangle, \beta_D\big)$ Boolean. The BVM in this case is

$$p(A|M, D, \langle\epsilon\rangle, \beta_D) = \int_{\hat{Y},Y} \rho(\hat{Y}|M, D) \cdot \Theta\Big(B\big(\hat{Y}, Y, \langle\epsilon\rangle, \beta_D\big)\Big) \cdot \rho(Y|D)\, d\hat{Y}\, dY, \tag{2.17}$$

where the compound Boolean $B\big(\hat{Y}, Y, \langle\epsilon\rangle, \beta_D\big)$ is equal to,

$$B\bigg(\frac{1}{n}\sum_i |\hat{y}_i - y_i| \leq \langle\epsilon\rangle\bigg) \wedge B\bigg(0.91 \leq \frac{1}{n}\sum_i \Theta\big(y_i \in [-c_{\hat{\alpha}}, c_{\hat{\alpha}}]_i\big) \leq 0.99\bigg), \tag{2.18}$$

where n is the number of data points in $\{y_i\} = Y$, and $[-c_{\hat{\alpha}}, c_{\hat{\alpha}}]_i$ is the model's 95% confidence interval at comparison location x_i. We treat the model's confidence intervals as certain quantities, which can be achieved effectively through enough Monte Carlo (MC) simulation of the model output pdfs (although this stipulation can be removed if needed).

Although the mathematical notation for the compound Boolean is a bit complicated, it is relatively easy to implement using *if* statements. This ease of programming allows the BVM to have a large capacity for representing complex and abstract validation scenarios in practice.

Expressing the BVM as an expectation value over $\rho(Y|D)\rho(\hat{Y}|M, D)$,

$$p(A|M, D, \langle\epsilon\rangle, \beta_D) = E\Big[\Theta\Big(B\big(\hat{Y}, Y, \langle\epsilon\rangle, \beta_D\big)\Big)\Big] \sim \frac{1}{K}\sum_{k=1}^{K} \Theta\Big(B\big(\hat{Y}^{(k)}, Y^{(k)}, \langle\epsilon\rangle, \beta_D\big)\Big), \tag{2.19}$$

allows one to compute the integral using standard statistical methods like MC. We use MC and $K = 3000$ samples in this toy example. In Figure 2-3, we implement the $\big(\langle\epsilon\rangle, \beta_D\big)$ Boolean and show that it is able to quantify both the average error and a model's probabilistic representation of the uncertain data, simultaneously.

We consider the data generated from,

$$y(x) = 1 + xe^{-\cos{(10x)}} + \sin{(10x)} + \epsilon_a(x), \tag{2.20}$$

where $\epsilon_a(x) \sim \mathcal{N}(0, 0.4^2)$ represents the aleatoric stochastic uncertainty due to the inherent randomness of the system. An instance of the aleatoric data Y without measurement uncertainty is depicted in red in Figures 2-3a and 2-3b. We consider the data to have an additional epistemic measurement uncertainty $\epsilon_e(x) \sim \mathcal{N}(0, 0.2^2)$ that contributes to the probability of whether or not the model agrees with the plotted instance of the aleatoric data Y.

The models plotted in Figures 2-3a and 2-3b are generated from,

$$\hat{y}(x; \vec{\alpha}) = \alpha_1 + \alpha_2 x e^{-\alpha_3 \cos(\alpha_4 x)} + \alpha_5 \sin(\alpha_6 x), \tag{2.21}$$

where $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)$ is the vector of model parameters. In Figure 2-3a, a deterministic model is considered and plotted in blue by treating the model parameters as completely certain numbers $\vec{\alpha} = (1, 1, 1, 10, 1, 10)$. In Figure 2-3b, we consider an uncertain model by treating the model parameters as uncertain values drawn from a multivariate Gaussian distribution having averages $\mu_{\vec{\alpha}} = (1, 1, 1, 10, 1, 10)$ and standard deviations $\sigma_{\vec{\alpha}} = (0.35, 0.3, 0.3, 0.3, 0.3, 0.3)$. We evaluate the probability that these data and model pairs pass the $\langle\epsilon\rangle$ Boolean versus the $\big(\langle\epsilon\rangle, \beta_D\big)$ Boolean by calculating their respective BVMs.

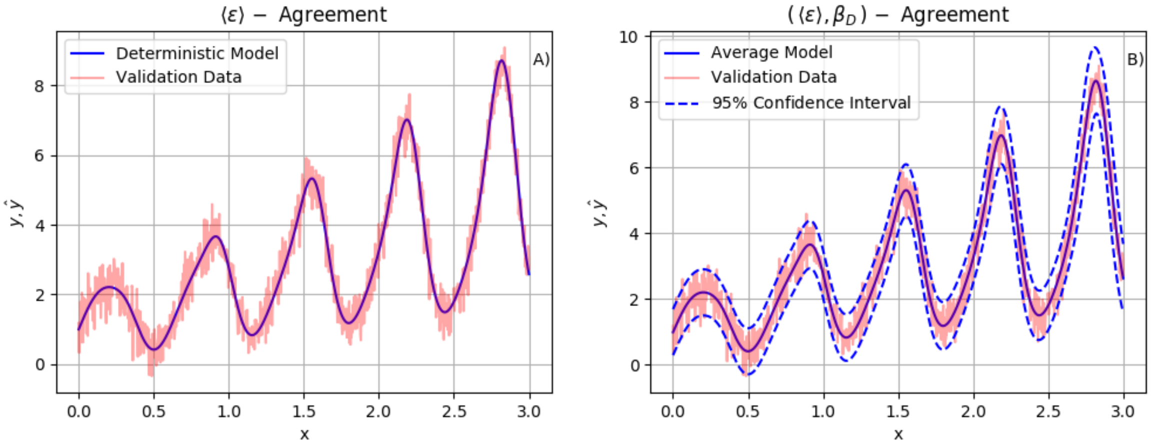


Figure 2-3: **Validating a deterministic model and an uncertain model according to two Boolean agreement functions.** (a) The deterministic model satisfies the $\langle\epsilon\rangle$ Boolean but fails to pass the $\big(\langle\epsilon\rangle, \beta_D\big)$ requirement given that its 95% confidence interval has a width of zero. (b) The uncertain model satisfies both the $\langle\epsilon\rangle$ and $\big(\langle\epsilon\rangle, \beta_D\big)$ agreement requirements given that its 95% confidence region has a nonzero width and represents better the uncertain data.

The deterministic model plotted in Figure 2-3a satisfies the average error validation requirement with $P(A|\langle\epsilon\rangle) = 0.99$ when a threshold of $\langle\epsilon\rangle = 0.46$ is used. This result is logical because the congregate standard deviation of the data is itself the combination of the aleatoric and epistemic uncertainties, i.e. $\sqrt{0.4^2 + 0.2^2} \approx 0.45$. This deterministic model fails to predict the uncertain fluctuations of the data because determinisic models have confidence intervals with zero width. Thus the deterministic model fails to agree according to the $(\langle\epsilon\rangle, \beta_D)$ Boolean and one finds $P(A|\langle\epsilon\rangle, \beta_D) = 0$ for any value of $\langle\epsilon\rangle$.

The uncertain model depicted in Figure 2-3b is able to pass both agreement definitions; however, our choice to evaluate each probable model path (rather than just the average model) against the epistemic uncertain data increases the threshold to about $\langle\epsilon\rangle = 0.9$ before an agreement probability of about $P(A|\langle\epsilon\rangle) = 0.96$ is achieved. When instead evaluating $\langle\epsilon\rangle$ against the average model, we obtained similar results to the deterministic model for this Boolean; $\langle\epsilon\rangle \sim 0.46$ and $P(A|\langle\epsilon\rangle) = 0.99$. The $\big(\langle\epsilon\rangle, \beta_D\big)$ BVM for the uncertain model is $P(A|\langle\epsilon\rangle, \beta_D) = 0.93$, because the uncertainty in the data more or less agrees with the confidence interval provided by the model. This model can be tested for agreement against other models or model parameter distributions for their respective definitions of agreement.

2.6.3 Exploring the BVM Ratio with the (γ, ϵ) BVM

In this section, we invent an agreement function to represent the visual inspection an engineer might perform graphically and use the BVM ratio for model selection under this definition of agreement. By quantifying this, a practitioner could visually validate a model without actually looking at the model-data pair, which can be helpful for high dimensional spaces that are beyond human comprehension/visualizability. We will proceed by introducing this agreement function and some simple models to test it on. We will then quantify this measure using the BVM in the completely certain and uncertain cases. The main purpose of this example is to explore the BVM Ratio while showcasing the conceptual flexibility of the framework.

To quantify something resembling the visual inspection an engineer might make graphically, we use two main criteria. We define the model to be accepted if *most* of the model and data point pairs lie relatively close to one another *and* if none of the point pairs deviate too far from one another. We therefore consider a compound

Boolean $B(\hat{Y}, Y, \gamma, \epsilon)$ that is true if a percentage larger than $\gamma\%$ ($\sim 90\%$) of the model output points $\hat{Y}$ lie within ϵ of the data points Y *and* 100% of the model output points lie within some multiple $m\epsilon$ of the data points, which rules out obvious model form error. We will call this compound Boolean the (γ, ϵ) Boolean. The values $\gamma\%$, ϵ, and m can be adjusted to the needs of the modeler. It should be noted that the ϵ in this metric makes point by point evaluations as opposed to the average $\langle\epsilon\rangle$ Boolean used in the previous example. We will perform the analysis for a variety of γ and ϵ values to explore the limits of the metric.

We will calculate the BVM for two different order polynomial models that approximate n data points taken from the cosine function $y_i = \cos(x_i)$, as an illustration. The points are evenly spaced in the range $x_i \in [0, \pi]$. The first model $\hat{y}_i^{(1)} = M_1(x_i; \vec{\alpha}) = \alpha_1 + \alpha_2 x_i^2 + \alpha_3 x_i^4$ and the second model $\hat{y}_i^{(2)} = M_2(x_i; \vec{\alpha}) = \alpha_1 + \alpha_2 x_i^2 + \alpha_3 x_i^4 + \alpha_4 x_i^6$ have uncertain parameters $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$.[5]

To formulate the BVM, we still need to formulate the model and data probability distributions. Because the Boolean expression $B(\hat{Y}, Y)$ is over the entire model and data functions, the model probability distribution is $p(\hat{Y}|M, D)$ and the data probability distribution is $p(Y|D)$. These are joint probabilities over all of the points $\big(\hat{Y} = \{\hat{y}_i\}, Y = \{y_i\}\big)$ that constitute a particular path $(\hat{Y}, Y)$ of the model or data, respectively. Because both models are linear in the uncertain coefficients $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, there is a one to one correspondence from the set of model parameters $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ to the set of the possible paths $\hat{Y}_{\alpha_1,\alpha_2,\alpha_3,\alpha_4}$ (given n is greater than the number of independent coefficients). This makes the uncertainty propagation from the uncertain model parameters to the full joint probability of the points on a path simple and results in the joint probability of the paths being equal to the joint probability of the uncertain input model parameters. For simplicity, we will let

$$p(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = \mathcal{N}(\mu_{\alpha_1}, \sigma_{\alpha_1})\,\mathcal{N}(\mu_{\alpha_2}, \sigma_{\alpha_2})\,\mathcal{N}(\mu_{\alpha_3}, \sigma_{\alpha_3})\,\mathcal{N}(\mu_{\alpha_4}, \sigma_{\alpha_4}),$$

where $\mathcal{N}(\mu, \sigma)$ is a normal distribution with a mean μ and a standard deviation σ.

[5] Note that $\alpha_4 = 0$ for model 1.

Thus, for each model M_j, we have,

$$p(\hat{Y}_{\alpha_1,\alpha_2,\alpha_3,\alpha_4}|M_j) = \frac{1}{Z}\exp\left(-\frac{(\alpha_1-\mu_{\alpha_1})^2}{2\sigma_{\alpha_1}^2}-\frac{(\alpha_2-\mu_{\alpha_2})^2}{2\sigma_{\alpha_2}^2}-\frac{(\alpha_3-\mu_{\alpha_3})^2}{2\sigma_{\alpha_3}^2}-\frac{(\alpha_4-\mu_{\alpha_4})^2}{2\sigma_{\alpha_4}^2}\right).$$

Because the problem is well understood, we discretize the integrals rather than estimating them with MC [61]. After discretization, the (γ,ϵ) BVM for each model M_j is,

$$p(A|M_j,D,\gamma,\epsilon) = \sum_{\hat{Y},Y} p(\hat{Y}|M_j)\cdot\Theta\big(B(\hat{Y},Y,\gamma,\epsilon)\big)\cdot p(Y|D). \tag{2.22}$$

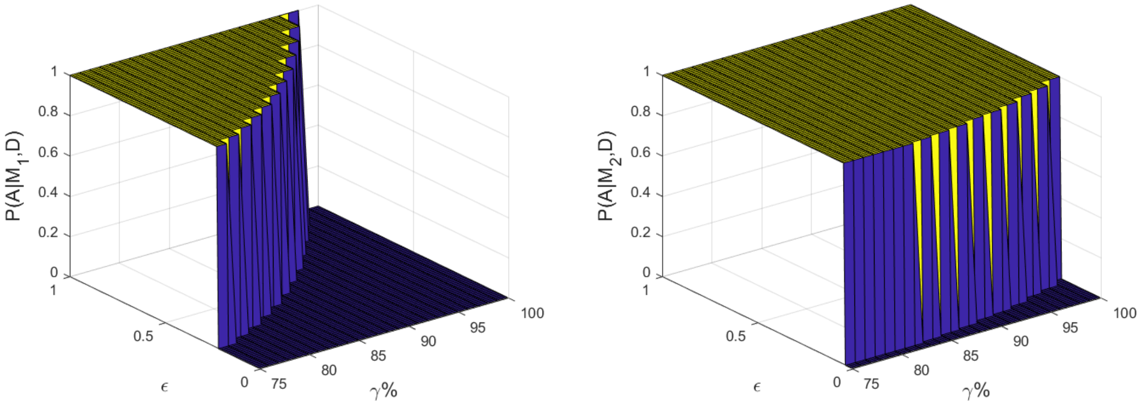
In principle, $\epsilon = (\epsilon_1, .., \epsilon_n)$ is an n-dimensional vector where each ϵ_i may be adjusted to impose more or less stringent agreement conditions on a point to point basis, which may be used to enforce reliability in regions of interest. In our example, we let all the components ϵ_i be equal (i.e. $\epsilon_i = \epsilon$ for all i). If the standard deviation of each data point $\sim \hat{\sigma}_i$ (aleatoric and/or measurement uncertainty) in the joint data pdf $p(Y|D)$ is much less than ϵ_i, and n is large, one may approximate the BVM as,

$$p(A|M_j,D,\gamma,\epsilon) \approx \sum_{\hat{Y}} p(\hat{Y}|M_j)\cdot\Theta\big(B(\hat{Y},Y',\gamma,\epsilon)\big), \tag{2.23}$$

which can greatly reduce the number of combinations one must calculate by effectively treating the data as known, deterministic, and equal to Y'. We will use this approximation as it does not take away from the main point of this example.

We will use the following numerics. In the completely certain (deterministic) case, we will let the parameters be the Taylor series coefficients $(\alpha_1,\alpha_2,\alpha_3,\alpha_4) = \left(1, -\frac{1}{2!}, \frac{1}{4!}, -\frac{1}{6!}\right)$ ($\alpha_4 = 0$ for model 1), and in the uncertain case, we let each coefficient have Gaussian uncertainty centered at their Taylor series coefficients with standard deviations $(\sigma_{\alpha_1},\sigma_{\alpha_2},\sigma_{\alpha_3},\sigma_{\alpha_4}) = (0.1, 0.05, 0.005, 0.0005)$ (and where $\sigma_{\alpha_4} = 0$ for model 1). We let each model output path have $n = 50$ points and we allow for 20 possible values per parameter $(\alpha_1,\alpha_2,\alpha_3,\alpha_4)$, which results in $20^3 = 8000$ possible paths for model 1 and $20^4 = 160,000$ for model 2. We let γ vary between 75% and 100% using an increment of 1% and let ϵ vary between 0 and 1 using an increment of 0.01. The value of m was chosen to be equal to 5, which imposes that no model path can have points that are greater than 5ϵ away while still be considered to agree with the data.

The BVM probability of agreement values as a function of the Boolean function parameters (γ, ϵ) are plotted in Figure 2-4 for model 1 and model 2 in the completely certain case:



(a) 4th order polynomial

(b) 6th order polynomial

Figure 2-4: **Completely Certain Case: The BVM probability of agreement between each of the models and data plotted in the (γ, ϵ) space.** Because here the models are deterministic, the BVM probability of agreement for each (γ, ϵ) pair is either zero or one. (a) The results for model 1 ($4^{\texttt{th}}$ order polynomial). (b) The results for model 2 ($6^{\texttt{th}}$ order polynomial). As expected, model 2 better fits the data in the (γ, ϵ) space as it has more BVM values equal to one than model 1, since it is overall closer to the cosine function being the next nonzero order polynomial in the Taylor series expansion. Neither model fits the data exactly as the BVM for both models at $(\gamma = 100\%, \epsilon = 0)$ is zero.

For a single (γ, ϵ) pair, the BVM ratio (a priori the models are assumed to be equally likely) is,

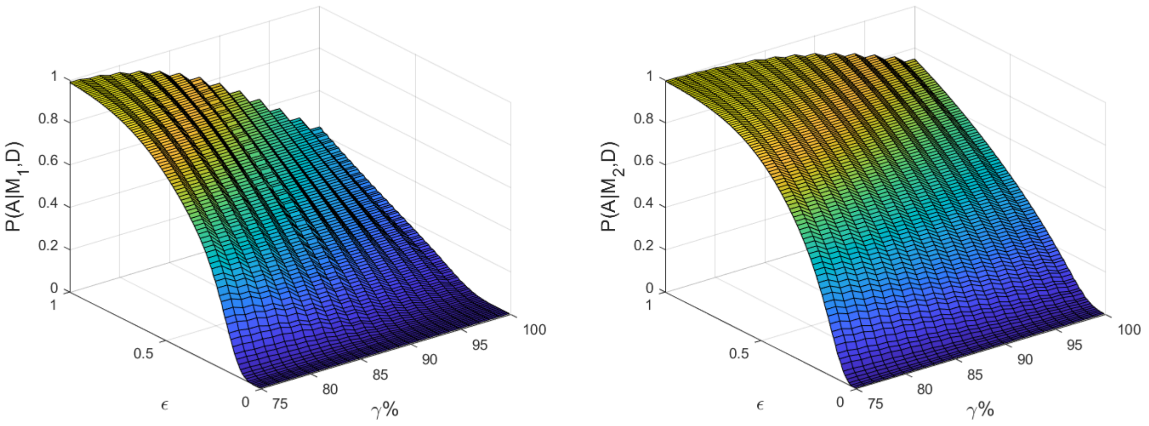
$$R\big(B(\gamma, \epsilon)\big) = \frac{p(A|M_1, D, \gamma, \epsilon)}{p(A|M_2, D, \gamma, \epsilon)}, \tag{2.24}$$

which, because the numerator and denominator are either 0 or 1 in the deterministic case, gives $R\big(B(\gamma, \epsilon)\big)$ equal to 1, 0, ∞, or 0/0 meaning that both models agree, model 1 does not agree but model 2 agrees, model 1 agrees but model 2 does not agree, or both models disagree, respectively. Thus, the BVM ratio for a single (γ, ϵ) pair between two deterministic models with completely certain data is not particularly insightful since they either agree or do not agree as defined by B. As it may not always be clear precisely what values of (γ, ϵ) one should choose to define agreement, one can meaningfully average (marginalize, analagous to (2.16)) over a viable volume in the (γ, ϵ) space, with $p(\gamma, \epsilon) = 1/V$, and arrive at an averaged Boolean BVM ratio,

$$R(B) = \frac{\sum_{\gamma,\epsilon} p(A|M_1, D, \gamma, \epsilon)}{\sum_{\gamma,\epsilon} p(A|M_2, D, \gamma, \epsilon)} = \frac{N_{1A}}{N_{2A}}, \tag{2.25}$$

which is simply a ratio of the number of agreements found for model 1, N_{1A}, to the number of agreements found for model 2, N_{2A}, in the selected (γ, ϵ) volume. In our deterministic example, $R(B) = 1108/2364 = 0.4687$ as model 2 better fits the data, as defined by B, for the chosen meaningful (γ, ϵ) volume (which is taken to be the whole tested volume in this toy example). The BVM ratio or the averaged Boolean BVM ratio may be used as a guide for selecting models in the deterministic case assuming that reasonable regions are chosen.

The BVM probability of agreement values as a function of (γ, ϵ) are plotted in Figure 2-5 for model 1 and model 2 in the uncertain model case:



(a) 4th order polynomial

(b) 6th order polynomial

Figure 2-5: **Uncertain Case: The BVM probability of agreement between each of the models and the data plotted in the (γ, ϵ) space.** Because the model paths are uncertain, the BVM probability of agreement for each (γ, ϵ) pair may take any value from zero to one. (a) The results for model 1 (4th order polynomial). (b) The results for model 2 (6th order polynomial). As expected, model 2 better fits the data in the (γ, ϵ) space as it has generally larger BVM values than model 1; however, the BVM values are about equal in cases of large values of ϵ and low values of γ (since the definition of agreement is less stringent and they both "agree") and in the case of demanding absolute equality ($\epsilon = 0$) as neither model fits the data exactly.

The BVM ratios $R\big(B(\gamma,\epsilon)\big) = p(A|M_1,D,\gamma,\epsilon)/p(A|M_2,D,\gamma,\epsilon)$ for the uncertain models are plotted as a function of (γ,ϵ) in Figure 2-6:

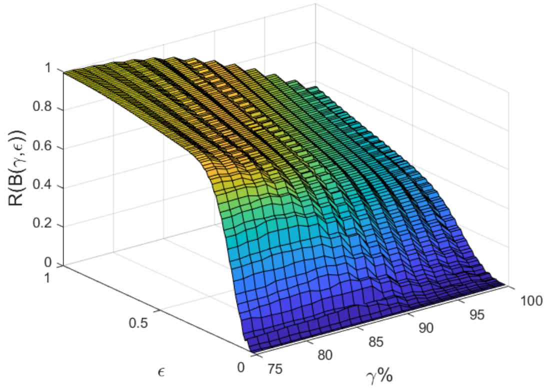


Figure 2-6: **The BVM ratios for the uncertain models plotted in the (γ,ϵ) space.** Model 2 is generally favored over model 1 as there exist no values greater than one on the plot. The amount the BVM ratio favors model 2 over model 1 decreases (i.e. the ratio increases and tends to one) as the metric becomes less and less stringent (i.e. as γ decreases and ϵ increases). The $\epsilon = 0$ line was removed because neither model agrees with the data exactly.

The averaged Boolean BVM ratio for the uncertain models is,

$$R(B) = \frac{\sum_{\gamma,\epsilon} p(A|M_1,D,\gamma,\epsilon)}{\sum_{\gamma,\epsilon} p(A|M_2,D,\gamma,\epsilon)} = 0.7471, \tag{2.26}$$

which conforms to the notion that model 2 is, generally speaking, the preferable model, and which may be communicated with this single number. We have given examples of the BVM ratio correctly selecting models according to abstract and new forms of validation.

2.7 Conclusion

This chapter presents the Bayesian Validation Metric (BVM) as a general model validation and testing tool. This metric is flexible enough to be used in different contexts for solving model validation problems. The BVM quantifies the probability of agreement between some model of interest and observed data, according to arbitrary quantified comparison functions of the model-data comparison values. Further, the BVM obeys all of the desirable validation criteria [28] and represents all of the standard validation metrics as special cases and generalizes them. Finally, using the BVM ratio, one can perform model selection based on arbitrary comparisons and agreement definitions. In other words, the BVM model testing framework generalizes the Bayesian model testing framework to arbitrary model-data contexts.

Chapter 3

Generalized Bayesian Regression and Model Learning

3.1 Introduction

In this chapter, we show how the BVM framework can be employed and expanded to a general calibration and validation framework by proposing a method for *generalized Bayesian regression and model learning* [58]. As we will see, this framework allows the users to perform Bayesian regression on any type of data distribution and based on arbitrary definitions of model-data agreement. This generalized approach has tackled some technical gaps that were found in Bayesian and standard regression methods, and is capable of representing and combining Bayesian regression, standard regression, and likelihood-based calibration techniques in a single framework. Using this framework, we give the users new insights into the interpretation of the predictive envelopes and provide them with more freedom and control over their meaning.

3.2 Background and Motivation

In this section, we will review least squares, likelihood-based, and Bayesian regression methods and discuss their advantages and disadvantages.

3.2.1 Standard Regression

Regression is the process of finding a model that fits some observed data based on a specific mathematical criterion (through the use of an objective function). Regression is one of the most important types of data analysis and is frequently used when making data-driven decisions. The simplest regression technique is linear regression.

Standard regression type methods have several positive and negative attributes. These methods are relatively easy to implement and can approximate model parameters even for reasonably high dimensional data and model parameter spaces (given one is not concerned with parameter or data uncertainty). When the data is uncertain, standard regression methods can generate parameter estimations along with their variances and covariances (analytically for simple cases and iteratively regressing randomly sampled uncertain data in others). In machine learning, it is common practice to regularize the objective function. Regularizing the objective function introduces bias in which the parameter estimations change (i.e. become biased estimators), the parameter variances become reduced, and the model's predictive envelope becomes more narrow and less representative of the data. Although this is not a problem for nonparametric models in which the parameters do not have physical interpretations, we find that regularization is problematic for parametric models because these parameters often represent physical quantities, e.g. the predicted mass of an exoplanet, the predicted circuit resistance due to the addition of an electrical load, or the predicted stiffness of a beam. It seems more natural that larger acceptable training errors should be correlated with an *increase* in the variance of a parameter rather a decrease because one is admitting that the model is not perfect. Currently, regularization causes parameters to become biased as well as "more certain" because the variance of the regressed model is reduced.

Finally, other than using generalization error type estimates (via training and testing error statistics), these methods do not offer any other methods for model selection in which one could easily include their prior knowledge in a principled way.

3.2.2 Likelihood-Based Methods

We give a brief review of likelihood-based methods below, but a more elaborate discussion is located in Appendix B. Likelihood-based methods for calibration and model learning use the likelihood function $\mathcal{L}(\vec{\alpha})$ to learn model parameters. The most common likelihood-based method is the maximum likelihood method.

These methods have similar advantages and disadvantages to standard regression methods. Likelihood-based methods do not allow the user to incorporate their prior knowledge of the model parameters into the framework and, unless there is uncertainty in the data, the learned parameters are point values. These methods also suffer the same conceptual issue of interpreting the affect of regularization on parametric models. One can however learn (or estimate) the prediction uncertainty of the model given the data.

3.2.3 Bayesian Regression and Model Testing

In this section, we present Bayesian regression and model testing (BMT) while introducing some probability notations to be used throughout this chapter. In Bayesian regression, rather than preforming regression to learn the model parameters, one performs regression to learn the posterior probability distribution of the model parameters. That is, one estimates a set of parameters $\vec{\alpha}$ in a model (or hypothesis) $M \equiv M(x; \vec{\alpha})$ for the data D (where x is the model input). The defining equation of Bayesian regression is the learning of the posterior parameter distribution from the prior via Bayes' Rule,

$$\rho(\vec{\alpha}|M) \xrightarrow{*} \rho(\vec{\alpha}|D, M) = \frac{\rho(D|\vec{\alpha}, M)\, \rho(\vec{\alpha}|M)}{\rho(D|M)}, \tag{3.1}$$

where, for reasons that will become obvious, we have borrowed the more explicit notation from Chapter 2. In Bayesian regression and model testing, these probabilities are named as follows:

$\rho(\vec{\alpha}|D, M) \equiv \mathcal{P}(\vec{\alpha})$ is the posterior probability of the parameter,

$\rho(D|\vec{\alpha}, M) \equiv \mathcal{L}(\vec{\alpha})$ is the likelihood function,

$\rho(\vec{\alpha}|M) \quad\ \ \equiv \pi(\vec{\alpha})$ is the prior probability,

$\rho(D|M) \quad\ \ \equiv \mathcal{Z}$ is the marginal likelihood or Bayesian evidence.

After learning the posterior distribution of the model parameters (given $\vec{\alpha}$ is the vector of parameters), we can evaluate the predictive distribution defined by:

$$\rho(\hat{y}|D, M) = \int_{\vec{\alpha}} \rho(\hat{y}|\vec{\alpha}, D, M) \cdot \rho(\vec{\alpha}|D, M)\, d\vec{\alpha}. \tag{3.2}$$

To perform Bayesian regression, one must calculate the Bayesian evidence, which is the marginal likelihood over $\vec{\alpha}$,

$$\mathcal{Z} = \rho(D|M) = \int_{\vec{\alpha}} \underbrace{\rho(D|\vec{\alpha}, M)}_{\mathcal{L}(\vec{\alpha})} \cdot \underbrace{\rho(\vec{\alpha}|M)\, d\vec{\alpha}}_{\pi(\vec{\alpha})\, d\vec{\alpha}}. \tag{3.3}$$

After performing regression and solving for the model parameters' values, rather than selecting the model with the lowest estimated generalization error as is done in standard regression, one instead uses BMT to select the model with the highest probability given the data. That is, for two Bayesian regressed models M_1 and M_2, BMT uses the Bayes ratio, R, and rank the data-informed posterior model probabilities. It can be expressed in several ways using Bayes Rule,

$$R \equiv \frac{p(M_1|D)}{p(M_2|D)} = \frac{\rho(D|M_1)\, p(M_1)}{\rho(D|M_2)\, p(M_2)} = \frac{\mathcal{Z}_1}{\mathcal{Z}_2} \frac{p(M_1)}{p(M_2)}.$$

If there is no reason to suspect that one model is more probable than another prior to observing the data, we may set the ratio of the prior probabilities of the model $p(M_1)/p(M_2) = 1$, *a priori*. In this case one gets,

$$R \to \frac{\mathcal{Z}_1}{\mathcal{Z}_2} \equiv K,$$

where K denotes the Bayes factor and is the ratio of model evidences. The Bayes factor is usually more accessible than R so it is usually used for model selection:
If $K > 1$, then the probability of M_1 given the observed data D is higher than the

probability of M_2 given D. In this case, we select model 1.

If $K < 1$, then the probability of M_2 given the observed data D is higher than the probability of M_1 given D. In this case, we select model 2.

If $K \approx 1$, then the probability of M_1 given the observed data D is equal to the probability of M_2 given D. In this case, both models are equally good or bad.

Bayesian regression has several positive and negative attributes. As a byproduct, Bayesian regression can perform model selection in a principled way that allows one to incorporate their prior knowledge into the selection process using BMT. Because Bayesian regression requires regressing probability distributions rather than just single model predictions, it can become intractable to calculate in general if the number of dimensions are large (as would standard regression if uncertainty is taken into account). Regularization in Bayesian regression is interpreted as coming from the uncertainty of the data and the uncertainty present in the prior parameters [4], which we view as being a potential drawback. If one wants to change the regularization it would require changing either of these uncertainties, or both, "artificially" because one would be tuning their prior probabilities *after* regression, which is a bit anti-Bayesian. Similar to standard regression, regularization can again lead to an unnatural reduction of the posterior variances of the parameters for parametric models.

Further, we highlight some technical gaps found in Bayesian regression and model testing. Although almost all instances of Bayesian regression use data probability distributions that have infinite tails, truncated (or bounded) data probability density functions (pdfs) are realistic in practice too. We find that truncated data pdfs are potentially problematic for Bayesian regression if the model is deterministic. In the extreme case of completely certain data, Bayesian regression methods usually do not terminate because the Bayesian evidence is zero in (3.1) since there are no possible combinations of parameter values that could exactly fit the data. This problem may also arise if the data uncertainties are bounded. In principle, standard regression methods can produce a solution regardless of the form of the data pdf. In what follows, we assume we are given a set of inputs X, a set of data points $Y = D$, and a set of model outputs $\hat{Y} = M(X; \vec{\alpha})$. All the sets are n-dimensional. In addition, we assume that the n data points were collected through independent experiments. We give explicit examples of the likelihoods below (see Appendix C) given the model is deterministic:

Infinite Tail Data Distributions

Data distributions with infinite tails result in likelihoods with infinite tails in (3.3). Some examples of infinite tail data distributions are Gaussian, Student-t, Laplace, canonical, and Poisson. For example, Gaussian distributed data (see Figure 1-1a) naturally has an infinite tailed likelihood function,

$$\mathcal{L}(\vec{\alpha}) = \frac{1}{\sqrt{(2\pi)^n|\Delta|}} e^{-\frac{1}{2}\big(M(X;\vec{\alpha}) - D\big)^T \Delta^{-1}\big(M(X;\vec{\alpha}) - D\big)}, \tag{3.4}$$

where Δ is the covariance matrix. Since the likelihood has infinite tails, the predicted model response $M(x_j;\vec{\alpha})$ has probabilistic flexibility around its corresponding data point D_j because it is uncertain. Even far from D, Bayesian regression is capable of estimating the posterior probability distributions of the model parameters in question as they are nonzero.

Truncated Tail Data Distributions

Data distributions with truncated tails naturally lead to truncated likelihoods in (3.3). For example, if the uncertain data is bounded to a region and is uniformly distributed, i.e. $D_j \sim \mathcal{U}(a_j, b_j)$ (see Figure 1-1b), then the likelihood function is,

$$\mathcal{L}(\vec{\alpha}) = \prod_{j=1}^{n} \frac{\Theta\big(a_j \leq M(x_j;\vec{\alpha}) \leq b_j\big)}{b_j - a_j}, \tag{3.5}$$

where $\Theta(\cdot)$ is the indicator function. In other words, for the likelihood $\mathcal{L}(\vec{\alpha})$ to be nonzero, the predicted model response $M(x_j;\vec{\alpha})$ at x_j must lie within the interval $\big[a_j, b_j\big]$ for all j simultaneously. The function space defined by the model and uncertain parameters is constrained by the data. This can make the probability of estimating a regressed posterior probability distribution of the model parameters very small, and in some cases impossible, because the likelihood may evaluate to zero for almost all combinations of $\vec{\alpha}$.

This point is exaggerated if the data is completely certain or deterministic, because the likelihood function becomes

$$\mathcal{L}(\vec{\alpha}) = \delta\big(M(X;\vec{\alpha}) - D\big), \tag{3.6}$$

where $\delta(\cdot)$ is the Dirac delta function. In this case, the model output and observed data only agree if their values are exactly equal, i.e. $M(X;\vec{\alpha}) \to D$ for all n points, which in most cases, is only possible if we overfit the data or the model is perfect. Thus, Bayesian regression will usually fail in this case, or if it succeeds, it only produces singular posterior distributions of the model parameters (i.e. $\sigma_{\vec{\alpha}} = 0$). When Bayesian regression fails, the Bayesian evidence is zero, which, although correct (the model does not support/fit the data), may not be the most useful type of answer for the modeler. It seems reasonable that a modeler would want both the benefits of Bayesian and standard regression simultaneously.

3.2.4 BVM Model Testing

We recall the Bayesian Validation Metric (BVM) introduced in Chapter 2. The BVM represents model to data validation in a general way using the probability of agreement,

$$\begin{aligned} p(A|M,D,B) &= \int_{\hat{z},z} p(A|\hat{z},M,z,D,B)\cdot\rho(\hat{z},z|M,D)\,d\hat{z}\,dz \\ &= \int_{\hat{z},z} \rho(\hat{z}|M,D)\cdot\Theta\big(B(\hat{z},z)\big)\cdot\rho(z|D)\,d\hat{z}\,dz, \end{aligned} \tag{3.7}$$

where $\hat{z}$ and z are the model and data comparison quantities, respectively. The "agreement kernel" $p(A|\hat{z},M,z,D) = \Theta\big(B(\hat{z},z)\big)$ is the indicator function of a user defined boolean function, $B(\hat{z},z)$, that defines the context of what is meant by "model to data agreement", by being true when $(\hat{z},z)$ agree or false otherwise. For simplicity, we will assume $\hat{z} \to \hat{y}$ and $z \to y$ are the model output and observed data respectively.

The BVM model testing framework was shown to generalize BMT where the probability of agreement plays the role of the evidence,

$$\mathcal{Z}(B) = p(A|M,D,B) = \int_{\vec{\alpha}} \underbrace{p(A|\vec{\alpha},M,D,B)}_{\mathcal{L}(\vec{\alpha},B)} \cdot \underbrace{\rho(\vec{\alpha}|M)\,d\vec{\alpha}}_{\pi(\vec{\alpha})\,d\vec{\alpha}}, \tag{3.8}$$

where $\mathcal{Z}(B)$ and $\mathcal{L}(\vec{\alpha},B)$ are the BVM evidence and likelihood, respectively, that have been modified by a user's definition of model-data agreement B. Equation (3.8) is a key insight in [58]. Analogous to the Bayesian model testing framework, we can perform BVM model testing between two models M_1 and M_2 using the probability of agreement defined above as follows:

$$R(B) \equiv \frac{p(M_1|A,D,B)}{p(M_2|A,D,B)} = \frac{p(A|M_1,D,B)\,p(M_1|D,B)}{p(A|M_2,D,B)\,p(M_2|D,B)} = \frac{\mathcal{Z}_1(B)}{\mathcal{Z}_2(B)}\frac{p(M_1|D,B)}{p(M_2|D,B)},$$

where $p(M_1|D,B)/p(M_2|D,B)$ is the ratio of prior probabilities of M_1 and M_2, which can often be set to unity, i.e. $p(M_1|D,B)/p(M_2|D,B) = 1$. In this case, we get

$$R(B) \to \frac{p(A|M_1,D,B)}{p(A|M_2,D,B)} = \frac{\mathcal{Z}_1(B)}{\mathcal{Z}_2(B)} = K(B),$$

where $R(B)$ denotes the BVM ratio and $K(B)$ denotes the BVM factor, which is analogous to the Bayes factor (as mentioned earlier in Chapter 2).

3.2.5 The Improved Reliability Metric

The reliability metric discussed in [47] is defined as the probability that the mean of the model prediction is within a tolerance ϵ of the mean of the data. This metric was later expanded in [51] to consider tolerances between each of the model and data point pairs, $|y_j - \hat{y}_j| \leq \epsilon_j$ for $j = 1, ..., n$, rather than comparing their means.

Consider a set of inputs X, a set of model outputs $\hat{Y}$ and a set of observed data points Y. Assume that all the sets are n-dimensional and that the data were collected through independent experiments. The improved reliability metric r_i is,

$$r_i = \int_Y \rho(Y|D) \left(\int_{Y-\epsilon}^{Y+\epsilon} \rho(\hat{Y}|M)\, d\hat{Y} \right) dY, \tag{3.9}$$

which can always be rewritten as,

$$r_i = \int_{\hat{Y},Y} \rho(\hat{Y}|M) \cdot \Theta\Big(\big|\hat{Y} - Y\big| \le \epsilon\Big) \cdot \rho(Y|D)\, d\hat{Y} dY. \tag{3.10}$$

This equation may be identified as a special case of the BVM (3.7) when,

$$\Theta(B(\hat{z}, z)) \to \Theta\big(B(\hat{Y}, Y)\big) = \Theta\Big(\big|\hat{Y} - Y\big| \le \epsilon\Big) = \prod_{j=1}^{n} \Theta\Big(\big|\hat{y}_j - y_j\big| \le \epsilon_j\Big), \tag{3.11}$$

and where $\hat{z} = \hat{Y}$, $z = Y$ (see Appendix A.1). Thus, this agreement kernel is based on the ϵ-Boolean. From (3.8) (see Appendix D for details), the BVM in this case is,

$$\mathscr{Z}(B) = p(A|M, D, B) = \int_{\vec{\alpha}} \bigg(\int_{Y} \Theta\Big(\big|M(X;\vec{\alpha}) - Y\big| \le \epsilon\Big) \cdot \rho(Y|D)\, dY \bigg) \cdot \rho(\vec{\alpha}|M)\, d\vec{\alpha}. \tag{3.12}$$

The ϵ-Boolean participates in several of our BVM regression examples in the following sections and thus reliability is automatically regressed into our model solutions through the improved reliability metric.

3.3 Generalized Bayesian Regression via the BVM

This section introduces BVM regression, which generalizes Bayesian and standard regression. This method has the ability to produce posterior parameter distributions and predictive envelopes for any data distribution, include prior knowledge about model parameters (if there is any), and regularize parameter solutions in a way that parameter uncertainty increases rather than decreases (as discussed in Section 3.2.1).

BVM regression consists of learning the posterior of a set of parameters $\vec{\alpha}$, given the agreement A and the Boolean function B, from the prior via Bayes' Rule,

$$\rho(\vec{\alpha}|M) \xrightarrow{*} \mathcal{P}(\vec{\alpha}|A) \equiv \rho(\vec{\alpha}|A, M, D, B) = \frac{p(A|\vec{\alpha}, M, D, B)\, \rho(\vec{\alpha}|M)}{p(A|M, D, B)}. \tag{3.13}$$

After learning the posterior distribution of the model parameters, we can evaluate the predictive distribution defined by:

$$p(\hat{y}|A,M,B) = \int_{\vec{\alpha}} p(\hat{y}|\vec{\alpha},A,M,B)\cdot\rho(\vec{\alpha}|A,M,D,B)\,d\vec{\alpha}. \tag{3.14}$$

Performing BVM regression requires evaluating the BVM probability of agreement. At the beginning of Appendix D, we give a derivation showing that (3.8) can be written as,

$$\mathcal{Z}(B) = p(A|M,D,B) = \int_{\vec{\alpha}}\left(\int_{Y}\Theta\Big(B\big(M(X;\vec{\alpha}),Y\big)\Big)\cdot\rho(Y|D)\,dY\right)\cdot\rho(\vec{\alpha}|M)\,d\vec{\alpha}, \tag{3.15}$$

which is analogous to (3.3) in form,[1] and where the comparison values are $\hat{z} = \hat{Y} = M(X;\vec{\alpha})$ and $z = Y$ (we assume we are dealing with a set of inputs, model outputs (i.e. a set of n points on a model curve), and n observed data points, where all the data points were collected through independent experiments).

BVM regression can reproduce Bayesian regression, standard regression, and likelihood-based methods as special cases. When the data and model outputs must be exactly equal to agree with one another $\big($i.e. $\delta(\hat{Y}-Y)\big)$, the BVM produces BMT as a special case and the regression solutions are given in Appendix C. Typical likelihood-based methods follow from the same "exactly equal" definition of model-data agreement. We find that the boolean function $B_{S.R}(\hat{Y}(\vec{\alpha}^*),Y)$ that reproduces standard regression is defined to be true iff $\vec{\alpha}^* = \underset{\vec{\alpha}}{\text{argmin}}\,\mathcal{E}\big(M(X;\vec{\alpha}),Y\big)$ for some objective function $\mathcal{E}(\cdot)$. This only gives nonsingular posterior parameter distributions and predictive model envelopes if the data is uncertain and/or if $\mathcal{E}\big(M(X;\vec{\alpha}),Y\big)$ does not have a unique global minimum.

If the objective function is convex, then we have a single minimum which results in one vector of parameters $\vec{\alpha}^*$ that makes $B_{S.R}(\hat{Y}(\vec{\alpha}^*),Y)$ true. However, when the cost function is non-convex, then multiple parameter vectors $\vec{\alpha}^*$ corresponding to different local minima, lead to a true $B_{S.R}(\hat{Y}(\vec{\alpha}^*),Y)$ and may be accepted due to the approximate nature of non-convex optimization methods. This results in multiple regressed solutions for the regression problem and approximates the posterior

[1] In terms of the BVM, the Bayesian evidence in BMT, i.e. Equation (3.3), may be interpreted as the probability that the uncertain data and model output are exactly equal, i.e. $\rho(D|M)\equiv\rho(\hat{Y}\equiv Y|M,D)\equiv\rho(A|M,D)$ which is Equation (C.1) derived in Appendix C.

parameters' distribution $\rho(\vec{\alpha}|A)$ (analogous to the accepted parameter samples in the Markov Chain Monte Carlo (MCMC) simulation). Marginalizing leads to the predictive posterior model distribution $p(\hat{y}|A, M, B)$ as in (3.14). Finding the predictive model output average is analogous to the results obtained in the ensemble methods in machine learning [8]. Because the BVM can reproduce these special cases and generate new ones by extending, combining, and modulating Boolean agreement functions, BVM regression may be seen as a generalized regression method.

Due to the flexibility of the BVM framework, there are many possible definitions of agreement that the user can define. We discussed some of these definitions in Chapter 2. We summarize them in Table 3.1 below.

	Agreement Boolean Function
$\epsilon-$Boolean	True iff $\big|y_j - M(x_j;\vec{\alpha})\big| \le \epsilon_j \ \forall \ j$
$(\gamma, \epsilon, \ell)-$Boolean	True iff $\big|y_j - M(x_j;\vec{\alpha})\big| \le \ell\epsilon_j \ \forall \ j$ and $\frac{1}{n}\sum_j \Theta\Big(\big|y_j - M(x_j;\vec{\alpha})\big| \le \epsilon_j\Big) \ge \gamma\%$
$\langle\epsilon\rangle-$Boolean	True iff $\frac{1}{n}\sum_j \big|y_j - M(x_j;\vec{\alpha})\big| \le \langle\epsilon\rangle$
$(\langle\epsilon\rangle, \hat{\alpha})-$Boolean	True iff $\frac{1}{n}\sum_j \big|y_j - M(x_j;\vec{\alpha})\big| \le \langle\epsilon\rangle$ and $0.91 \le \frac{1}{n}\sum_j \Theta(y_j \in [-c_{\hat{\alpha}}, c_{\hat{\alpha}}]_j) \le 0.99$

Table 3.1: Some examples of agreement Boolean functions.

To address the concerns we raised about Bayesian and standard regression depicted in Figure 1-1, consider using the $\epsilon-$Boolean with the agreement kernel,

$$\Theta\Big(B\big(M(x_j;\vec{\alpha}), y_j\big)\Big) = \begin{cases} 1, & \text{if } \big|y_j - M(x_j;\vec{\alpha})\big| \le \epsilon_j \\ 0, & \text{otherwise} \end{cases}$$

for all $j = 1, \ldots, n$, where ϵ_j may be adjusted and tuned to impose more or less strict agreement conditions which may be used by the modeler to enforce reliability in some region or to be more tolerant of training errors at instance x_j. For simplicity, We assume that $\epsilon_j = \epsilon$ for all j. Note that this is (3.11) in Section 3.2.5. In other words, we use the special case of the BVM, the improved reliability metric with evidence derived in (3.12), to derive our theoretical solutions.[2] Utilizing this BVM definition allows us to solve the truncated tail data distributions problem in Bayesian regression

[2] Note that by choosing to adopt a different agreement kernel (or Boolean function) as in Section 3.4.2, we generalize the results derived above; this is the power of the BVM.

in a simple way – details in Appendix D.[3]

Truncated Tail Data Distributions Solution Summary

Let the data be known to have the truncated pdf $y_j \sim \mathcal{U}(a_j, b_j)$, for $j = 1, \ldots, n$. By using the ϵ-Boolean, we introduce leniency into the regression in that it no longer needs to exactly pass through all intervals $[a_j, b_j]$ simultaneously to count as a "fit". This produces likelihood functions such as,

$$\mathcal{L}(\vec{\alpha}, B) = \prod_{j=1}^{n} \frac{u_j - l_j}{b_j - a_j}, \tag{3.16}$$

where l_j and u_j are defined by the boundaries of the intersection of the data uncertainty and the model's tolerance ϵ,

$$\left[l_j,\ u_j\right] = \left[M(x_j;\vec{\alpha}) - \epsilon,\ M(x_j;\vec{\alpha}) + \epsilon\right] \cap \left[a_j,\ b_j\right] \qquad j = 1, \ldots, n.$$

An illustration of how the BVM works with truncated data distributions is shown in Figure 3-1 below. For example, at instance x_2, the interval $\left[l_2,\ u_2\right]$ is found by intersecting the intervals $\left[a_2,\ b_2\right]$ and $\left[M(x_2;\vec{\alpha}) - \epsilon,\ M(x_2;\vec{\alpha}) + \epsilon\right]$. Note that this applies to the instances x_j for all j. In this case, the likelihood is nonzero, resulting in a nonzero evidence (3.15). Thus, given this agreement definition, the probability of finding a model given the truncated data is nonzero.

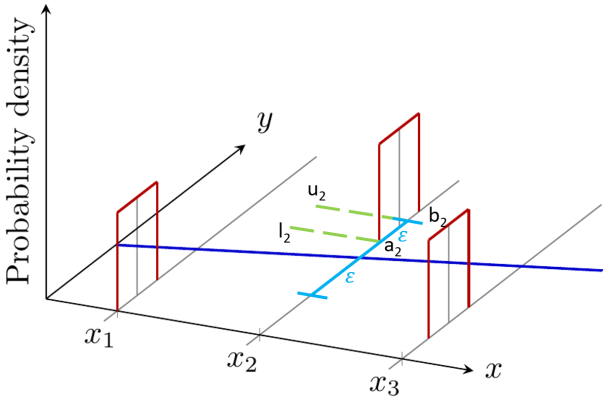


Figure 3-1: **Truncated tail data distributions solution.** Using BVM regression results in a nonzero probability of finding a model given the observed truncated data.

[3] A complete derivation for the infinite tail Gaussian data distribution is given in Appendix D.1.

Now, if we consider the special case when the data is completely certain, deterministic, i.e. $Y = D$, then the likelihood function is

$$\mathcal{L}(\vec{\alpha}, B) = \Theta\Big(B\big(M(X;\vec{\alpha}), D\big)\Big), \tag{3.17}$$

which can be seen as a relaxed general form of the delta function adopted in the Bayesian model testing (where $\epsilon = 0$), which implies that the model output must be within ϵ from the observed measurements in order for them to agree. An analogous ϵ-Boolean solution exists for standard regression methods which leads to nonsingular parameter distributions whether the regression is regularized or not.

Tolerant agreement as a new kind of regularization

The purpose of regularization is to better represent one's expectations of unobserved data using the chosen model or model class. Using BVM regression and nonzero agreement tolerances (e.g. $\epsilon > 0$ in the ϵ-Boolean), we can broaden the model's prediction envelope to better represent our expectations of the data. Increasing agreement tolerances naturally increases the posterior variance of the parameters, which differs from standard regularization methods and can be used to avoid conceptual issues of interpreting regularized physical parameters. It should also be noted that this is done without changing the prior distributions of the parameters nor the given probability distributions of the data. This becomes a useful feature in our first example.

3.4 Implementation and Examples

3.4.1 Computing the BVM Evidence

Like the Bayesian evidence, the BVM evidence is computationally expensive to calculate when one has many model parameters to learn. Several approaches were adopted to solve this problem. Markov Chain Monte Carlo (MCMC) is a computational technique used for Bayesian methods that has been widely studied and improved [6, 17, 35, 37, 42, 48] as it is considered an indispensable tool for Bayesian inference.

Other techniques include the Nested Sampling method [10, 54] and the MultiNest algorithm [11].

We will approximate the BVM evidence and generate the posterior model parameter distributions (for the purpose of generating model's predictive envelopes) using MCMC. MCMC takes the following inputs: the likelihood function $\mathcal{L}$ and the prior distribution π of the model parameters, a model M and a set of input/output data points $\{X, Y\}$. The Bayesian terms $(\mathcal{Z}, \mathcal{L})$ have analogous BVM terms $(\mathcal{Z}(B), \mathcal{L}(B))$ and it is therefore straightforward to extend the MCMC algorithm to BVM calculations to obtain posterior parameter distributions. These distributions are used to generate a model's predictive envelope. Any adaptation to the standard MCMC algorithm will be discussed in text with their corresponding example.

3.4.2 BVM Regression Examples

Exploratory Example 1

We consider the case study investigated in [3] using a bacterial growth model. The data is obtained by operating a continuous flow biological reactor at steady-state conditions. The observations are as follows:

$\boldsymbol{x}$ **(mg/L COD)**	28	55	83	110	138	225	375
$\boldsymbol{y}$ **(1/h)**	0.053	0.060	0.112	0.105	0.099	0.122	0.125

Table 3.2: The observations we aim to fit.

where y is the growth rate at substrate concentration x. We replicate the results found in [3] using the nonlinear Monod model to fit the data, i.e,

$$\hat{y} = M(x; \vec{\alpha}) = \frac{\alpha_1 x}{\alpha_2 + x} \tag{3.18}$$

where α_1 is the maximum growth rate (h^{-1}: per hour), and α_2 is the saturation constant (mg/L COD: the Chemical Oxygen Demand, measured in milligrams per liter).

We run MCMC on the likelihoods derived in (3.16) and (3.17) corresponding to the different types of data distributions discussed above, i.e. bounded or truncated uniform data distributions, and completely certain observation points (we also consider normal or Gaussian data distributions with infinite tails). We assume that the vector of parameters $\vec{\alpha}$ has some Gaussian prior distribution. We use the $\epsilon-$Boolean agreement function and find that BVM regression is able to construct posterior inferences of the model parameters for each type of data measurement distributions, unlike Bayesian model testing and standard regression techniques that fail at this task for truncated and completely certain data, as discussed before. We summarize the results in Table 3.3 below.

Data Distribution	BVM regression	Standard regression	Bayesian regression
Infinite Tail	✓	✓	✓
Truncated Tail	✓	✓	✗
Completely Certain	✓	✗	✗

Table 3.3: High probability of the model producing posterior parameter distributions and predictive envelopes for different types of data distributions using the three approaches. BVM regression is capable of producing posterior distributions of the model parameters for any type of data distributions.

Using the BVM regressed parameters' distributions, we can make predictions of y for new values of x, i.e., $p(\hat{y}|A, M, B)$ from (3.14). In addition, instead of just computing a point estimate of the fit, we should also study the predictive posterior distribution of the model, (also called the predictive envelope). As an illustration of the predictive posterior distribution of our BVM regressed model, we plot the predictive envelopes of the nonlinear Monod model described in (3.18), treating the data as completely certain and using the ϵ-Boolean function with a tolerance $\epsilon = 0.03$.

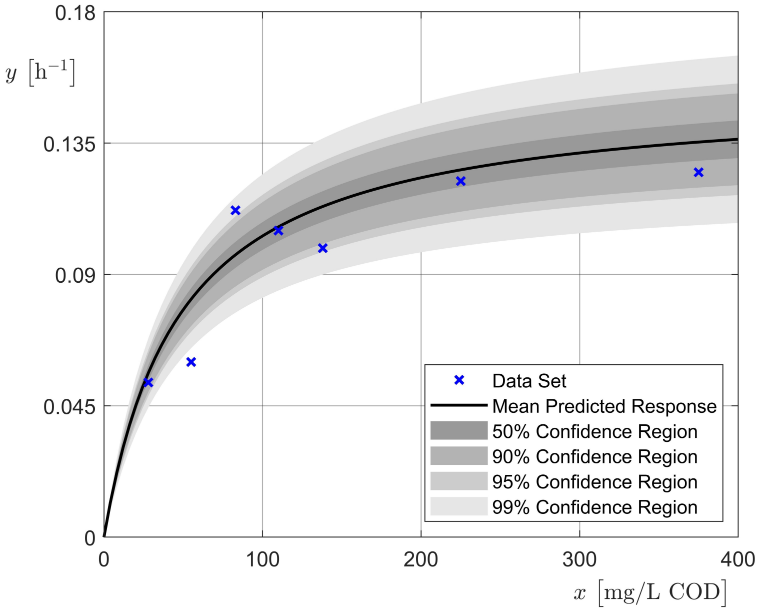


Figure 3-2: **Predictive envelopes of the model in the absence of data uncertainty using the BVM.** As tabulated in Table 3.3, Bayesian regression fails to produce a candidate model solution as the data is completely certain and standard regression produces a single deterministic solution with no model uncertainty.

The black curve shows the predicted response, which is the model fit calculated using the mean values of the parameters α_1 and α_2 in the chain. The gray shaded areas correspond to 50%, 90%, 95%, and 99% predictive posterior regions (by computing the model fit for a randomly selected subset of the chain). In other words, the gray regions span 0.675, 1.645, 2, 3 standard deviations on either side of the mean response, respectively. We will leave the interpretation of the predictive envelopes for our compound Boolean agreement function example in Section 3.4.2.

The value of the tolerance ϵ chosen affects the shape of the model parameters' distributions and thus the predictive envelope. A smaller tolerance implies stricter agreement conditions between the model response and the observed data, which results in less uncertainty in the predictive posterior distributions of the model parameters and a narrower envelope. On the other hand, a larger tolerance implies a more flexible agreement conditions, and results in more uncertainty in the predictive distributions, a wider envelope and a less predictive power. Thus, increasing ϵ can always result in finding a model given the data. To avoid getting very wide envelopes relative to the spread of the data, we start with a very small ϵ when running the MCMC simulation. We then keep increasing ϵ until the MCMC algorithm starts achieving a reasonably small acceptance rate for the new candidates in the chain.

Since this model has just two adaptive parameters, namely α_1 and α_2, we can plot the prior and posterior distributions directly in parameter space. We explore the dependence between the parameters' posterior distributions and the value of the tolerance ϵ. Figure 3-3 shows the results of BVM learning for the Monod model in (3.18) as the value of ϵ is decreased. For comparison, the optimal parameter values $\alpha_1 = 0.14542$ and $\alpha_2 = 49.053$ computed using standard regression are shown by a yellow cross in the first row of Figure 3-3.

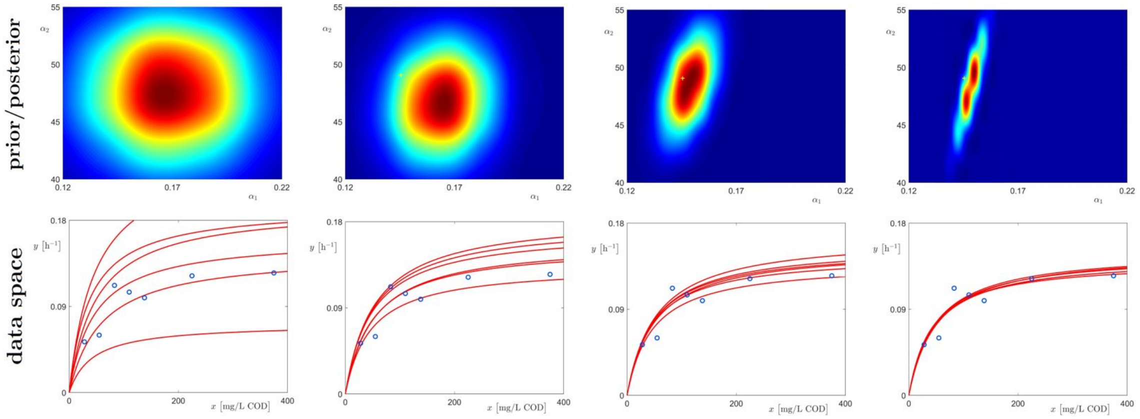


Figure 3-3: **Illustration of BVM Learning for the Monod model for decreasing values of ϵ.** In the first row is the prior/posterior parameter distribution in (α_1, α_2) space. The data points are shown by a blue circle in the second row. The first column corresponds to the situation before any data point is observed and shows a plot of the prior distribution in (α_1, α_2) space together with six samples of the model response $M(x; \vec{\alpha})$ (red lines) in which the values of α_1 and α_2 are randomly drawn from the prior. In the second, third and fourth columns, we see the situation after running our BVM learning using MCMC, with a tolerance $\epsilon = 0.03$, $\epsilon = 0.025$ and $\epsilon = 0.02$, respectively. The posterior has now been influenced by the agreement tolerance ϵ, this gives a relatively compact posterior distribution. Samples from this posterior distribution lead to the functions shown in red in the second row.

As Figure 3-3 shows, the smaller the tolerance is, the narrower and sharper the posterior distributions of the parameters are, the closer the red lines get to each other, and the lower the uncertainty is. This explains the shape of the predictive envelopes as was discussed before. Thus, by varying ϵ, one can tune the model response posterior distribution to be more or less representative of the data. We will elaborate more on this in Section 3.4.3. Note that in our example, when ϵ goes below about 0.017, no solution seems to be possible and hence the probability of finding a model given the observed data becomes zero. In this case, the analyst may choose to work with any tolerance beyond this threshold, depending on his specifications and agreement requirements.

Once we generate the posterior distributions of the model parameters and the predictive envelopes, we can measure and estimate the reliability and accuracy of our computational model using a validation metric (including the BVM). By doing so, we determine how accurate is our model representation of the real world.

Exploratory Example 2

After showing how the BVM can be used to perform regression on any type of data distribution to generate posterior model parameters' distributions and predictive envelopes, we now focus on how the user can choose the Boolean function to define the model-data agreement.

In this example, we will use the compound Boolean as presented in Chapter 2, Section 2.6.2.[4] In that case, the definition of agreement requires the model to pass an average square error threshold of $\langle\epsilon\rangle$ as well as a check for probabilistic model configuration. The latter states that $95\% \pm 4\%$ of the uncertain observations (data) should lie inside the model's $1-\hat{\alpha} = 95\%$ confidence interval. Note that we impose the $\pm 4\%$ tolerance to prevent the scenario where all 100% of the data points lie within an overly wide confidence interval, being marked as "agreeing". We denote this compound Boolean function by $B(\hat{Y}, Y, \langle\epsilon\rangle, \hat{\alpha})$ and it is equal to,

$$B\left(\frac{1}{n}\sum_i |\hat{y}_i - y_i| \leq \langle\epsilon\rangle\right) \wedge B\left(0.91 \leq \frac{1}{n}\sum_i \Theta(y_i \in [-c_{\hat{\alpha}}, c_{\hat{\alpha}}]_i) \leq 0.99\right), \tag{3.19}$$

where n is the number of data points in the set Y, and $[-c_{\hat{\alpha}}, c_{\hat{\alpha}}]_i$ is the model's 95% confidence interval at instance x_i (this is the $(\langle\epsilon\rangle, \hat{\alpha})$–Boolean function in Table 3.1).

[4] It should be noted that in Chapter 2, a model is simply validated according to this compound metric – here we calibrate the model with respect to it instead.

Note that, although this compound Boolean seems to be complex, it is relatively easy to code and implement.

The BVM probability of agreement in this case can be expressed as,

$$\mathcal{Z}(B) = p(A|M, D, B, \langle\epsilon\rangle, \hat{\alpha}) = \int_{\vec{\alpha}} \underbrace{\left(\int_Y \Theta\Big(B\big(M(X; \vec{\alpha}), Y, \langle\epsilon\rangle, \hat{\alpha}\big)\Big) \cdot \rho(Y|D)\, dY \right)}_{\mathcal{L}(\vec{\alpha}, B)} \cdot \rho(\vec{\alpha}|M) d\vec{\alpha}$$

Note that the likelihood $\mathcal{L}(\vec{\alpha}, B)$ can be expressed as an expectation value over $\rho(Y|D)$,

$$\mathcal{L}(\vec{\alpha}, B) = E\Big[\Theta\Big(B\big(M(X; \vec{\alpha}), Y, \langle\epsilon\rangle, \hat{\alpha}\big)\Big)\Big] \sim \frac{1}{K}\sum_{k=1}^{K} \Theta\Big(B\big(M(X; \vec{\alpha}), Y^{(k)}, \langle\epsilon\rangle, \hat{\alpha}\big)\Big), \quad (3.20)$$

where $Y^{(k)}$ denotes the k^{th} set of data points drawn randomly from the probability distribution of Y. This allows us to approximate the integral using a statistical method like Monte-Carlo (MC). In this example, we use MC with $K = 50$.

We implement the compound Boolean, $B(\hat{Y}, Y, \langle\epsilon\rangle, \hat{\alpha})$, and show its ability to combine and quantify the average error as well as the probabilistic model representation of the uncertain data observations. We generated data using,

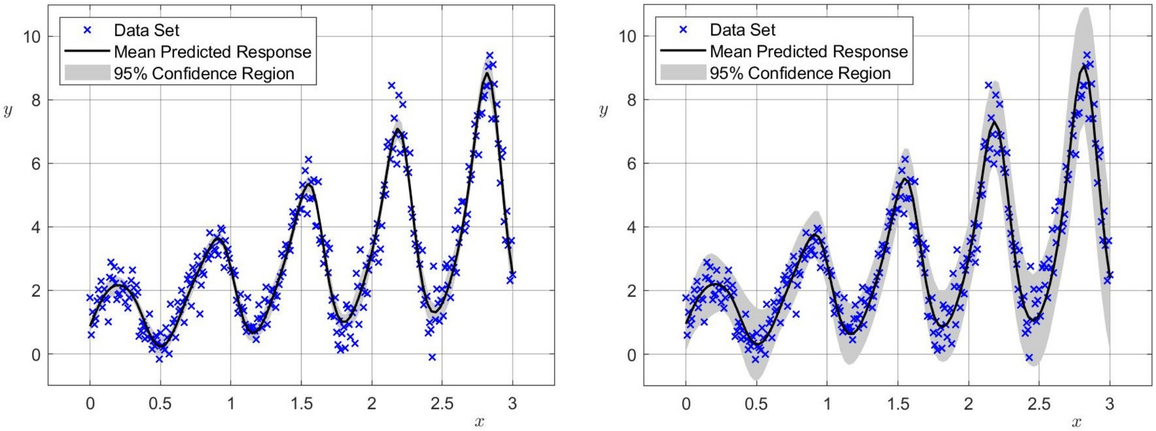
$$y(x) = 1 + xe^{-\cos(10x)} + \sin(10x) + \epsilon_a(x),$$

where $\epsilon_a(x) \sim \mathcal{N}(0, 0.4^2)$ for $x \in [0, 1.5]$ and $\epsilon_a(x) \sim \mathcal{N}(0, 0.6^2)$ for $x \in [1.5, 3]$, which represents the aleatoric stochastic uncertainty due to the system's randomness. We also assume the presence of epistemic measurement uncertainty in the data [50] with an additional normal distribution $\mathcal{N}(0, 0.5^2)$ about each data point.

To solve this example, we consider the following deterministic non-linear model,

$$\hat{y} = M(x; \vec{\alpha}) = \alpha_1 + \alpha_2 x e^{-\alpha_3 \cos(\alpha_4 x)} + \alpha_5 \sin(\alpha_6 x), \quad (3.21)$$

where $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)$ is the vector of model parameters having normally distributed prior distributions with means $\mu_{\vec{\alpha}} = (0, 0, 0, 9, 0, 9)$ and standard deviations $\sigma_{\vec{\alpha}} = (1, 1, 1, 0.5, 1, 0.5)$. We then conduct two experimental simulations. We first run the MCMC algorithm for 5000 iterations with 10% burn-in using Bayesian regression and plot the results in Figure 3-4a. We then repeat the simulation using the approximate likelihood $\mathcal{L}(\vec{\alpha}, B)$ from (3.20) with a threshold of $\langle\epsilon\rangle = 0.7$. The results are shown in Figure 3-4b.



(a) Bayesian regression.

(b) BVM regression.

Figure 3-4: **Comparison between Bayesian regression and BVM regression.** (a) Bayesian regression under infinite tail data distribution. Note that the 95% confidence interval is very narrow and standard regression method produces a nearly identical result. (b) BVM regression using the compound Boolean. In this case, the 95% confidence region is much wider and represents the data more accurately. Note that this probabilistic model passes both agreement conditions imposed by the compound Boolean $B(\hat{Y}, Y, \langle\epsilon\rangle, \hat{\alpha})$. Starting with a very small $\langle\epsilon\rangle$ in the MCMC simulation, we tune $\langle\epsilon\rangle$ by gradually increasing its value until both elements of the compound Boolean are naturally satisfied.

The BVM regression framework offers new insights into the interpretation of the predictive envelopes of Bayesian and standard regression. It is clear in Figure 3-4a that the Bayesian and standard regression methods generate predictive envelopes that would not accurately predict new target points. Surprisingly, these envelopes actually quantify the uncertainties in the *least square error solution* due to the presence of data uncertainty rather than a measure of predictive uncertainty. By being careful in how we define the model-data agreement (as in Figure 3-4b), we were able to construct predictive envelopes that satisfy our desire in representing new target points probabilistically. In other words, using the BVM regression framework gives the user more control over the predictive envelopes and what uncertainties they represent.

3.4.3 Discussion

As we have shown in the exploratory examples above, the results are sensitive to the tolerance ϵ. This arises from the fact that ϵ represents the modeler's tolerance on the difference between the model outputs and the observed data. Thus, any other parameter that is included in the Boolean function B defined by the modeler will have an effect on the parameters' posterior distributions and the predictive envelopes. As shown in the previous examples, we can take advantage of this feature for modeling.

We analyse the effect of the tolerance ϵ as follows. Changing the agreement tolerance ϵ affects the acceptance rate in the MCMC iterations, i.e. a larger tolerance yields a larger variance of accepted "candidate" samples. The increased variance in the accepted samples produces wider posterior model parameter distributions. A smaller tolerance implies the converse.

Parameters not regressed inside the Boolean function B (ϵ in this case) play a role similar to hyperparameters in Machine Learning. By tuning the hyperparameters, the modeler can make the predictive envelopes more representative of the data, and improve the overall performance of the model when compared to a randomly selected test (or validation) data set.

Note that in the case of the ϵ–Boolean, the user can increase ϵ indefinitely and still "regress" the model; although the predictive envelopes will be much wider than the data spread and hence less representative of the data. While widening the predictive envelopes can be useful for reliability and safety, if they are widened too much, the model can lose some of its predictive utility. To balance the trade-off between safety and utility, we regressed the model in (3.21) with respect to the compound Boolean (3.19) in Section 3.4.2. This Boolean forced the model to be regressed in such a way that $95\%\pm4\%$ of the (uncertain) data points lie within the model's 95% confidence region while simultaneously satisfying the $\langle\epsilon\rangle$ requirement (for diversity we used $\langle\epsilon\rangle$ instead of ϵ although nothing prevents us from doing so in principle). As stated in the caption of Figure 3-4, we gradually tuned the value of $\langle\epsilon\rangle$ (hyperparameter tuning) to simultaneously satisfy both requirements imposed by the compound Boolean function.

3.5 Conclusion

This chapter presents a generalized Bayesian regression and model learning technique that is capable of probabilistically regressing (learning) model parameter distributions while satisfying arbitrary definitions of model-data agreement within the BVM framework. Using this technique, we can perform Bayesian regression based on any type of data distribution and construct predictive envelopes that are more representative of the data, which improves the overall performance of the model under question.

Chapter 4

Conclusions and Recommendations

4.1 Conclusions

This thesis presents the BVM, a general model validation and testing tool. We demonstrated the versatility of the BVM toward expressing and solving model validation and calibration problems. The BVM quantifies the probability that a model is valid for arbitrary quantifiable definitions of model-data agreement using arbitrary comparison functions of the model-data comparison values. The BVM was shown: to obey all of the desired validation metric criteria [28] (which is a first), to be able to represent all of the standard validation metrics as *special cases*, to supply improvements and generalizations to those special cases, and to be a tool for quantifying the validity of a model in novel model-data contexts. The latter was demonstrated by the validation metrics we invented and quantified in our examples.

In addition, it was shown that one can perform model selection using the BVM ratio. The BVM model testing framework was shown to generalize the Bayesian model testing framework to arbitrary model-data contexts and with reference to arbitrary comparisons and agreement definitions. That is, the BVM ratio may be used to rank models directly in terms of the relevant model-data validation context. The problem of model-data validation may be reduced to the problem of finding/defining the four BVM inputs: $(\hat{z}, z)$, $\rho(\hat{z}, z|M, D)$, $f(\hat{z}, z)$, and $B(f)$, and computing their BVM value. We find that the BVM is a useful tool for performing model validation and testing.

Finally, the BVM framework can be expanded to probabilistically regress model parameter distributions that satisfy arbitrary definitions of agreement. Particularly, in the calibration stage of model development, we can use the BVM to perform regression and model learning on data with any type of uncertainty, generate posterior parameter distributions, and model predictive envelopes, according to user-specified definitions of model-data agreement. The BVM regression framework proved its potential in offering new insights into the interpretation of the predictive envelopes of the Bayesian regression, standard regression, and likelihood-based techniques, and hence providing the analyst with more freedom and control over the predictive envelopes and their meaning. In short, we find the BVM to constitute a generalized framework for probabilistic model calibration and validation allowing us to address several potential shortcomings in the calibration and validation techniques found in the literature.

4.2 Recommendations

We provide some suggestions for researchers who are interested in building on the contributions discussed in this thesis to improve and advance this area of research.

All the work presented in this thesis relies heavily on parametric models. Thus, one possible future research topic is to expand the BVM framework to nonparametric regression (e.g. Gaussian processes). Another line of research involves developing neural networks within the BVM framework. Particularly, one can investigate and explore ways to integrate the BVM probability of agreement into the loss function of the neural networks.

Because the BVM is an open framework where the definition of agreement is composable and user-defined, there is room for the quantification of further agreement/validation requirements as the need arises in data analysis and model reliability.

Finally, we emphasize the importance of practicing statistical responsibility by being explicit in the definition of agreement between the models and data in the field of reliability. The BVM framework forces the modeler to be explicit in their definitions, assumptions, and criteria when performing model calibration and validation.

Appendix A

Representing the Known Validation Metrics with the BVM

In the following sections, we will show some of the special cases of the Bayesian Validation Metric (BVM). Subsequent improvements or immediate generalizations of the metrics using (2.2) are presented when applicable. A detailed review of the majority of these metrics may be found in [28] and the references therein. Tables 2.1 and 2.2 in Section 2.5 outline the results.

A.1 Reliability Metric and Probability of Agreement

There are a few validation metrics related to the reliability metric present in the literature. The reliability metric $r = p(|\langle\hat{y}\rangle - \mu_y| < \epsilon)$ [47] is equal to the probability that the data and the model expectation values are within a tolerance of size ϵ. Their "probability of agreement" introduced in [57] is closely related to r, but instead expresses the quantity as "the probability the data and the model expectation values *agree* within a tolerance (or sliding tolerance) of ϵ". The reliability metric was expanded in [51] to account for model outputs and data rather than simply comparing the mean of the model prediction against the mean of the data. The improved reliability metric is equal to,

$$r_i = \int_{-\infty}^{\infty} \rho(Y|D) \int_{Y-\epsilon(Y)}^{Y+\epsilon(Y)} \rho(\hat{Y}|M)\, d\hat{Y}\, dY, \tag{A.1}$$

where $\rho(\hat{Y}|M)$ and $\rho(Y|D)$ are the full joint probability distributions of the model outputs and the data, respectively. This metric quantifies the probability that the error is less than a value $\epsilon(y)$ on a point to point basis.

The BVM is the reliability metric when the comparison values are $\hat{z} = \langle \hat{y} \rangle$ and $z = \mu_y$, and B takes the form of an inequality, being true if $-\epsilon \leq \langle \hat{y} \rangle - \mu_y \leq \epsilon$. If we would like to use a sliding interval of "tolerance" or "error acceptance", denote it by $\big[c_-(\mu_y), c_+(\mu_y)\big]$, where $c_- < c_+$, and the BVM is,

$$\begin{aligned} p(A|M,D) &= \int_{\langle \hat{y} \rangle, \mu_y} \rho(\langle \hat{y} \rangle|M) \cdot \Theta\Big(c_-(\mu_y) \leq \langle \hat{y} \rangle \leq c_+(\mu_y)\Big) \cdot \rho(\mu_y|D)\, d\langle \hat{y} \rangle\, d\mu_y \\ &= \int_{\mu_y} \int_{\langle \hat{y} \rangle = c_-(\mu_y)}^{c_+(\mu_y)} \rho(\langle \hat{y} \rangle|M) \cdot \rho(\mu_y|D)\, d\langle \hat{y} \rangle\, d\mu_y = r. \end{aligned} \tag{A.2}$$

This is the reliability metric if $c_\pm = \mu_y \pm \epsilon$ is a constant and the sliding interval is symmetric about μ_y.

The BVM is the improved reliability metric when $\hat{z} = \hat{Y}$, $z = Y$, and when the Boolean $B(\hat{z}, z)$ is true iff $|\hat{y} - y| \leq \epsilon(y)$ for all $\hat{y}, y$ pairs. That is,

$$p(A|M,D) = \int_{\hat{Y},Y} \rho(\hat{Y}|D) \cdot \rho(Y|M) \cdot \Big(\prod_{\hat{y},y \text{ pairs}} \Theta\big(|\hat{y} - y| \leq \epsilon(y)\big)\Big)\, d\hat{Y}\, dY = r_i. \tag{A.3}$$

The BVM quantifies the probability of square error (or difference) is less than some ϵ by considering,

$$p(A|M,D) = \int_{\hat{Y},Y} \rho(\hat{Y}|D) \cdot \Theta\big(|\hat{Y} - Y| \leq \epsilon\big) \cdot \rho(Y|M)\, d\hat{Y}\, dY. \tag{A.4}$$

Nothing in the BVM requires the variables to be continuous or ordered, so the natural generalization is to let the Boolean expression be true if "The value $\hat{z}$ is in the subset $S(z)$, which is the set of $\hat{z}$'s agreeing with z". For example, if $\hat{z}$'s are strings,

$S(z)$ might be the set of words or phrases in $\hat{z}$ that are reasonably synonymous with z. This gives the straightforward generalization to accommodate arbitrary data types,

$$p(A|M,D) = \sum_{\hat{z},z} p(\hat{z}|M) \cdot \Theta\big(\hat{z} \in S(z)\big) \cdot p(z|D) = \sum_{z} \sum_{\hat{z}\in S(z)} p(\hat{z}|M) \cdot p(z|D), \quad \text{(A.5)}$$

by using sets rather than intervals.

A.2 Frequentist Validation Metric

To include the frequentist validation metric in the BVM, we will have to express the comparison variables $\hat{z}$ and z and their respective probabilities. The result can be replicated by letting: $z = \mu_y$ be the Student's t-distribution and $\hat{z} = \langle\hat{y}\rangle$ have a Dirac delta distribution $\rho(\hat{z}|M,D) = \rho(\langle\hat{y}\rangle|M,D) = \delta\big(\langle\hat{y}\rangle - \langle\hat{y}\rangle'\big)$ where $\langle\hat{y}\rangle'$ is the known value of the computational model's expected output. Because the frequentist validation metric does not force the modeler to define what is meant by *agreement*, we represent this freedom by keeping $B(\hat{z},z)$ general. This gives,

$$\begin{aligned} p(A|M,D) &= \int_{\hat{z},z} \rho(\hat{z}|M,D) \cdot \Theta\big(B(\hat{z},z)\big) \cdot \rho(z|D)\, d\hat{z}\, dz \\ &= \int_{\langle\hat{y}\rangle,\mu_y} \delta\big(\langle\hat{y}\rangle - \langle\hat{y}\rangle'\big) \cdot \Theta\Big(B\big(\langle\hat{y}\rangle,\mu_y\big)\Big) \cdot \left(\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{(\mu_y - \overline{y})^2}{\nu\overline{s}^2/N}\right)^{-\frac{\nu+1}{2}} \right) d\langle\hat{y}\rangle\, d\mu_y, \end{aligned}$$

where $\overline{y}$ (the population average), $\overline{s}$ (the population standard deviation), and ν (the degrees of freedom) are the parameters of the data's t-distribution. Making the coordinate transformations $\langle\hat{y}\rangle \to \overline{E} = \langle\hat{y}\rangle - \overline{y}$ and $\mu_y \to E = \langle\hat{y}\rangle - \mu_y$ gives,

$$\begin{aligned} p(A|M,D) &= \int_{\overline{E},E} \delta\big(\overline{E} - \overline{E}'\big) \cdot \Theta\Big(B\big(\overline{E},E\big)\Big) \cdot \left(\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{(E - \overline{E})^2}{\nu\overline{s}^2/N}\right)^{-\frac{\nu+1}{2}} \right) d\overline{E}\, dE \\ &= \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \int_{E'} \Theta\Big(B\big(\overline{E}',E'\big)\Big) \cdot \left(1 + \frac{(E' - \overline{E}')^2}{\nu\overline{s}^2/N}\right)^{-\frac{\nu+1}{2}} dE', \qquad \text{(A.6)} \end{aligned}$$

with $\overline{E}' = \langle\hat{y}\rangle' - \overline{y}$ and $E' = \langle\hat{y}\rangle' - \mu_y$. Given that judgments of agreement in the frequentist validation metric are expected to be made based on the confidence level $1 - \alpha$ that E' is within the confidence interval, this may be factored into $B\big(\overline{E}',E'\big) \to B\big(\overline{E}',E',\alpha\big)$, as well as other user-defined terms toward expressing agreement. Equation (A.6) is thought to be the full BVM representation of the frequentist validation metric.

The immediate generalization to the frequentist validation metric offered by the BVM is to let the model output expectation value $\hat{z} = \langle \hat{y} \rangle$ have some amount of uncertainty. The uncertainty is perhaps Gaussian or Student's t-distributed in $\mu_{\langle \hat{y} \rangle}$ (the true model expectation value) due to only having a finite number of Monte Carlo samples and/or uncertainty induced by discretization error. Because $\hat{z} = \mu_{\langle \hat{y} \rangle}$ and $z = \mu_y$ are both uncertain in general, one generalizes to,

$$p(A|M, D, B) = \int_{\mu_{\langle \hat{y} \rangle}, \mu_y} \rho(\mu_{\langle \hat{y} \rangle}|M, D) \cdot \Theta\Big(B\big(\mu_{\langle \hat{y} \rangle}, \mu_y\big)\Big) \cdot \rho(\mu_y|D)\, d\mu_{\langle \hat{y} \rangle}\, d\mu_y. \quad \text{(A.7)}$$

It is interesting to note the consequence of defining a reasonable Boolean expression of agreement on the BVM representation of the frequentist metric. A natural agreement function B for the metric is one that is true if $E_\mu = |\mu_{\langle \hat{y} \rangle} - \mu_y| \leq \epsilon$.

The BVM then gives,

$$\begin{aligned} p(A|M, D) &= \int_{\mu_{\langle \hat{y} \rangle}, \mu_y} \rho(\mu_{\langle \hat{y} \rangle}|M, D) \cdot \Theta\big(E_\mu \leq \epsilon\big) \cdot \rho(\mu_y|D)\, d\mu_{\langle \hat{y} \rangle}\, d\mu_y \\ &= \int_{\mu_{\langle \hat{y} \rangle}} \int_{\mu_y = \mu_{\langle \hat{y} \rangle} \pm \epsilon} \rho(\mu_{\langle \hat{y} \rangle}|M, D) \cdot \rho(\mu_y|D)\, d\mu_{\langle \hat{y} \rangle}\, d\mu_y = r. \end{aligned} \quad \text{(A.8)}$$

Thus, the frequentist validation metric is the reliability metric [47] and the "probability of agreement" [57] when reasonable accuracy requirements are imposed on the acceptable difference between the expectation values.

A.3 Area and Binned Probability Difference Metric

The BVM is able to represent and generalize the area metric by letting the comparison values $(\hat{z}, z)$ be the cumulative distribution functions (cdfs) in question to be compared $(F(\hat{y}|M), F(y|D))$. The area metric is,

$$d\big[F(\hat{y}|M), F(y|D)\big] = \int_{-\infty}^{\infty} \big|F(\hat{y}|M) - F(y = \hat{y}|D)\big|\, d\hat{y}. \quad \text{(A.9)}$$

The area metric may be represented by the BVM in a simple way. First allow the comparison value function to be the area metric functional,

$$f(z, \hat{z}) \equiv d\big[F(\hat{y}|M), F(y|D)\big], \tag{A.10}$$

and define agreement with a Boolean,

$$p(A|M, D) = \Theta\Big(B\big(d[F(\hat{y}|M), F(y|D)]\big)\Big). \tag{A.11}$$

Here, the cdfs are treated as completely certain $\big(\delta(\hat{z} - F(\hat{y}|M))$ and $\delta(z - F(y|D))$ are Dirac delta functionals$\big)$. The functional form of the Boolean may be decided given the user's specific validation requirements; however, satisfying some kind of ϵ threshold $d[F(\hat{y}|M), F(y|D)] \leq \epsilon$ seems logical. Generalizations to the area metric may be represented analogously with the BVM.

When the values of the cdfs are uncertain, the BVM becomes,

$$p(A|M, D) \leftarrow \int \Theta\Big(B\big(d[F(\hat{y}|M), F(y|D)]\big)\Big) \cdot \rho\big(F(\hat{y}|M), F(y|D)\big)\, dF(\hat{y}|M)\, dF(y|D), \tag{A.12}$$

which may pose computational challenges and require random sampling; however, in some cases it may be permissible to treat the model cdf as known. The extension to uncertain cdfs is usually not considered in the literature; however, the theoretical generalization to the uncertain case is apparent in the BVM framework. As an approximation, one may consider discretizing the area metric by breaking it into K comparison points,

$$d\big[F(\hat{y}|M), F(y = \hat{y}|D)\big] \approx \sum_{i=1}^{K} \big|F(\hat{y}_i|M) - F(y_i = \hat{y}_i|D)\big|. \tag{A.13}$$

The uncertainty in the cumulative distribution of the data $\rho(z|D)$ is now a finite joint product pdf $\rho\big(F(y_1|D), \ldots, F(y_K|D)\big|D\big)$ over possible cdf values in the K bins, constrained by $F(y_i|D) \leq F(y_{i+1}|D)$. This metric may also be represented with the BVM.

Alternatively, one may consider a binned probability difference comparison functional,

$$d_m\big[p(\hat{y}|M), p(y=\hat{y}|D)\big] = \sum_{i=1}^{K} \big|p(\hat{y}_i|M) - p(y_i = \hat{y}_i|D)\big|. \tag{A.14}$$

Let

$$p(z|D) = p\big(p_{y_1}, \ldots, p_{y_K}\big|D, \textstyle\sum_k p_{y_k} = 1\big)$$

be the uncertainty for the probability estimate in each data bin due to the random process. Given that the pdf of the model is set to a (presumably known) single pdf function $p(\hat{y}|M)$, the BVM is,

$$\begin{aligned} p(A|M,D) &= \int_{\hat{z},z} \delta\big(\hat{z} - \{p(\hat{y}|M)\}\big) \cdot \Theta\Big(B\big(d_m[\hat{z},z]\big)\Big) \cdot p(z|D)\, d\hat{z}\, dz \\ &= \int_{p_{y_1},\ldots,p_{y_K}} \Theta\Big(B\big(d_m\big[\{p(\hat{y}|M)\},\{p(y|D)\}\big]\big)\Big) \cdot p(p_{y_1},\ldots,p_{y_K}|D) \prod_i (dp_{y_i}). \end{aligned}$$

Given the form of the distribution $p(p_{y_1}, \ldots, p_{y_K}|D)$ is well-known (i.e. a Dirichlet distribution after n independent observations of y), the BVM can be treated as an expectation value,

$$p(A|M,D) = E\Big[\Theta\Big(B\big(d_m\big[\{p(\hat{y}|M)\},\{p(y|D)\}\big]\big)\Big)\Big]_{\{p(y|D)\}} \tag{A.15}$$

and estimated with Monte Carlo. When K is large, we may face the curse of dimensionality if the probabilities (bin heights) are themselves uncertain.

The number of bins K plays a role similar to ϵ in that its choice affects the definition/context of agreement represented by the BVM. The binned pdf metric is most informative when K is large because one is checking each region of the pdf for agreement. On the other hand, perhaps when data is limited, there may be instances when having $K = 2$ bins is useful, if for example one was interested in testing a model for representing binary type probabilities (like pass/fail or positive/negative). Using a favorable agreement in this simple binary context to imply that the model works for *otherwise untested* $K > 2$ comparisons/contexts/validations, is statistical misrepresentation and should be avoided (see Section 2.3.3). The BVM requires one to explicitly state the comparison values, the comparison value function, and the definition of agreement to compute $p(A|M, D)$ such that confusion may be avoided and model validation comparisons may be justified.

Thus, the choice of K ultimately determines the granularity of the definition of agreement being tested. If one uses methods similar to [19] to select K, then one is letting the complexity of the data and the number of data points determine the stringency of the definition of agreement.

A.4 Probability Density Function Comparison Metrics

Another way to gauge the agreement between uncertain data and models is through a pdf comparison metric, $G(\rho_D||\rho_M)$ [34]. Examples of $G(\rho_D||\rho_M)$ are the negative relative entropy or KL divergence,

$$D_{KL}(\rho_D||\rho_M) = \int_y \rho(y|D) \log \left(\frac{\rho(y|D)}{\rho(\hat{y} = y|M)} \right) dy, \tag{A.16}$$

the Symmetrized KL divergence $S_{KL}(\rho_D, \rho_M) = D_{KL}(\rho_D||\rho_M) + D_{KL}(\rho_M||\rho_D)$, the Jensen-Shannon divergence $D_{JS}(\rho_D, \rho_M)$, the Hellinger Metric $H(\rho_D, \rho_M)$, the Fisher information distance $\ell(\rho_D, \rho_M)$, and the Wasserstein distance $W(\rho_D, \rho_M)$. These metrics give a notion of "closeness" between the pdfs that can be used for validation.

The BVM may represent these validation metrics by letting the comparison value function $f(\hat{z}, z)$ be the pdf comparison metric,

$$f(\hat{z}, z) \equiv G(\rho_D||\rho_M),$$

where $\hat{z} \equiv \rho_M$ and $z \equiv \rho_D$ are the model and data pdfs, respectively. In the absence of model and data uncertainty (i.e. the pdfs are treated as known functions, e.g. ρ_D

is a gaussian with mean = 0 and variance = 1), the BVM is simply,

$$p(A|M,D) = \Theta\Big(B\big(f(\hat{z},z)\big)\Big) = \Theta\Big(B\big(G(\rho_D||\rho_M)\big)\Big),$$

which either meets the specifications for them to agree (defined by B), or does not (e.g. passing a tolerance threshold). Following the structure of the BVM, if there are uncertainties in the functional forms of the pdfs, they may be included into the BVM,

$$p(A|M,D) \leftarrow \int_{\rho_D,\rho_M} \Theta\Big(B\big(G(\rho_D||\rho_M)\big)\Big) \cdot \rho(\rho_D,\rho_M)\, d\rho_D\, d\rho_M.$$

Uncertainties in the data pdfs may come from a lack of data and uncertainties in the model may come from parametric uncertainty $\big($e.g. a Gaussian pdf model $\rho_{M|\mu}$ with an uncertain mean $\rho(\mu)\big)$.[1] As with the area metric, these metrics may be discretized from pdf to probability comparison metrics, and uncertainties in the pdfs themselves are typically not discussed/quantified in the literature.

A.5 Statistical Hypothesis Testing

Normally when statistical hypothesis testing is performed, one constructs the pdf of the relevant test statistic of the data $\rho(z|D)$ and then assumes the null hypothesis is true ($M = D$) counterfactually, which is enforced by setting the "to be tested population" (the model outputs here) pdf to be equal to the pdf of the test statistic of the data $\rho(\hat{z}|M) \to \rho(\hat{z}|M = D)$. However, in the present case, we are interested in the general modeling case in which one is able to extract a pdf of the outputs of the model, which we would like to test against data before assuming that they are equal. We will first represent the classical statistical hypothesis test using the BVM and then later supply a version that is more relevant to model validation problems.

[1] One should note that here there is the potential for two different models. One in which $\rho_M \equiv \int \rho_{M|\mu}\rho(\mu)\, d\mu$ is marginalized over (in which case the model pdf is "certain" if integrated analytically), and another in which the model pdf is gaussian $\rho_{M|\mu}$ but there is model parametric uncertainty of the form $\rho(\mu)$.

Classical statistical hypothesis testing. In classical hypothesis testing, one constructs the pdf of a relevant test statistic $S_y \equiv z$ of the data $\rho(z|D)$. The null hypothesis is that the model is equal to the data, which is enforced by setting $\rho(\hat{z}|M) = \rho(\hat{z}|M = D)$. Further, the null hypothesis is not rejected if the test statistic from the model $\hat{z}$ falls within the critical region $[-c_\alpha, c_\alpha]$ that corresponds to the probability $\int_{-c_\alpha}^{c_\alpha} \rho(z|D)\, dz = 1 - \alpha$ of the data. The case of not rejecting the null hypothesis is represented by the Boolean expression $B(\hat{z})$, which is true if $-c_\alpha \leq \hat{z} \leq c_\alpha$ is true – defining "agreement" in this case. This results in the following BVM,

$$\begin{aligned} p(A|M = D, D) &= \int_{\hat{z},z} \rho(\hat{z}|M = D) \cdot \Theta\big(B(\hat{z})\big) \cdot \rho(z|D)\, d\hat{z}\, dz \\ &= \int_{\hat{z}} \rho(\hat{z}|M = D) \cdot \Theta\big(-c_\alpha \leq \hat{z} \leq c_\alpha\big)\, d\hat{z} = 1 - \alpha, \quad \text{(A.17)} \end{aligned}$$

because the model was assumed to be equal to the data counterfactually in the null hypothesis and $B(\hat{z}, z) = B(\hat{z})$ only.

The probability of type I error, i.e. rejecting the counterfactually assumed true null hypothesis when it is actually true, is equal to α. It should be noted that this is not equal to the probability of finding the model value outside of the data's confidence interval because it is unknown if the null hypothesis $\big($in what results in $\rho(\hat{z}|M) \to \rho(\hat{z}|M = D)\big)$ is actually true, or not, because the null hypothesis was merely assumed to be true counterfactually. It should further be noted that with probability α the data's test statistic is outside of its *own* confidence interval. Thus, the probability α both indicates type I error *and* a systematic type error that the wrong sort of comparison is being made, i.e. the wrong Boolean expression was chosen, because, why would we care if the model is within a certain confidence interval if the data is not even within that interval?[2] The probability of type II error, i.e. that the null hypothesis was accepted when it is actually false, is equal to $\beta_M(\alpha) = 1 - \int_{-c_\alpha}^{c_\alpha} \rho(\hat{z}|M)\, d\hat{z}$, which in the classical case is difficult to calculate directly because one does not have access to the actual model pdf $p(\hat{z}|M)$ in frequentist probability.

[2] Independent of whether or not the null hypothesis is true.

Improved statistical hypothesis testing for validation. For the validation cases we are interested in, both $\rho(\hat{z}|M)$ and $\rho(z|D)$ are quantified, and therefore assuming that $\rho(\hat{z}|M) \to \rho(\hat{z}|M = D)$ would irresponsibly throw away any information sent through the model. We therefore offer the improved statistical hypothesis test for validation using the BVM, which uses both the model and data pdfs. We call this BVM the statistical power BVM.

For the modified statistical hypothesis test, let the definition of *agreement* be a compound Boolean expression that is true iff both $-c_\alpha \leq \hat{z} \leq c_\alpha$, that the model test statistic lies in the data's confidence interval, *and* $-c_{\hat{\alpha}} \leq z \leq c_{\hat{\alpha}}$ that the data statistic lies in the model's confidence interval, which corresponds to the probability $\int_{-c_{\hat{\alpha}}}^{c_{\hat{\alpha}}} \rho(\hat{z}|M)\, d\hat{z} = 1 - \hat{\alpha}$ of the model. By not assuming $M = D$, we remove the possibility that either type I or type II errors can occur; however, systematic errors, that the Boolean expression meant to define agreement is nonsensical, still exist. Thus, while more or less adhering to the type of tests one might perform for model validation using statistical hypothesis testing, we get,

$$\begin{aligned} p(A|M,D) &= \int_{\hat{z},z} \rho(\hat{z}|M,D) \cdot \Theta\big(-c_\alpha \leq \hat{z} \leq c_\alpha\big) \cdot \Theta\big(-c_{\hat{\alpha}} \leq z \leq c_{\hat{\alpha}}\big) \cdot \rho(z|D)\, d\hat{z}\, dz \\ &= \big(1 - \beta_M(\alpha)\big) \cdot \big(1 - \beta_D(\hat{\alpha})\big), \end{aligned} \tag{A.18}$$

which is the probability that both the model and the data lie within one another's confidence intervals. The value $1 - \beta(\alpha)$ is called the statistical power of the test, but here we have access to both the statistical power of the data and the model. The probability the model and data do not agree as defined by B is given by,

$$p(\overline{A}|M,D) = 1 - p(A|M,D) = \beta_D(\hat{\alpha}) + \beta_M(\alpha) - \beta_D(\hat{\alpha})\beta_M(\alpha), \tag{A.19}$$

which occurs if either $\hat{z}$, z, or both are outside of one another's confidence intervals. The probability for systematic error (that $\hat{z}$, z, or both are outside of their own confidence intervals) is equal to $\alpha + \hat{\alpha} - \alpha\hat{\alpha}$; however, there is no conceptual issue with setting α, $\hat{\alpha}$, or both equal to 0 as long as both distributions do not span the entire range of possible values (as in such a case the model and data would always agree, which means the test has zero resolving power). Setting $\alpha = \hat{\alpha} = 0$ removes the chance of systematic error from the analysis.

Overall, the use of confidence intervals is suboptimal unless both the model and data pdfs are strictly unimodal. Therefore, rather than using a confidence interval, one may use a "confidence set", which we define as the smallest set of $\hat{z}$ (as well as for z) values whose probability adds to $1 - \hat{\alpha}$ (and similarly $1 - \alpha$). Confidence sets are generated by adding the largest probabilities of $\hat{z}$ (z) until a confidence level of $1 - \hat{\alpha}$ ($1 - \alpha$) is met. A confidence interval is only equal to the confidence set if the distribution is unimodal. As the confidence set is the smallest set of values adding up to the confidence level, this set of values is more informative than a confidence interval, which may include many 0 or low probability events. Using a confidence set improves the resolving power of the metric further.

The statistical power BVM (A.18) is more informative than the classical statistical hypothesis test (A.17) because it utilizes the model pdf while also removing type I, type II, and (optionally) systematic errors from the test; however its overall resolving power is weak when compared to other metrics. By rewriting the pair of overlapping confidence intervals (or sets) as the set of values in a single "overlap interval" $I = [-c_\alpha, c_\alpha] \cap [-c_{\hat{\alpha}}, c_{\hat{\alpha}}]$, one may see that (A.18) is a particular case of (A.5) with $S(y) = I$ for $y \in I$ and being the null set otherwise. Thus, the statistical power BVM may be seen as a special case of the generalized reliability metric suggested by the BVM in (A.5) that effectively has large and unvarying tolerance intervals (sets). The use of confidence intervals as well as confidence sets in defining agreement effectively coarse grain the probabilities, which removes their informative features. The most stringent definition of agreement between the model and the data leads to Bayesian model testing, which we prove in the next section, as it is indeed equal to the generalized model reliability metric with a zero tolerance for differences $\epsilon = 0$.

A.6 Bayesian Model Testing

In Bayesian model testing, rather than assuming a particular model is true, one lets the available data determine which model is most likely given that data. Represented probabilistically, out of a set of possible models, Bayesian model testing selects the model with the maximum posterior probability $p(M|Y)$, i.e. the probability of a (previously calibrated [52]) model M given the data set $Y = \{y_i\}$ from data source D. As the selection rule for models requires comparing values of $p(M|Y)$, one often constructs the posterior odds ratio,

$$R = \frac{p(M|Y)}{p(M'|Y)}, \tag{A.20}$$

and selects the model M with the highest value of R relative to a chosen base model M'. Using Bayes' Theorem, the posterior odds ratio may be recast as,

$$R = \frac{p(Y|M)p(M)}{p(Y|M')p(M')} = K\frac{p(M)}{p(M')}, \tag{A.21}$$

where $K \equiv p(Y|M)/p(Y|M')$ is known as the Bayes factor, which is the ratio of the likelihoods of the models. The ratio of the prior model probabilities $p(M)/p(M')$ is the ratio of the belief that model M is true prior to examining the data, relative to M'. If there is reason to believe that one model is more probable than another due to prior (perhaps statistical) knowledge of the system under investigation, then the Bayesian framework allows one to take this into account through the prior model probabilities. As it is common for the R values to be potentially many orders of magnitude greater than one, the prior model probabilities may be overwhelmed by the Bayes factor. In any case, if there is no reason to prefer one model over another, one can let the prior model probabilities be equal a priori.

Given we are testing some model function $\hat{y} = M(\vec{x}, \vec{\alpha})$, we have access to the forward propagation of data and parameters through a given model, the problem of calculating R may be rerouted to the computation of K. Using the potentially multidimensional inputs to our models $(\vec{x}, \vec{\alpha})$, which represent the input data and the model parameters respectively, the Bayes factor is calculated through forward propagation (marginalization) of these inputs through both models,

$$K = \frac{p(Y|M)}{p(Y|M')} = \frac{\int_{\vec{x},\vec{\alpha}} p(Y|\vec{x},\vec{\alpha},M)\cdot\rho(\vec{x},\vec{\alpha}|M)\,d\vec{x}\,d\vec{\alpha}}{\int_{\vec{x}',\vec{\alpha}'} p(Y|\vec{x}',\vec{\alpha}',M')\cdot\rho(\vec{x}',\vec{\alpha}'|M')\,d\vec{x}'\,d\vec{\alpha}'}. \tag{A.22}$$

The prior probability of the inputs to the model $\rho(\vec{x},\vec{\alpha}|M)$ are the input probability distributions used to propagate uncertainty through the model. The probability $p(Y|\vec{x},\vec{\alpha},M)$ is the probability of the data given by the knowledge of the model function $\hat{y} = M(\vec{x},\vec{\alpha})$; however, it should be noted that in general the data $Y = \{y_i\} \neq \hat{Y}$, is not the set of model outputs $\hat{Y}$, as the data Y was collected from the experiment rather than from model outputs. Thus, one must impose the assumptions under which a model is built, i.e. it is built for the purpose of approximating $y_i \approx M(x_i,\vec{\alpha})$, as we will see later. Thus, a more verbose representation of K is,

$$K = \frac{p(\hat{Y}|M)}{p(\hat{Y}|M')} = \frac{\rho(\hat{Y}|M)}{\rho(\hat{Y}|M')} = \frac{\int_{\vec{x},\vec{\alpha}} \rho(\hat{Y}=Y|\vec{x},\vec{\alpha},M)\cdot\rho(\vec{x},\vec{\alpha}|M)\,d\vec{x}\,d\vec{\alpha}}{\int_{\vec{x}',\vec{\alpha}'} \rho(\hat{Y}=Y|\vec{x}',\vec{\alpha}',M')\cdot\rho(\vec{x}',\vec{\alpha}'|M')\,d\vec{x}'\,d\vec{\alpha}'}, \tag{A.23}$$

where $p(\hat{Y}|M)$ is understood to be the sum of the model and the data probabilities that jointly output the same values, i.e. the probability that any of the possible model and data values turn out to be the same. The form of $\rho(\hat{Y} = Y|\vec{x},\vec{\alpha},M)$ is usually assumed to be something that works (and even better if it is also simple) [45], like the product of Gaussian distributions,

$$\rho(\hat{Y}=Y|\vec{x},\vec{\alpha},M) = \frac{1}{Z}\exp\left(-\frac{1}{2\sigma^2}\sum_i \left(M(x_i;\vec{\alpha}) - y_i\right)^2\right), \tag{A.24}$$

where σ^2 is interpreted as the measurement uncertainty of the data.[3] One usually computes the integrals in K using various sampling algorithms such as nested sampling [10, 53] or another Markov Chain Monte Carlo technique. As an added bonus, Bayesian model testing has an inbuilt Occam's razor mechanism which penalizes needlessly complex models, i.e. ones that would overfit the data by using a large number of uncertain model parameters $\vec{\alpha}$. A clear explanation may be found in [5].

[3] The dependence on the data set/experiment D is therefore implied.

We can represent Bayesian model testing using the Bayesian Validation Metric. Let the model comparison value $\hat{z}$ be $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_n)$ that in principle corresponds to $z = Y = (y_1, \ldots, y_n)$, a validation set of data. The Boolean expression B is considered to factor into a set of "and" statements over the individual model output and data points $B(\hat{Y}, Y) = B(\hat{y}_1, y_1) \wedge \ldots \wedge B(\hat{y}_n, y_n)$, where each B is true iff $\hat{y}_i = y_i$ exactly. The Bayesian Validation Metric in this case is then,

$$\begin{aligned} p(A|M, D) &= \int_{\hat{Y}, Y} \rho(\hat{Y}|M, D) \cdot \Theta\big(B(\hat{Y}, Y)\big) \cdot \rho(Y|D) \, d\hat{Y} \, dY \\ &= \int_{\hat{Y}, Y} \rho(\hat{Y}|M, D) \cdot \delta_{\hat{Y}, Y} \cdot \rho(Y|D) \, d\hat{Y} \, dY. \end{aligned} \tag{A.25}$$

In general, computational models may be described by a model function $\hat{Y} = M(\vec{x}, \vec{\alpha})$, and given the model pdf was constructed through forward propagation of the uncertainties $\rho(\vec{x}, \vec{\alpha})$, the model output pdf is

$$\rho(\hat{Y}|M, D) = \int_{\vec{x}, \vec{\alpha}} \rho(\hat{Y}|\vec{x}, \vec{\alpha}, M, D) \cdot \rho(\vec{x}, \vec{\alpha}) \, d\vec{x} \, d\vec{\alpha}. \tag{A.26}$$

Substituting (A.26) into (A.25), using the trick (2.11) and (2.12) – that

$$\int_{Y-\epsilon}^{Y+\epsilon} \rho(\hat{Y}|M, D) \, d\hat{Y} \overset{\epsilon \to 0^+}{\longrightarrow} p(\hat{Y} = Y|M, D) = \rho(\hat{Y} = Y|M, D) \, d\hat{Y},$$

and integrating over Y, one finds,

$$\begin{aligned} p(A|M, D) &= \int_{\vec{x}, \vec{\alpha}, Y} p(\hat{Y} = Y|\vec{x}, \vec{\alpha}, M, D) \cdot \rho(\vec{x}, \vec{\alpha}) \cdot \rho(Y|D) \, d\vec{x} \, d\vec{\alpha} \, dY \\ &= \int_{\vec{x}, \vec{\alpha}} p(\hat{Y}|\vec{x}, \vec{\alpha}, M, D) \cdot \rho(\vec{x}, \vec{\alpha}) \, d\vec{x} \, d\vec{\alpha} \\ &= \left(\int_{\vec{x}, \vec{\alpha}} \rho(\hat{Y}|\vec{x}, \vec{\alpha}, M, D) \cdot \rho(\vec{x}, \vec{\alpha}) \, d\vec{x} \, d\vec{\alpha} \right) d\hat{Y} \\ &= \rho\big(\hat{Y} \equiv Y \big| M, D\big) \, d\hat{Y}, \end{aligned} \tag{A.27}$$

which is what is meant by, and is equal to, the probability in the numerator of the

Bayes factor (A.23),

$$p(A|M,D) = \rho\big(\hat{Y} \equiv Y\big|M,D\big)\,d\hat{Y} \equiv p\big(\hat{Y} \equiv Y\big|M\big). \tag{A.28}$$

That is, we have shown that Bayesian model testing is a special case of the generalized model reliability metric in the case of exact agreement (A.5), i.e. $\epsilon = 0$, as can be seen by investigating (A.25).

Thus, a generalization of Bayesian model testing is to let the definition of agreement have a tolerance $\epsilon > 0$ such that the Bayes factor becomes,

$$K = \frac{p(A|M,D)}{p(A|M',D)} \quad\longrightarrow\quad K(\epsilon) = \frac{p(A|M,D,\epsilon)}{p(A|M',D,\epsilon)}, \tag{A.29}$$

where $p(A|M,D,\epsilon)$ is Equation (A.2). This derivation suggests that the BVM can be used analogously to Bayesian model testing, except with arbitrary definitions of agreement $\epsilon \to B$, that is, we may construct the BVM factor,

$$K(B) = \frac{p(A|M,D,B)}{p(A|M',D,B)}. \tag{A.30}$$

Using Bayes' Theorem, $p(M|A,D,B) = p(A|M,D,B)p(M|D,B)/p(A|D,B)$, we may further construct the BVM ratio,

$$R(B) = \frac{p(M|A,D,B)}{p(M'|A,D,B)} = \frac{p(A|M,D,B)\,p(M|D,B)}{p(A|M',D,B)\,p(M'|D,B)} = K(B)\frac{p(M|D,B)}{p(M'|D,B)}, \tag{A.31}$$

for the purpose of model testing under a general definition of agreement B, i.e. we can do model selection under any definition of agreement with the BVM ratio. The ratio $p(M|D,B)/p(M'|D,B)$ is the ratio of prior probabilities of M and M'. Again, if there is no reason to suspect that one model is a priori more probable than another, one may let $p(M|D,B)/p(M'|D,B) = 1$, and then $R(B) \to K(B)$. With the BVM ratio, one could in principle compare data and models with different data types to perform model testing or selection.

Appendix B

Likelihood-Based Methods

Likelihood-based methods for calibration and model learning use the likelihood function $\mathcal{L}(\vec{\alpha})$ to learn model parameters. The most common likelihood-based method is the max likelihood method, that, due to the monotonicity of the log function is cast as

$$\vec{\alpha}^* = \arg\max_{\vec{\alpha}} \Big(- \log \mathcal{L}(\vec{\alpha}) \Big). \tag{B.1}$$

Although the approach is rather general, in practice one considers likelihood functions generated from stochastic model function

$$\hat{y}_i = M(x_i; \vec{\alpha}) = M(x_i; \alpha_1, \ldots, \alpha_m) + \alpha_0, \tag{B.2}$$

where $M(x_i; \alpha_1, \ldots, \alpha_m)$ is a deterministic model function, α_0 is drawn from $\alpha_0 \sim \mathcal{N}(0, \sigma_{\alpha_0}^2)$, and $\sigma_{\alpha_0} \in \vec{\alpha} \leftarrow (\sigma_{\alpha_0}, \alpha_1, \ldots, \alpha_m)$ may be treated as a learnable parameter. The likelihood of this data given the model (is the true underlying model) is then a product of Gaussians,

$$\mathcal{L}(\vec{\alpha}) = \rho(D|\vec{\alpha}, M) = \prod_{i=1}^{n} \mathcal{N}\big(M(x_i; \alpha_1, \ldots, \alpha_m), \sigma_{\alpha_0}^2\big), \tag{B.3}$$

which implicitly involves setting the certain data, $Y = D$, equal to the model output exactly, $\hat{Y} = Y$, as the true probability described by (B.2) is the model output drawn from the model $\hat{y}_i \sim \mathcal{N}\big(M(x_i; \alpha_1, ..., \alpha_m), \sigma_{\alpha_0}^2\big)$. If the observed data D has measurement or other types of uncertainty, nothing prevents this uncertainty from being built further into $\mathcal{L}(\vec{\alpha})$.

Appendix C

Bayesian Model Testing

We derive Equations (3.4) – (3.6) mentioned in Section 3.2.3. In the Bayesian model testing framework, the model output and the observed data are defined to agree only if their values are exactly equal. Thus, Bayesian model testing is a special case of the BVM where the agreement kernel is equal to the kronecker delta function (exact agreement) with continuous indices, i.e. $\Theta\big(B(\hat{Y},Y)\big) = \delta_{\hat{Y},Y} = \prod_{i=1}^{n} \delta_{\hat{y}_i,y_i}$. Since Bayesian model testing deals with probability densities, we have the following expression for the probability density of agreement (3.7):

$$\begin{aligned}\rho(A|M,D) = \frac{p(A|M,D)}{dA} &= \frac{1}{dA}\int_{\hat{Y},Y} \rho(\hat{Y}|M,D)\cdot\Theta\big(B(\hat{Y},Y)\big)\cdot\rho(Y|D)d\hat{Y}dY \\ &= \int_{\hat{Y},Y} \rho(\hat{Y}|M,D)\cdot\frac{\delta_{\hat{Y},Y}}{dA}\cdot\rho(Y|D)d\hat{Y}dY.\end{aligned}$$

The kronecker delta $\delta_{\hat{Y},Y}$ and the dirac delta $\delta(\hat{Y}-Y)$ functions are related as follows:

$$\frac{\delta_{\hat{Y},Y}}{dA} = \delta(\hat{Y}-Y) = \prod_{i=1}^{n}\delta(\hat{y}_i - y_i).$$

Thus, the probability density of agreement becomes,

$$
\begin{aligned}
\rho(A|M,D) &= \int_{\hat{Y},Y} \rho(\hat{Y}|M,D)\cdot\delta(\hat{Y}-Y)\cdot\rho(Y|D)d\hat{Y}dY \\
&= \int_{\hat{Y},Y}\left(\int_{\vec{\alpha}}\rho(\hat{Y}|\vec{\alpha},M)\rho(\vec{\alpha}|M)d\vec{\alpha}\right)\cdot\delta(\hat{Y}-Y)\cdot\rho(Y|D)d\hat{Y}dY \\
&= \int_{\vec{\alpha}}\left(\int_{\hat{Y},Y}\rho(\hat{Y}|\vec{\alpha},M)\cdot\delta(\hat{Y}-Y)\cdot\rho(Y|D)d\hat{Y}dY\right)\cdot\rho(\vec{\alpha}|M)d\vec{\alpha} \\
&= \int_{\vec{\alpha}}\left(\int_{\hat{Y},Y}\underbrace{\delta\big(\hat{Y}-M(X;\vec{\alpha})\big)}_{\hat{Y}=M(X;\vec{\alpha})}\cdot\delta(\hat{Y}-Y)\cdot\rho(Y|D)d\hat{Y}dY\right)\cdot\rho(\vec{\alpha}|M)d\vec{\alpha} \\
&= \int_{\vec{\alpha}}\left(\int_{Y}\delta\big(M(X;\vec{\alpha})-Y\big)\cdot\rho(Y|D)dY\right)\cdot\rho(\vec{\alpha}|M)d\vec{\alpha}. \qquad \text{(C.1)}
\end{aligned}
$$

C.1 Normally Distributed Data

If we assume the data to be normally distributed, i.e. $Y \sim \mathcal{N}(D,\Delta)$, we get,

$$\rho(Y|D) = \frac{1}{\sqrt{(2\pi)^n|\Delta|}}e^{-\frac{1}{2}(Y-D)^T\Delta^{-1}(Y-D)},$$

where n is the dimension of the training data set, Δ is the covariance matrix, and D is the observed data values.

Therefore, using (C.1), we have,

$$
\begin{aligned}
\rho(A|M,D) &= \int_{\vec{\alpha}}\left(\int_{Y}\delta\big(M(X;\vec{\alpha})-Y\big)\cdot\frac{1}{\sqrt{(2\pi)^n|\Delta|}}e^{-\frac{1}{2}(Y-D)^T\Delta^{-1}(Y-D)}dY\right)\cdot\rho(\vec{\alpha}|M)d\vec{\alpha}, \\
&= \int_{\vec{\alpha}}\underbrace{\frac{1}{\sqrt{(2\pi)^n|\Delta|}}e^{-\frac{1}{2}\big(M(X;\vec{\alpha})-D\big)^T\Delta^{-1}\big(M(X;\vec{\alpha})-D\big)}}_{\mathcal{L}(\vec{\alpha})}\cdot\underbrace{\rho(\vec{\alpha}|M)d\vec{\alpha}}_{\pi(\vec{\alpha})d\vec{\alpha}}.
\end{aligned}
$$

Therefore, the likelihood function to be used in the MCMC algorithm is

$$\mathcal{L}(\vec{\alpha}) = \frac{1}{\sqrt{(2\pi)^n|\Delta|}}e^{-\frac{1}{2}\big(M(X;\vec{\alpha})-D\big)^T\Delta^{-1}\big(M(X;\vec{\alpha})-D\big)},$$

which is Equation (3.4) presented in Section 3.2.3.

C.2 Uniformly Distributed Data

We first note that

$$dY \equiv d^nY \equiv \prod_{j=1}^{n} dy_j, \qquad j = 1, \ldots, n. \tag{C.2}$$

Now, we assume the data to be uniformly distributed, i.e.

$$y_j \sim \mathcal{U}(a_j, b_j), \qquad j = 1, \ldots, n.$$

Then, the probability density $\rho(Y|D)$ becomes:

$$\rho(Y|D) = \prod_{j=1}^{n} \rho(y_j|D) = \prod_{j=1}^{n} \frac{\Theta\big(a_j \le y_j \le b_j\big)}{b_j - a_j}. \tag{C.3}$$

Notice that we can generalize $\rho(y_j|D) = \prod_{j=1}^{n} \Theta\big(a_j \le y_j \le b_j\big)\rho(y_j|D)$ to any bounded probability density function (pdf). Therefore, using (C.1), we have,

$$\begin{aligned}\rho(A|M, D) &= \int_{\vec{\alpha}} \left(\prod_{j=1}^{n} \int_{y_j} \delta\Big(M(x_j; \vec{\alpha}) - y_j\Big) \cdot \frac{\Theta\big(a_j \le y_j \le b_j\big)}{b_j - a_j}\, dy_j \right) \cdot \rho(\vec{\alpha}|M) d\vec{\alpha}, \\ &= \int_{\vec{\alpha}} \underbrace{\left(\prod_{j=1}^{n} \frac{\Theta\big(a_j \le M(x_j; \vec{\alpha}) \le b_j\big)}{b_j - a_j} \right)}_{\mathcal{L}(\vec{\alpha})} \cdot \underbrace{\rho(\vec{\alpha}|M) d\vec{\alpha}}_{\pi(\vec{\alpha}) d\vec{\alpha}}.\end{aligned}$$

Therefore, the likelihood function to be used in the MCMC algorithm is

$$\mathcal{L}(\vec{\alpha}) = \prod_{j=1}^{n} \frac{\Theta\big(a_j \le M(x_j; \vec{\alpha}) \le b_j\big)}{b_j - a_j},$$

which is Equation (3.5) presented in Section 3.2.3.

C.3 Completely Certain Data

If we consider the data to be completely certain, deterministic, i.e. $Y = D$, then, the probability density $\rho(Y|D)$ becomes $\rho(Y|D) = \delta(Y - D)$, and thus, using (C.1), we have,

$$\begin{aligned}\rho(A|M, D) &= \int_{\vec{\alpha}} \left(\int_Y \delta\big(M(X; \vec{\alpha}) - Y\big) \cdot \delta(Y - D) dY \right) \cdot \rho(\vec{\alpha}|M) d\vec{\alpha} \\ &= \int_{\vec{\alpha}} \underbrace{\delta\big(M(X; \vec{\alpha}) - D\big)}_{\mathcal{L}(\vec{\alpha})} \cdot \underbrace{\rho(\vec{\alpha}|M) d\vec{\alpha}}_{\pi(\vec{\alpha}) d\vec{\alpha}}.\end{aligned}$$

Therefore, the likelihood function to be used in the MCMC algorithm is

$$\mathcal{L}(\vec{\alpha}) = \delta\big(M(X; \vec{\alpha}) - D\big),$$

which is Equation (3.6) presented in Section 3.2.3.

Appendix D

BVM Model Selection

We derive Equations (3.15) – (3.17) presented in Section 3.3. We show how we can apply Bayesian model selection on any data distribution using the BVM probability of agreement. Starting from the original definition of probability of agreement (3.7), we have,

$$\begin{aligned}
p(A|M, D, B) &= \int_{\hat{Y},Y} \rho(\hat{Y}|M, D) \cdot \Theta\big(B(\hat{Y}, Y)\big) \cdot \rho(Y|D) d\hat{Y} dY \\
&= \int_{\hat{Y},Y} \left(\int_{\vec{\alpha}} \rho(\hat{Y}|\vec{\alpha}, M) \rho(\vec{\alpha}|M) d\vec{\alpha} \right) \cdot \Theta\big(B(\hat{Y}, Y)\big) \cdot \rho(Y|D) d\hat{Y} dY \\
&= \int_{\vec{\alpha}} \left(\int_{\hat{Y},Y} \rho(\hat{Y}|\vec{\alpha}, M) \cdot \Theta\big(B(\hat{Y}, Y)\big) \cdot \rho(Y|D) d\hat{Y} dY \right) \cdot \rho(\vec{\alpha}|M) d\vec{\alpha} \\
&= \int_{\vec{\alpha}} \left(\int_{\hat{Y},Y} \delta\big(\hat{Y} - M(X; \vec{\alpha})\big) \cdot \Theta\big(B(\hat{Y}, Y)\big) \cdot \rho(Y|D) d\hat{Y} dY \right) \cdot \rho(\vec{\alpha}|M) d\vec{\alpha} \\
&= \int_{\vec{\alpha}} \left(\int_{Y} \Theta\Big(B\big(M(X; \vec{\alpha}), Y\big)\Big) \cdot \rho(Y|D) dY \right) \cdot \rho(\vec{\alpha}|M) d\vec{\alpha},
\end{aligned}$$

which is Equation (3.15) derived in Section 3.3. From (C.2), we know that

$$dY \equiv d^n Y \equiv \prod_{j=1}^{n} dy_j, \qquad j = 1, \dots, n.$$

The probability density $\rho(Y|D)$ can be expressed as:

$$\rho(Y|D) = \prod_{j=1}^{n} \rho(y_j|D).$$

We also note that the compound Boolean under question can be expressed as:

$$\Theta\Big(B\big(M(X;\vec{\alpha}),Y\big)\Big)=\prod_{j=1}^{n}\Theta\Big(B\big(M(x_j;\vec{\alpha}),y_j\big)\Big)\qquad j=1,\ldots,n.$$

Thus, we rewrite the BVM probability of agreement as follows:

$$p(A|M,D,B)=\int_{\vec{\alpha}}\left(\prod_{j=1}^{n}\int_{y_j}\Theta\Big(B\big(M(x_j;\vec{\alpha}),y_j\big)\Big)\cdot\rho(y_j|D)\,dy_j\right)\cdot\rho(\vec{\alpha}|M)d\vec{\alpha}.$$

We will use the $\epsilon-$Boolean indicator function defined as:

$$\Theta\Big(B\big(M(x_j;\vec{\alpha}),y_j\big)\Big)=\begin{cases}1, & \text{if } \big|y_j-M(x_j;\vec{\alpha})\big|\le\epsilon\\ 0, & \text{otherwise}\end{cases}$$

where $j=1,\ldots,n$.

Then, the indicator function can be rewritten as:

$$\begin{aligned}\Theta\Big(B\big(M(x_j;\vec{\alpha}),y_j\big)\Big)&=\Theta\Big(\big|y_j-M(x_j;\vec{\alpha})\big|\le\epsilon\Big)\\&=\Theta\Big(M(x_j;\vec{\alpha})-\epsilon\le y_j\le M(x_j;\vec{\alpha})+\epsilon\Big),\end{aligned}$$

where $j=1,\ldots,n$. Therefore, the BVM probability of agreement can be expressed as:

$$p(A|M,D,B)=\int_{\vec{\alpha}}\left(\prod_{j=1}^{n}\int_{y_j}\Theta\Big(M(x_j;\vec{\alpha})-\epsilon\le y_j\le M(x_j;\vec{\alpha})+\epsilon\Big)\cdot\rho(y_j|D)\,dy_j\right)\cdot\rho(\vec{\alpha}|M)d\vec{\alpha}.\tag{D.1}$$

D.1 Normally Distributed Data

Note that the Boolean $\Theta\Big(M(x_j;\vec{\alpha})-\epsilon\le y_j\le M(x_j;\vec{\alpha})+\epsilon\Big)$ is equal to 1 only when y_j belongs to the interval $\Big[M(x_j;\vec{\alpha})-\epsilon,\ M(x_j;\vec{\alpha})+\epsilon\Big]$.

Thus, the BVM probability of agreement becomes:

$$p(A|M,D,B)=\int_{\vec{\alpha}}\left(\prod_{j=1}^{n}\int_{M(x_j;\vec{\alpha})-\epsilon}^{M(x_j;\vec{\alpha})+\epsilon}\rho(y_j|D)\,dy_j\right)\cdot\rho(\vec{\alpha}|M)d\vec{\alpha}.$$

Now, we assume that the data is normally distributed, i.e.

$$y_j \sim \mathcal{N}(D_j, \sigma_j^2), \qquad j = 1, \ldots, n.$$

Then, the probability density $\rho(Y|D)$ becomes:

$$\rho(Y|D) = \prod_{j=1}^{n} \rho(y_j|D) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \mathrm{e}^{-\frac{1}{2}\left(\frac{y_j - D_j}{\sigma_j}\right)^2}.$$

Thus, we rewrite the BVM probability of agreement as follows:

$$\begin{aligned}
p(A|M, D, B) &= \int_{\vec{\alpha}} \left(\prod_{j=1}^{n} \int_{M(x_j;\vec{\alpha})-\epsilon}^{M(x_j;\vec{\alpha})+\epsilon} \rho(y_j|D)\, dy_j \right) \cdot \rho(\vec{\alpha}|M) d\vec{\alpha} \\
&= \int_{\vec{\alpha}} \left(\prod_{j=1}^{n} \int_{M(x_j;\vec{\alpha})-\epsilon}^{M(x_j;\vec{\alpha})+\epsilon} \frac{1}{\sqrt{2\pi}\sigma_j} \mathrm{e}^{-\frac{1}{2}\left(\frac{y_j - D_j}{\sigma_j}\right)^2} dy_j \right) \cdot \rho(\vec{\alpha}|M) d\vec{\alpha} \\
&= \int_{\vec{\alpha}} \underbrace{\prod_{j=1}^{n} \left(F\Big(M(x_j;\vec{\alpha}) + \epsilon\Big) - F\Big(M(x_j;\vec{\alpha}) - \epsilon\Big) \right)}_{\mathcal{L}(\vec{\alpha}, B)} \cdot \underbrace{\rho(\vec{\alpha}|M) d\vec{\alpha}}_{\pi(\vec{\alpha})d\vec{\alpha}},
\end{aligned}$$

where $F(x)$ is the cumulative distribution function (cdf) expressed as:

$$F(x) = \Phi\left(\frac{x - D}{\sigma}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, i.e. $\mathcal{N}(0, 1)$, and expressed as:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}}\, dt.$$

Therefore, the likelihood function to be used in the MCMC algorithm is

$$\mathcal{L}(\vec{\alpha}, B) = \prod_{j=1}^{n} \left(F\Big(M(x_j;\vec{\alpha}) + \epsilon\Big) - F\Big(M(x_j;\vec{\alpha}) - \epsilon\Big) \right).$$

D.2 Uniformly Distributed Data

If we assume the data to be uniformly distributed (C.3), then the BVM probability of agreement (D.1) becomes:

$$p(A|M,D,B) = \int_{\vec{\alpha}} \left(\prod_{j=1}^{n} \int_{y_j} \Theta\Big(M(x_j;\vec{\alpha}) - \epsilon \leq y_j \leq M(x_j;\vec{\alpha}) + \epsilon\Big) \cdot \frac{\Theta(a_j \leq y_j \leq b_j)}{b_j - a_j}\, dy_j \right) \cdot \rho(\vec{\alpha}|M) d\vec{\alpha}.$$

Note that the product $\Theta\Big(M(x_j;\vec{\alpha}) - \epsilon \leq y_j \leq M(x_j;\vec{\alpha}) + \epsilon\Big) \cdot \Theta\Big(a_j \leq y_j \leq b_j\Big)$ is equal to 1 only when y_j belongs to both intervals $\Big[M(x_j;\vec{\alpha}) - \epsilon,\ M(x_j;\vec{\alpha}) + \epsilon\Big]$ and $\Big[a_j,\ b_j\Big]$.

Let l_j and u_j be such that

$$\Big[l_j,\ u_j\Big] = \Big[M(x_j;\vec{\alpha}) - \epsilon,\ M(x_j;\vec{\alpha}) + \epsilon\Big] \cap \Big[a_j,\ b_j\Big] \qquad j = 1, \ldots, n$$

Thus, the BVM probability of agreement becomes:

$$\begin{aligned} p(A|M,D,B) &= \int_{\vec{\alpha}} \left(\prod_{j=1}^{n} \int_{y_j} \frac{\Theta(l_j \leq y_j \leq u_j)}{b_j - a_j}\, dy_j \right) \cdot \rho(\vec{\alpha}|M) d\vec{\alpha} \\ &= \int_{\vec{\alpha}} \left(\prod_{j=1}^{n} \int_{l_j}^{u_j} \frac{1}{b_j - a_j}\, dy_j \right) \cdot \rho(\vec{\alpha}|M) d\vec{\alpha} \\ &= \int_{\vec{\alpha}} \underbrace{\left(\prod_{j=1}^{n} \frac{u_j - l_j}{b_j - a_j} \right)}_{\mathcal{L}(\vec{\alpha},B)} \cdot \underbrace{\rho(\vec{\alpha}|M) d\vec{\alpha}}_{\pi(\vec{\alpha})d\vec{\alpha}}. \end{aligned}$$

Therefore, the likelihood function to be used in the MCMC algorithm is

$$\mathcal{L}(\vec{\alpha},B) = \prod_{j=1}^{n} \frac{u_j - l_j}{b_j - a_j},$$

which is Equation (3.16) presented in Section 3.3.

D.3 Completely Certain Data

If we consider the data to be completely certain, deterministic, i.e. $Y = D$, then, the probability density $\rho(Y|D)$ becomes $\rho(Y|D) = \delta(Y - D)$, and thus, using (3.15), we have,

$$\begin{aligned} p(A|M, D, B) &= \int_{\vec{\alpha}} \left(\int_{Y} \Theta\Big(B\big(M(X;\vec{\alpha}), Y\big)\Big) \cdot \delta(Y - D) dY \right) \cdot \rho(\vec{\alpha}|M) d\vec{\alpha} \\ &= \int_{\vec{\alpha}} \underbrace{\Theta\Big(B\big(M(X;\vec{\alpha}), D\big)\Big)}_{\mathcal{L}(\vec{\alpha}, B)} \cdot \underbrace{\rho(\vec{\alpha}|M) d\vec{\alpha}}_{\pi(\vec{\alpha}) d\vec{\alpha}} . \end{aligned}$$

Therefore, the likelihood function to be used in the MCMC algorithm is

$$\mathcal{L}(\vec{\alpha}, B) = \Theta\Big(B\big(M(X;\vec{\alpha}), D\big)\Big),$$

which is Equation (3.17) presented in Section 3.3.

Bibliography

[1] Brian M Adams, WJ Bohnhoff, KR Dalbey, JP Eddy, MS Eldred, DM Gay, K Haskell, Patricia D Hough, and Laura P Swiler. Dakota, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: version 5.0 user's manual. *Sandia National Laboratories, Tech. Rep. SAND2010-2183*, 2009.

[2] Mark A. Beaumont, Wenyang Zhang, and David J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[3] Paul Mac Berthouex and Linfield C. Brown. *Statistics for Environmental Engineers.* Lewis Publishers, 2002.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag, Berlin, Heidelberg, 2006.

[5] Ariel Caticha. Entropic inference and the foundations of physics. *Brazilian Chapter of the International Society for Bayesian Analysis-ISBrA, Sao Paulo, Brazil*, 2012.

[6] Radu V. Craiu and Jeffrey S. Rosenthal. Bayesian computation via markov chain monte carlo. *Annual Review of Statistics and Its Application*, 1(1):179–201, 2014.

[7] Bert J Debusschere, Habib N Najm, Philippe P Pébay, Omar M Knio, Roger G Ghanem, and Olivier P Le Maı tre. Numerical challenges in the use of polynomial chaos representations for stochastic processes. *SIAM journal on scientific computing*, 26(2):698–719, 2004.

[8] Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[9] Anthony William Fairbank Edwards. *Likelihood.* CUP Archive, 1984.

[10] F. Feroz and M. P. Hobson. Multimodal nested sampling: an efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, 384(2):449–463, Jan 2008.

[11] F. Feroz, M. P. Hobson, and M. Bridges. Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, 398(4):1601–1614, Oct 2009.

[12] Scott Ferson, William L. Oberkampf, and Lev Ginzburg. Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering*, 197(29):2408 – 2430, 2008. Validation Challenge Workshop.

[13] A. E. Gelfand and D. K. Dey. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.

[14] John Geweke. Bayesian model comparison and validation. *American Economic Review*, 97(2):60–64, May 2007.

[15] R Ghanem, D Higdon, and H Owhadi. The uncertainty quantification toolkit (uqtk), handbook of uncertainty quantification, 2016.

[16] Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice.* Chapman and Hall/CRC, 1995.

[17] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[18] Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[19] Kevin H Knuth. Optimal data-based binning for histograms. *arXiv preprint physics/0605197*, 2006.

[20] Kevin H. Knuth, Michael Habeck, Nabin K. Malakar, Asim M. Mubeen, and Ben Placek. Bayesian evidence and model selection. *Digital Signal Processing*, 47:50–67, Dec 2015.

[21] Olivier Le Maître and Omar M Knio. *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics.* Springer Science & Business Media, 2010.

[22] Guesuk Lee, Wongon Kim, Hyunseok Oh, Byeng D Youn, and Nam H Kim. Review of statistical model calibration and validation—from the perspective of uncertainty structures. *Structural and Multidisciplinary Optimization*, pages 1–26, 2019.

[23] Peter M Lee. *Bayesian statistics.* Arnold Publication, 1997.

[24] Thomas Leonard and John SJ Hsu. *Bayesian methods: an analysis for statisticians and interdisciplinary researchers*, volume 5. Cambridge University Press, 2001.

[25] Chenzhao Li and Sankaran Mahadevan. Role of calibration, validation, and relevance in multi-level uncertainty integration. *Reliability Engineering & System Safety*, 148:32 – 43, 2016.

[26] Wei Li, Wei Chen, Zhen Jiang, Zhenzhou Lu, and Yu Liu. New validation metrics for models with multiple correlated responses. *Reliability Engineering & System Safety*, 127:1 – 11, 2014.

[27] You Ling and Sankaran Mahadevan. Quantitative model validation techniques: New insights. *Reliability Engineering & System Safety*, 111:217 – 231, 2013.

[28] Yu Liu, Wei Chen, Paul Arendt, and Hong-Zhong Huang. Toward a better understanding of model validation metrics. *Journal of Mechanical Design*, 133(7):071005, 2011.

[29] Scott M. Lynch. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer-Verlag, New York, 2007.

[30] David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002.

[31] Sankaran Mahadevan and Ramesh Rebba. Validation of reliability computational models using bayes networks. *Reliability Engineering & System Safety*, 87(2):223 – 232, 2005.

[32] Alberto Malinverno and Victoria A Briggs. Expanded uncertainty quantification in inverse problems: Hierarchical bayes and empirical bayes. *Geophysics*, 69(4):1005–1016, 2004.

[33] Jean-Michel Marin, Pierre Pudlo, Christian P. Robert, and Robin J. Ryder. Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, Nov 2012.

[34] Kathryn A Maupin, Laura P Swiler, and Nathan W Porter. Validation metrics for deterministic and probabilistic data. *Journal of Verification, Validation and Uncertainty Quantification*, 3(3):031002, 2018.

[35] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[36] Joshua Mullins, You Ling, Sankaran Mahadevan, Lin Sun, and Alejandro Strachan. Separation of aleatory and epistemic uncertainty in probabilistic model validation. *Reliability Engineering & System Safety*, 147:49 – 59, 2016.

[37] Radford M. Neal. An improved acceptance procedure for the hybrid monte carlo algorithm. *Journal of Computational Physics*, 111(1):194–203, March 1994.

[38] William L. Oberkampf and Matthew F. Barone. Measures of agreement between computation and experiment: Validation metrics. *Journal of Computational Physics*, 217(1):5 – 36, 2006. Uncertainty Quantification in Simulation Science.

[39] William L Oberkampf, Timothy G Trucano, and Charles Hirsch. Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews*, 57(5):345–384, 12 2004.

[40] Inseok Park, Hemanth K. Amarchinta, and Ramana V. Grandhi. A Bayesian approach for quantification of model uncertainty. *Reliability Engineering & System Safety*, 95(7):777–785, 2010.

[41] M Parno and A Davis. MUQ: MIT Uncertainty Quantification Library, 2018.

[42] Matthew D. Parno and Youssef M. Marzouk. Transport map accelerated markov chain monte carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):645–682, Jan 2018.

[43] Yudi Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.

[44] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[45] Ben Placek. *Bayesian detection and characterization of extra-solar planets via photometric variations*. PhD thesis, 2014.

[46] Ben Placek, Kevin H. Knuth, and Daniel Angerhausen. EXONEST: BAYESIAN MODEL SELECTION APPLIED TO THE DETECTION AND CHARACTERIZATION OF EXOPLANETS VIA PHOTOMETRIC VARIATIONS. *The Astrophysical Journal*, 795(2):112, oct 2014.

[47] Ramesh Rebba and Sankaran Mahadevan. Computational methods for model reliability assessment. *Reliability Engineering & System Safety*, 93(8):1197 – 1207, 2008.

[48] Christian P. Robert, Víctor Elvira, Nick Tawn, and Changye Wu. Accelerating mcmc algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(5):e1435, 2018.

[49] Christopher J. Roy and William L. Oberkampf. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25):2131 – 2144, 2011.

[50] Shankar Sankararaman and Sankaran Mahadevan. Model validation under epistemic uncertainty. *Reliability Engineering & System Safety*, 96(9):1232 – 1241, 2011. Quantification of Margins and Uncertainties.

[51] Shankar Sankararaman and Sankaran Mahadevan. Assessing the reliability of computational models under uncertainty. In *54th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, page 1873, 2013.

[52] Shankar Sankararaman and Sankaran Mahadevan. Integration of model verification, validation, and calibration for uncertainty quantification in engineering systems. *Reliability Engineering & System Safety*, 138:194–209, 2015.

[53] D. S. Sivia and J. Skilling. *Data Analysis - A Bayesian Tutorial.* Oxford Science Publications. Oxford University Press, 2nd edition, 2006.

[54] John Skilling. Nested sampling for general bayesian computation. *Bayesian Anal.*, 1(4):833–859, 12 2006.

[55] D. Sornette, A. B. Davis, K. Ide, K. R. Vixie, V. Pisarenko, and J. R. Kamm. Algorithm for model validation: Theory and applications. *Proceedings of the National Academy of Sciences*, 104(16):6562–6567, 2007.

[56] M Stefano and B Sudret. Uqlab user manual-polynomial chaos expansions. report uqlab-v0. 9-104, chair of risk, safety and uncertainty quantification. *ETH Zurich*, 2015.

[57] Nathaniel Stevens. Assessment and comparison of continuous measurement systems. 2014.

[58] Tony Tohme, Kevin Vanslette, and Kamal Youcef-Toumi. Generalized bayesian regression and model learning. *arXiv preprint arXiv:1911.11715*, 2019.

[59] Paul Troughton and Simon J. Godsill. Bayesian model selection for time series using markov chain monte carlo. In *ICASSP*, pages 3733–3736, 1997.

[60] T.G. Trucano, L.P. Swiler, T. Igusa, W.L. Oberkampf, and M. Pilch. Calibration, validation, and sensitivity analysis: What's what. *Reliability Engineering & System Safety*, 91(10):1331 – 1357, 2006. The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004).

[61] Kevin Vanslette, Abdullatif Al Alsheikh, and Kamal Youcef-Toumi. Why simple quadrature is just as good as monte carlo. *Monte Carlo Methods and Applications*, 26(1), 2020.

[62] Kevin Vanslette, Tony Tohme, and Kamal Youcef-Toumi. A general model validation and testing tool. *Reliability Engineering & System Safety*, 195, March 2020.

[63] Chong Wang and Hermann G. Matthies. Novel model calibration method via non-probabilistic interval characterization and bayesian theory. *Reliability Engineering & System Safety*, 183:84 – 92, 2019.

[64] CJ Wild and GAF Seber. *Nonlinear regression.* New York: Wiley, 1989.

[65] Danqing Wu, Zhenzhou Lu, Yanping Wang, and Lei Cheng. Model validation and calibration based on component functions of model output. *Reliability Engineering & System Safety*, 140:59 – 70, 2015.

[66] Ruoxue Zhang and Sankaran Mahadevan. Bayesian methodology for reliability model acceptance. *Reliability Engineering & System Safety*, 80(1):95 – 103, 2003.

[67] Lufeng Zhao, Zhenzhou Lu, Wanying Yun, and Wenjin Wang. Validation metric based on mahalanobis distance for models with multiple correlated responses. *Reliability Engineering & System Safety*, 159:80 – 89, 2017.