Acta Numerica (2018), pp. 353–450
doi:10.1017/S096249291800003X
© Cambridge University Press, 2018
Printed in the United Kingdom

Adaptive multiscale predictive modelling

J. Tinsley Oden
Institute for Computational Engineering and Sciences,
The University of Texas at Austin,
TX 78712, USA
E-mail: oden@ices.utexas.edu

The use of computational models and simulations to predict events that take place in our physical universe, or to predict the behaviour of engineered systems, has significantly advanced the pace of scientific discovery and the creation of new technologies for the benefit of humankind over recent decades, at least up to a point. That 'point' in recent history occurred around the time that the scientific community began to realize that true predictive science must deal with many formidable obstacles, including the determination of the reliability of the models in the presence of many uncertainties. To develop meaningful predictions one needs relevant data, itself possessing uncertainty due to experimental noise; in addition, one must determine model parameters, and concomitantly, there is the overriding need to select and validate models given the data and the goals of the simulation.

This article provides a broad overview of predictive computational science within the framework of what is often called the science of uncertainty quantification. The exposition is divided into three major parts. In Part 1, philosophical and statistical foundations of predictive science are developed within a Bayesian framework. There the case is made that the Bayesian framework provides, perhaps, a unique setting for handling all of the uncertainties encountered in scientific prediction. In Part 2, general frameworks and procedures for the calculation and validation of mathematical models of physical realities are given, all in a Bayesian setting. But beyond Bayes, an introduction to information theory, the maximum entropy principle, model sensitivity analysis and sampling methods such as MCMC are presented. In Part 3, the central problem of predictive computational science is addressed: the selection, adaptive control and validation of mathematical and computational models of complex systems. The Occam Plausibility Algorithm, OPAL, is introduced as a framework for model selection, calibration and validation. Applications to complex models of tumour growth are discussed.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

CONTENTS

PART 1: Introduction to predictive science

1 Introductory thoughts and comments 355
2 Predictive computational science 356
3 The logical theory of probability of R. T. Cox and E. T. Jaynes 363
4 Bayesian framework for statistical calibration and validation of parametric models 374

PART 2: Principles of Bayesian model calibration and validation

5 The prediction pyramid: Bayesian statistical calibration and validation parametric models 381
6 Information theory and the principle of maximum entropy 391
7 Monte Carlo methods: MCMC and sensitivity analysis 400

PART 3: Adaptive modelling: model selection, OPAL and adaptive control of multiscale models

8 $\mathcal{M}$ uncertainty 415
9 Adaptive control of multiscale models 423
10 Selection, calibration, and validation of families of stochastic parametric models: prediction of tumour growth 435

References 445

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

PART ONE
Introduction to predictive science

1. Introductory thoughts and comments

1.1. Comments on the title

The title of this article needs some elaboration. In particular, the adjective 'predictive' sets the entire exposition in a non-conventional direction, as it immediately brings into discussion the idea that some models, or some multiscale modelling, may not be truly predictive of realities at all. The terms *predictive modelling or predictive science* have come to mean the disciplines, philosophy, processes and mathematics underlying what is necessary to forecast realities, actual events and phenomena that occur in the physical universe. A central issue is that for scientific prediction, one must cope with a daunting list of uncertainties in every aspect of the prediction process or, equivalently, of the process of developing predictive models of physical events. We discuss the foundations of predictive computational science and methods for quantifying uncertainty in this article. Some of these ideas are surveyed in Oden, Babuška and Faghihi (2018) and Oden (2017), and portions of Part 1 of this exposition are drawn from these works.

Another adjective in the title is *multiscale*, a term appearing frequently in contemporary literature, in reference to the fact that many physical phenomena (in fact, all physical phenomena we observe in our day-to-day existence), are manifested by events that occur at many different spatial and temporal scales. How to correctly model the complex events at one spatial scale and how they may influence events at other scales stands as one of the most elusive and challenging goals of modern computational science. Indeed, this question has stood at the centre of physical chemistry, thermodynamics and statistical mechanics for over a century.

Finally, the word *adaptive* also appears in our title, meaning to 'change or adapt so as to control and improve'. A predictive model must be able to predict events or realities with a level of accuracy acceptable to the modeller, and this demands that the modeller provide a metric or norm measuring the distance of a (computer) prediction of a quantity of interest from a reality or from an exact solution of the model, plus an error tolerance defining the level of error acceptable to the modeller. If that level of accuracy is not met, one seeks to adapt the model to improve accuracy and meet the error tolerance or reject it in the favour of another model. Adaptive modelling is thus an exercise in control theory, in which a system is controlled to meet well-defined objectives. In contemporary predictive adaptive modelling, the control paradigms are severely complicated by the overpowering uncertainties in observational data, model parameters and, importantly, in the actual

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

selection of the appropriate model itself. When one adds the complexities that occur when different model structures exist as different spatial and temporal scales, a new field emerges – and a brief journey into that field is one goal of this exposition.

1.2. The structure of this exposition

This work deals with several fundamental topics within the broad area of predictive modelling, all built around relatively new computational methods. Thus, while numerical algorithms are at the centre of all parts of this essay, the scope is a much larger one, not only covering the structure of key algorithms but also the broader philosophy of predictability in the presence of uncertainties. Because of the wide scope of this presentation, a tutorial style is employed in many places, so as to make the work accessible to a broad audience.

2. Predictive computational science

Summing up the points laid down in the opening comments, we define predictive computational science as the scientific discipline concerned with assessing the predictability of mathematical and computational models of events that occur in our physical universe in the presence of uncertainties in all of the factors that determine the reliability and scientific significance of the prediction. Prediction is a fundamental mission of basic science and the essence of engineering, which is concerned with the application of science in the prediction at the performance of engineered systems under design conditions: the overarching goal of engineering for centuries. It is also at the core of medical science, which seeks to predict the outcome of therapies and medical procedures.

What is a 'science-based prediction'? Indeed, what is science? The word *science* is derived from the Latin, *scientia*, meaning knowledge. Thus, the *Oxford English Dictionary* defines science as the activity concerned with the systematic acquisition of knowledge: the enterprise that builds and organizes knowledge in forming testable explanations and predictions about events that take place, or can take place, in the universe. Engineering and medicine are the activities concerned with the application of scientific and practical knowledge for the benefit or needs of mankind.

Exactly how knowledge is acquired is a subject debated for millennia. To Socrates and Plato (*c.* 370 BCE), humankind is endowed with pre-existing knowledge, knowledge that is already present in our minds and is accessible to our intuitive powers. Aristotle had a quite different view: in his *Organon* describing his six works on logic, he asserts that the acquisition of knowledge comes through sensation and perception. *Sensation* is the use of one's

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

five senses to experience reality via empirical events and thus interpretation by the consciousness of the subject acquiring knowledge. To this he adds *perception*, the recognition that the event has a sensation. The term *observation* refers to the acquisition of information through one's senses, or through aids to one's senses made possible by using instruments, all to either observe natural events or events orchestrated in experimental set-ups. Aristotle was the father of logic – the science of reasoning – and he laid down the rules of deductive logic as the foundations of algebra and much of mathematics.

Deductive reasoning, or deductive logic, began with the two-sided logic advanced by Aristotle. It is embodied in a process of reasoning beginning with one or more general statements (axioms or principles) and following the rules of logic to reach logical conclusions. It thus provides the formalisms for mathematics and mathematical proof. It (deduction) is sometimes called 'top-down' logic. It is infallible and not influenced by empirical observations. On the other hand, scientific knowledge is based on *inductive logic*, the process of reasoning by generalizing or extrapolating from initial information or hypotheses. Inductive logic is 'bottom-up' logic, an open system including domains of epistemic uncertainty, thus allowing a conclusion to be false.

Thus, from antiquity there came the notions that knowledge can arise from observations of natural events or events manufactured in experiments, and through logical conceptualizations, usually mathematical in form, that model or characterize in some way the observations, with the hope of predicting events in the future. Thus, a scientific theory generally rests on a mathematical model. But not all philosophies accepted these ideas.

Philosopher David Hume (1711–1776) launched the scientific philosophy of *scepticism* whereby he asserted that knowledge can only be obtained through the study of empirical observations and data. He thus rejected inductive logic as a means to gain knowledge: if inductive logic can allow a possibly false conclusion, how can it be a source of knowledge? Fortunately, the great theoretical scientists of the eighteenth, nineteenth and twentieth centuries (Maxwell, Einstein, Dirac, Boltzmann, ...) ignored this point of view and went on to vastly increase our knowledge of the universe through miraculous advances in theoretical science. The paradox between experimental evidence and scientific theory was more or less put to rest by the philosopher Karl Popper (1902–1994) who introduced the *principle of falsification*. According to this principle, any scientific claim, any hypothesis put forth as a possible scientific theory, cannot qualify as a scientific theory unless it is possible that it be falsified, that is, it can possibly be disproved when new conflicting information is acquired. Quite the opposite to Hume, famous astronomer Arthur Eddington (1882–1944) announced that 'one should never accept any experimental data that has not been confirmed by theory' – an affront to Hume's philosophy and an acknowledgement of

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

the great advances in theoretical physics made by his associate Paul Dirac, who discovered antimatter through purely theoretical arguments four years before it was observed experimentally. To the great probabilist and physicist E. T. Jaynes (1922–1998), all knowledge is inductive: inductive logic must precede the performance of meaningful experiments or the interpretation of observations of natural events. Coveney, Dougherty and Highfield (2016, p. 3) assert that 'causality is not part of science'; and 'science concerns the mathematical description of behaviour': 'reason guiding experiment' (as emphasized by Jaynes 2003), 'new reason and then experiment'. It is exactly this strategy that we follow in model validation and selection in later sections.

Today it is generally recognized that knowledge is acquired through observation and theory, the two classical pillars of science. These are the foundations of the scientific method, the body of techniques for investigating phenomena and acquiring new knowledge based on *observation* (empirical or measurable evidence) and *theory* (the formulation, testing, and modification of mathematical models based on inductive logic laid down to explain observed or perceived events).

Throughout this work, theory is presumed to be characterized by mathematical models or models derived from surrogate models that arise from simplifying and subjective assumptions. When possible, such models are generally based on long-standing scientific principles complemented by empirical laws, all expressed in mathematical constructions.

A third pillar of science, *computational science*, is recognized as a new means of acquiring knowledge through computer simulations where experiments are impossible or inadequate for acquiring observational data. There are countless examples in which computer simulations based on models derived from accepted physical laws and empirical relationships provide a path to knowledge, assuming that the models are 'reliable' in some sense. One issue to be addressed in some depth here is model validity, taken up in Part 2.

These cornerstones of predictive science – the collection and interpretation of observational and experimental data, the creation and use of theory and models, and the use of computer simulations – are by no means independent: they are intrinsically intertwined and inseparably connected. Inductive reasoning precludes the design of experiments: theory and models underlie the design and construction of instruments and measuring devices, and data and mathematical abstractions of theoretical models form the foundation of computer simulations. All are fundamental to science-based prediction.

Efforts to expand and enrich the processes of prediction to take such uncertainties into account has gradually begun to reshape much of scientific and mathematical research over the last two decades, and particularly over the last decade, and a large and growing literature is devoted to the

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

subject. As representative, see AIAA (1998), ASME (2006), Oden (2014), Roache (1998), Oberkampf and Roy (2010), Babuška, Tempone and Nobile (2008), Szabó and Babuška (2011), Babuška and Oden (2004), Ainsworth and Oden (2000), Oden *et al.* (2018), Oden (2017) and Oden, Moser and Ghattas (2010). Predictive computational science concerns the development of methods for handling randomness in observational data, model selection, model calibration, model validation, and verification of computational models in the presence of uncertainties. It therefore embraces the development of methods for solving the forward prediction problem for target outputs, the quantities of interest, methods for solving statistical inverse problems for parameter identification, and quantifying the uncertainty in the predictions. In the jargon of early literature on coping with uncertainties, the field was given the acronym 'VV & UQ' (validation, verification and uncertainty quantification). More recently, this label has been truncated to simply 'UQ' – uncertainty quantification – a simplification that some say has unfortunately moved attention away from some of the most important issues of all: model selection and validation.

This exposition presents a broad overview of the field of adaptive multiscale predictive modelling in three parts. In Part 1, an introductory account of what we refer to as predictive computational science is given. This includes a summary of ideas underlying the theory of logical probability of Cox and Jaynes as interpreted by Van Horn and others, that leads logically to a Bayesian framework for dealing with uncertainties in predictive modelling. Bayesian approaches to model calibration, validation and model selection are then discussed, together with a brief look at frequentist approaches. Part 2 contains a review of fundamental principles and rules of model validation, an introduction to information entropy and the Maximum Entropy Principle, a very brief review of Monte Carlo methods, Markov chain Monte Carlo methods, and parameter sensitivity analysis. Part 3 covers methods of model selection, adaptive modelling, *a posteriori* estimates of modelling errors, and methods of adaptive multiscale modelling. Here OPAL, the Occam Plausibility Algorithm, is presented as a framework for systematically implementing many of the rules and principles of model selection, calibration and validation. Applications to the selection and validation of complex models of tumour growth are described as a canonical example of how various tools and theories covered can be implemented to identify predictive models of complex physical events.

2.1. Mathematical models

At the heart of predictive computational science is the idea of a *mathematical model* of a class of physical events of interest, a mathematical abstraction of reality generally based on a combination of scientific principles and empirical

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

laws. Mathematical models stand as surrogates or as representatives of theories – the mathematical abstractions designed to approximate perceived realities – and to provide the essential component in prediction of future events.

It is convenient (indeed, essential for our purposes) to consider mathematical models as representing an abstract mathematical problem, such as the following. Given a vector $\boldsymbol{\theta}$ and a domain S in which the problem is posed, find a function $u(\boldsymbol{\theta}, S) \in U$ such that

$$\mathcal{A}(\boldsymbol{\theta}, S; u(\boldsymbol{\theta}, S)) = 0, \tag{2.1}$$

where $\mathcal{A}$ is a collection of operators (algebraic, differential, partial-differential, integral, ...), including possibly constraints, boundary and initial conditions, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$ is a vector of model parameters taken from a parameter space Θ, S is the scenario in which the problem is posed, and $u(\boldsymbol{\theta}, S)$ is the solution of equation (2.1) which belongs to a function space U for specific $\boldsymbol{\theta}$ and S.

The notion of a scenario is intended to describe the features of the mathematical problem which can be exactly specified, independent of parameters, such as the domain of the solution and certain boundary and initial conditions, the idea being that the same model may be used in several different scenarios. We use the term *scenario* interchangeably to refer to both the domains in which the mathematical model (2.1) is solved and the actual physical environments in which experimental data are collected, or in which the target quantities to be predicted reside. Later we will argue that it makes sense to consider a hierarchy of scenarios, beginning with component or unit calibration tests in scenarios S_c to initially adjust model parameters, then moving to more complex subsystem validation scenarios S_v to test the accuracy with which the model can predict observables (to see if the model is 'valid'), and then the prediction scenario S_p, involving the full physical system and model, in which the forward problem (2.1) is solved to predict the *quantities of interest* (QoIs). A revealing example of various scenarios emerges in models of biomedical systems in which *in vitro* data are selected in laboratory experiments (*e.g.* cell proliferation in a Petri dish supplied with concentration of nutrients, or *in vitro* experiments with laboratory animals) and the models are there expected to be validated using *in vivo* data collected through biopsies or emerging imaging techniques. These are the calibration scenarios S_c (the *in vitro* experiments) and the validation scenarios S_v (the *in vivo* experiments). The model may then be called upon to make an out-of-data prediction in the prediction scenario S_p. We discuss these ideas in more detail later.

To develop a framework for addressing uncertainties, we accept the idea that the model parameters $\boldsymbol{\theta}$ may generally be random vectors. Thus, the solutions $u(\boldsymbol{\theta}, S)$ are random or stochastic functions. Problem (2.1),

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

therefore, describes an infinite class of models parametrized by $\boldsymbol{\theta}$. The mathematical problem characterized by (2.1) is called the *forward problem*. It describes a projection of data 'forward' into a prediction $u(\boldsymbol{\theta}, S)$. Thus, $u(\boldsymbol{\theta}, S)$ is the solution of the forward problem. Given observational data, and a mathematical model, the problem of determining the model parameters that best fit the data, in some statistical sense, is the (Bayesian) inverse problems, which we take up in Part 2.

Two basic issues arise at this point.

(i) Solutions $u(\boldsymbol{\theta}, S)$ to the forward problem must exist for every $\boldsymbol{\theta} \in \Theta$. This often-ignored fact leads to conditions on the space Θ of parameters. Generally, the solution will be a member of an appropriate space U of random functions.

(ii) The fundamental goal of constructing and solving the model (2.1) is not only to determine the solution of the forward problem, but to use these solutions to compute specific QoIs, generally described as functionals Q on the space U containing the solution. This calculation is made in the full prediction scenario, $S = S_p$. Thus, S_p pertains to the full system in which the model extrapolates all relevant information to predict an outcome beyond those events available through the analysis of observational data:

$$Q : U \to \mathbb{R} : Q(u(\boldsymbol{\theta}, S_p)) = \bar{Q}(\boldsymbol{\theta}). \tag{2.2}$$

Clearly, $\bar{Q}$ is, in general, a random function.

The determination of the validity of the model (2.1) as a means of predicting $\bar{Q}$ of (2.2) must, by definition, involve determining the accuracy with which model predictions agree with experimental data – the observations. These data are collected in a *model validation scenario* S_v, which is characterized by an experimental set-up designed to, hopefully, represent how well the model can predict the QoI. As noted earlier, at a lower unit-test level, initial characterizations of model parameters are done in calibration scenario S_c, wherein experimental data are used to adjust parameters so as to make model predictions agree with experimental measurements. The full calibration–validation–prediction process is symbolized in a 'prediction pyramid' described later (see Figure 5.1).

2.2. Sources of uncertainty in predictions

So what is the anatomy of scientific prediction within the context laid down thus far, and how do we cope with the multitude of uncertainties in each step of the prediction process? We wish to predict a specific reality in a possibly very complex scenario S_p using a computational model which is based on a mathematical model of a real situation from which the target

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

QoI can be specified. We must therefore select a model that we hope can deliver a sufficiently accurate estimate of the QoI and for which we have some confidence in its predictability. The predictability of our model rests on knowledge gained through observation – and our access to observational data – and our inductive reasoning embodied in our construction of a model. The model, in general, will be characterized by a set of parameters which themselves are generally unknown, and which at best involve uncertainties. Then we must discretize the mathematical model to produce a computational model, a corruption necessary to render the model into a form that can be processed by digital computers.

Collecting all these sources of uncertainty, and following, in part, the classification plan of Terenin and Draper (2015), we list the following sources of uncertainty in predictive computational science and in making prediction of realities.

(1) $\mathcal{P}$ *uncertainty.* This is so-called *systemic uncertainty* – uncertainty concerning the appropriate system of reasoning or the mathematical framework in which to study and quantify uncertainty in a logically consistent way. For example, do we frame our predictive process in the language of probability theory (if so, which one?), possibility theory, fuzzy sets, Dempster–Shafer theory or interval analysis? Which logical system produces a self-consistent way to deal with uncertainty in prediction?

(2) $\mathcal{Y}$ *uncertainty.* This embodies uncertainty in observational data. All data recorded from physical experiments can be polluted by experimental noise. How do we cope with uncertainties in data?

(3) $\mathcal{M}$ *uncertainty.* One of the most elusive and important sources of uncertainty is the selection of the mathematical model itself: $\mathcal{M}$ uncertainty. We agree with the famous quote of George E. P. Box (Box 1979, Box and Draper 1987): 'Essentially, all models are wrong, but some are useful.' One of the most important missions of model selection and validation processes is to identify such useful models.

(4) $\boldsymbol{\theta}$ *uncertainty.* Model parameter $\boldsymbol{\theta}$ uncertainty: the uncertainty in parameter values or in parametric families of distributions – a key source of uncertainty in predictive models.

(5) h *uncertainty.* This is uncertainty brought about by the discretization process that produces the computational model – 'h' representing an appropriate discretization parameter. This uncertainty can sometimes be estimated *a posteriori* through error estimation or bounded through *a priori* estimates or at least studied in some quantitative sense through various verification processes.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

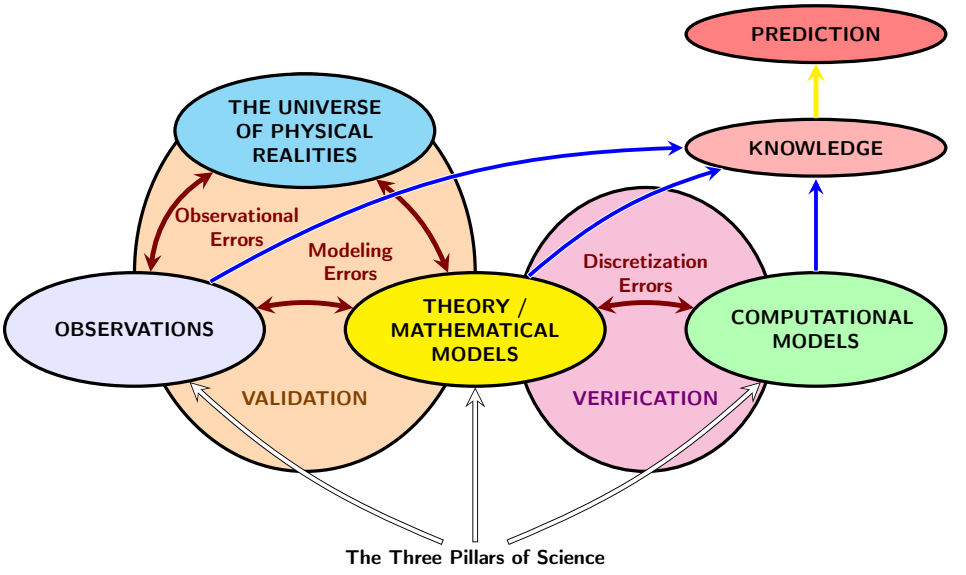


Figure 2.1. The imperfect paths to knowledge and prediction. A schematic of the imperfect paths to knowledge that involve uncertainties in all phases of the process for gaining knowledge and making predictions. (From Oden *et al.* 2010.)

The various paths to knowledge, made imperfect by the uncertainties discussed, are represented symbolically in Figure 2.1.

To lay the philosophical foundations for our view of systemic ($\mathcal{P}$) uncertainty, we provide a review of the logical theory of probability of Cox and Jaynes, which points to a Bayesian framework for studying most of the sources of uncertainty. Then we review briefly the Fisher/frequentist approach to problems in statistical analysis.

3. The logical theory of probability of R. T. Cox and E. T. Jaynes

We are familiar with the 'classical' (frequentist) theory of probability of Kolmogorov built on the notion of a triple representing a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, consisting of the set Ω of all possible outcomes of an action or experiment designed to characterize random events, the sigma algebra $\mathcal{F}$ embodying the set of all events, each containing possible outcomes $\omega \in \Omega$ or no outcomes at all ($\emptyset \in \mathcal{F}$), and a probability measure $\mathbb{P}$ to reflect the frequency with which an outcome will be observed over an infinite number of samples, thus inherently appealing to an infinite sequence of samples and to a limit process depending on certain topological properties.

An alternative and quite different mathematical framework to Kolmogorov's for developing a theory of probability was begun by R. T. Cox (1946),

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

expanded on in subsequent accounts (Cox 1961, Cox 1979), and enriched and extended by E. T. Jaynes in his treatise *Probability Theory: The Logic of Science*, published posthumously (Jaynes 2003). A significant literature has accumulated on this so-called subjective or logical theory of probability based on Bayesian inference, including works on the Bayesian approach to scientific reasoning by Howson and Urbach (2006), counterexamples to Cox's original examples by Halpern (1999*a*), 'patches' (corrections) to the theorem by Halpern (1999*b*) and alternative proofs and interpretations by Arnborg and Sjodin (2001), Knuth and Skilling (2012), Zimmermann and Cremers (2011) and Dupre and Tipler (2009). A 'guide to Cox's theorem' was provided by a comprehensive account of Van Horn (2003), and a recent paper by Terenin and Draper (2015) is devoted to 'rigorizing and extending the Cox–Jaynes derivation of probability'. We briefly outline the essential features of Cox–Jaynes theory, drawing from Jaynes (2003), Van Horn (2003) and Terenin and Draper (2015).

3.1. Plausible reasoning

The notion of logical or Cox–Jaynes probability begins with the concept of *plausible reasoning*. The idea is that in science, and indeed, in everyday life, we rarely have enough information to permit deductive reasoning. Instead, we make plausible conclusions based on past experience. Certain reasoning processes assess the plausibility of a proposition accepting that it may not be a certainty. Thus, we regard the logical approach to scientific reasoning as one way of assessing the plausibility of a proposition (as opposed to the frequency of an event) that can be true or false or that can be assigned a degree of plausibility.

The appropriate mathematical setting for such considerations is that furnished by a *Boolean algebra*.

Propositional Boolean algebra. A propositional Boolean algebra C consists of the 6-tuple

$$C \sim \{\mathcal{A}, \wedge, \vee, \neg, \mathbf{0}, \mathbf{1}\}, \tag{3.1}$$

wherein

- $\mathcal{A}$ = a set of true propositions, which might be finite, countable or uncountable;
- two binary operations, $\wedge$ and $\vee$, and a unitary operation $\neg$ on the propositions in $\mathcal{A}$, *i.e.*,
  - $A \wedge B$ = the *conjunction* of A and B, meaning that both A and B are true,
  - $A \vee B$ = the *disjunction* of A and B, meaning that at least one of A and B is true ('A or B'),

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

– $\neg A = A\neg =$ the negation of A, meaning A is false;

- two identity elements **0** and **1**.

We also require the following.

(i) C is *closed* under countable applications of $\wedge$ and $\vee$.

(ii) For all propositions $(A, B, C) \in \mathcal{A}$, the following five Boolean axioms hold.

Associative laws:

$$A \wedge (B \wedge C) = (A \wedge B) \wedge C, \quad A \vee (B \vee C) = (A \vee B) \vee C.$$

Commutative laws:

$$A \wedge B = B \wedge A, \quad A \vee B = B \vee A.$$

Identity elements:

$$A \wedge \mathbf{1} = A, \quad A \vee \mathbf{0} = A.$$

Complementation or negation:

$$A \wedge (\neg A) = \mathbf{0}, \quad A \vee (\neg A) = \mathbf{1}.$$

Distributive laws:

$$A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C),$$
$$A \vee (B \wedge C) = (A \vee B) \wedge (A \vee C).$$

The conjunction of two propositions $A \wedge B$ is also written AB, *i.e.* as a 'product', and the disjunction $A \vee B$ is written as the sum $A + B$, all in analogy with the ('two-valued') algebra of numbers. In this context, the Boolean algebra of propositions is said to follow the *rules of Aristotelian logic*.

The plausibility of a proposition A, given that a proposition B is true, is characterized by a real-valued function $pl(.|.)$ on pairs of propositions:

$$pl : \mathcal{A} \times \mathcal{A}/\mathbf{0} \quad \to \mathbb{R}^+. \tag{3.2}$$

Thus $pl(A|B)$ is the *degree of plausibility of proposition A given that B is true*. If B is true, A becomes plausible, while evidence does not guarantee that A is true, but by observation of one or more of its consequences gives confidence that A is plausible. Also, if, given B, A is more plausible than C, we have $pl(A|B) > pl(C|B)$. Other fundamental rules will follow.

Jaynes' example (Jaynes 2003). Suppose we state the proposition $A \sim$ 'it will start raining at 10.00 a.m. at the latest', and $B \sim$ 'the sky will become cloudy before 10.00 a.m.' Plausible reasoning not only establishes whether a proposition is more or less plausible, it evaluates the degree of plausibility in some way. For example the darkness of the clouds may suggest the degree of plausibility of rain.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

3.2. Logic and philosophical theories of probability

Logic is the science of reasoning, an intellectual discipline with origins dating back millennia, that provides the cornerstones to mathematics, science and all forms of human knowledge.

As noted in Section 2, deductive logic or logical deduction is a form of reasoning from one or more premises or axioms so as to deduce a logically certain conclusion. Deductive logic is infallible and irrevocable when its conclusions are reached by following strict laws of reasoning. Mathematics is a discipline founded on deductive logic: one lays down axioms and definitions and then deduces their consequences following rules of logic. According to Jaynes (2003), however, all knowledge is inductive. In deductive reasoning, if A and B are elements or propositions, one writes $A \Rightarrow B$ to mean if A is true, B is logically deduced from A. According to Jaynes (2003, p. 4), in almost all situations, 'we do not have the right kind of information to allow this kind of reasoning'. In the formal logic of plausible reasoning $A \Rightarrow B$ means only that A and B have the same 'truth value'.

If A and B are true propositions, a proposition C which is a function f of A and B, $(C = f(A, B))$, can be either true or false. Since A and B take on values $[TT, TF, FT, FF]$, there are 16 logical functions: $f(T, T) = T, f(T, T) = F, \ldots$ *etc.* Thus a logical function of n propositions has 2^n function values. One can therefore construct truth tables analogous to the theory of elementary logic of operations on real numbers, to cover possible truth values of such logic functions.

Plausible reasoning not only establishes whether a proposition is more or less plausible, it evaluates the degree of plausibility in some way (*e.g.* the darkness of the clouds may suggest the plausibility of rain). Examples of the Boolean algebra of propositions are as follows.

- $A, B, C, \ldots$ = propositions = unambiguous, simple, definitive, statements that are either true or false (two-valued Aristotelian logic)
- As noted earlier, conjunction 'cojoins' two propositions into a 'logical product':
$$A \wedge B = AB \quad \text{(both } A \text{ and } B \text{ are true)}.$$
- Disjunction, called the 'logical sum', 'disjoins' two propositions:
$$A \vee B = A + B \quad \text{(at least one of } A \text{ and } B \text{ are true)}.$$
- Negation describes 'denial':
$$\neg A \equiv A \quad \text{is false.}$$
- We understand equality and identity of propositions as
$$A = B \quad (A \text{ and } B \text{ are equal in truth (or 'truth value'))},$$
$$A \equiv B \quad (A \text{ and } B \text{ are equal by definition}).$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

Some properties:

- $A \vee \neg A =$ a necessary (or logical) truth (A is either true or false),
- $A \wedge \neg A =$ a necessary (or logical) falsehood (A is true and false),
- $\neg A \wedge B = AB$ is false,
- $\neg A \wedge \neg B =$ both A and B are false,
- if $C = A \wedge B$, then $\neg C = \neg A \vee \neg B$,
- if $D = A \vee B$, then $\neg D = \neg A \wedge \neg B$.

These properties are recognized as analogous to elementary set properties with A, B, C, D sets, $AB = A \cup B$, $A + B = A \cap B$, $\bar{A} = A'$, *etc.* $A \Rightarrow B$ (A implies B) does not assert that either A or B is true, it means $A \wedge \bar{B}$ (A and (B is false)) is false or $\bar{A} \vee B$ (either A is false or B is true) is true.

3.3. Van Horn's guide to Cox's theorem

It is informative to follow a few arguments given by Van Horn in his guide to Cox's theorem (Van Horn 2003). There are two central notions.

(i) The *state of information* X summarizes the information one has on what is called the 'atomic proposition Λ': those that cannot be decomposed into other propositions. These atomic propositions form the *basis* for the state of information X.

(ii) The *plausibility of a proposition* $A \in \mathcal{A}$, given the state of information X, is a number $pl(A|X)$: the state of information obtained from A by adding the additional information that X is true, thus written (AX) or $(A|X)$.

We may now lay down several assumptions. In the notation inspired by Van Horn, there are five 'rules', R1, ..., R5.

R1. The plausibility $pl(A|X)$ of A given X is a single non-negative real number. Moreover, there exists a number T (the plausibility of some known true proposition) such that $pl(A|X) \leq T$. We say that X is *consistent* if there is no proposition A for which both $pl(A|X) = T$ and $pl(\neg A|X) = T$ (are true).

R2. Plausibilities are consistent with so-called propositional calculus, in the sense that the following hold.

(i) If $A \Leftrightarrow A'$, meaning that '$A \Leftrightarrow B$' is a *tautology* (meaning a fundamental truth in formal logic), then $pl(A|X) = pl(A'|X)$.

(ii) If A is a tautology, then $pl(A|X) = T$.

(iii) $pl(A|B, C, X) = pl(A|(B \wedge C), X)$.

(iv) If X is consistent and $pl(\neg A|X) < T$, then A, X is also consistent.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

R3. There is a non-increasing function S such that $pl(\neg A|X) = S(pl(A|X))$ for all A and consistent X.

R4 (universality). There exists a non-empty set $Y_\circ$ that is a dense subset of $(S(T), T)$. Thus, for real numbers $a, b, c \in Y_\circ$ there is a consistent X and three atomic propositions A_1, A_2, A_3 such that $a = pl(A_1|X)$, $b = pl(A_2|A_1, X)$ and $c = pl(A_3|A_2, A_1, X)$.

Rules R1–R4 are set up to establish more or less commonsense properties of the plausibility of a proposition. R1 establishes the reasonable notion that real numbers can be used to represent degrees of plausibility, and that larger numbers T represent higher degrees of plausibility, representing the plausibility of certain known true propositions. The idea that for a given state of information X, two plausibilities A and B are equally plausible, or one is more plausible than the other, is an assumption of *universal comparability*. The idea is not without some controversy. Some (*e.g.* Shafer 1976) propose that a 'two-dimensional' logic may be called for, in which both the 'degree of belief' of a proposition as well as the belief in its negation is justified. Jaynes (2003) and Van Horn (2003), among others, provide strong arguments in favour of R1.

For R2, we simply demand that if two propositions are equivalent, we can use them interchangeably, and that the plausibility of any proposition known to be a tautology should be assigned a value reflecting this knowledge. Condition (iii) in R2 is obvious, and in (iv), if we cannot say with certainty that the negation $\neg A$ of A is true, there remains the fact that A may be true and, then, cannot contradict assertion on $\neg A$.

In R3 (negation), we demand that A and $\neg A$ are opposite answers to the same question: if A is known to be true, then $\neg A$ is false: $pl(A|X)$ and $pl(\neg A|X)$ cannot vary independently in two-dimensional Aristotelian logic.

In R4, the goal is to lay down rules that enable plausible reasoning to fall under a universal system of logic applicable to any problem domain in which meaningful propositions can be formulated, allowing the range of the set of allowed plausibilities to be dense in $(S(T), T)$). Van Horn (2003) comments that Halpern (1999*b*) argues that R4 is problematic as it cannot be satisfied on finite domains, an assertion Van Horn discounts. We return to this point later.

Conjunction (the product rule of plausibilities). We now come to the centrepiece of Cox's theorem, the conjunction or product rule for propositions. Given consistent X and the conjunction $A \wedge B = AB$ of two propositions conditioned on X, what can be said about the relationship of $pl(A \wedge B|X)$ with the plausibilities $pl(A|B, X)$ and $pl(B|X)$? A systematic list of arguments leads to the following.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

R5. Given $S(T)$ and a tautology T, there exists a continuous function $F : [S(T), T]^2 \to [S(T), T]$, strictly increasing on $(S(T), T]$, such that

$$pl(A \wedge B|X) = F\big[pl(A|B, X), pl(B|X)\big] \tag{3.3}$$

for any A, B and consistent X.

The establishment of (3.3) follows some surprisingly intuitive arguments. We begin by writing down all the combinations of A, B and X that $pl(A \wedge B|X)$ could be a function of (omitting 'pl' temporarily, for simplicity), they are $(A|X)$, $(B|X)$, $(A|B, X)$ and $(B|A, X)$ and all combinations of these numbers, 15 in all. Because $A \wedge B = B \wedge A$, only nine of these survive. Van Horn (2003), following Tribus (2016), rules out as many of these as possible:

(1) $(A \wedge B|X) = F[(A|X)]$,
(2) $(A \wedge B|X) = F[(A|BX)]$,
(3) $(A \wedge B|X) = F[(A|X), (A|BX)]$,
(4) $(A \wedge B|X) = F[(A|BX), (B|AX)]$,
(5) $(A \wedge B|X) = F[(A|X), (B|X)]$.

Remarkably, all five of these candidates lead to a contradiction and can, therefore, be discarded. For example, in (1), if A is a tautology (a fundamental truth) and B is a 'atomic' proposition, then $A \wedge B$ is equivalent to B and $(A|X) = T$, so $S(T) = (B|X)$. But X can always be chosen to make $(B|X)$ any value in $Y_\circ$. So we have a contradiction. Therefore (1) cannot hold. Similarly, in (3), for instance, if A is a tautology and B is any atomic proposition, then $S(T) = (B|X)$, a contradiction. Similar arguments can be made for all these candidates leaving only four possibilities out of the original 15:

(I) $(A \wedge B|X) = F[(A|BX), (B|X)]$,
(II) $(A \wedge B|X) = F[(A|B, X), (B|X), (A|X)]$,
(III) $(A \wedge B|X) = F[(A|B, X), (B|X), (B|A, X)]$,
(IV) $(A \wedge B|X) = F[(A|B, X), (B|X), (B|A, X), (A|X)]$.

Van Horn refers to the arguments of Tribus (2016) who claims that (III) and (IV) can be ruled out because 'one cannot avoid dealing with plausibility expressions that involve some inconsistent state of information Y, and hence there is unavoidable ambiguity' (Van Horn 2003, p. 16). Also, rule (II) is eliminated for certain similar reasons, leaving only (I). But Van Horn asserts that 'there is no completely compelling reason for choosing any particular one of the four remaining candidates'. He leans toward (I) as the other candidates merely add additional conditions to those used in (I) and because Jaynes (2003) provides an intuitive argument in support of (I).

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

Assuming (I) holds, next consider the plausibility $A \wedge B \wedge C|X)$, X being consistent. Applying rule (I), with 'B' now replaced by $(B \wedge C)$, gives

$$\begin{aligned}(A \wedge (B \wedge C)|X) &= F[(A|(B \wedge C), X), ((B \wedge C)|X)] \\ &= F[(A|B, C, X), F[(B|C), X), (C|X)]]. \end{aligned} \tag{3.4}$$

Now use the equivalent grouping of the propositions on the left-hand side:

$$\begin{aligned}((A \wedge B) \wedge C|X) &= F[((A \wedge B)|C, X), (C|X)] \\ &= F[F[(A|B, CX), (B|C, X)](C|X)]. \end{aligned} \tag{3.5}$$

But $A \wedge (B \wedge C) = (A \wedge B) \wedge C$, so the functions derived in (3.4) and (3.5) must coincide. Setting

$$\begin{aligned} x &= (A|B, C, X), \\ y &= (B|C, X), \quad \text{and} \\ z &= (C|X), \end{aligned}$$

where $x, y, z \in [0, T]$, we are looking for a function F that for numbers x, y, z is such that

$$F(x, F(y, z)) = F(F(x, y), z). \tag{3.6}$$

Since F is continuous and *dense*, this relation holds for all $x, y, z \in [0, T]$. This property of denseness is fundamental to Cox's theorem. The fact that '$\wedge$' is associative ensures that F is associative, which limits choices for F to be 'isomorphic to multiplication'. A little inventory: if we can find any function F satisfying (3.6), then, by rule (I), we will have determined how the 'product' $A \wedge B$, conditional on X, depends on $(A|BX)$ and $(B|X)$. Thus, we establish (3.3).

Equation (3.6) is called the *associativity equation*; it is an example of a *functional equation*. Such equations are studied in Aczel (1966), where the following theorem can be found.

Theorem 3.1 (Aczel's theorem). Let $f : (a, b]^2 \to (a, b]$ be a continuous, strictly increasing function in both arguments a and b, with $a, b \in \mathbb{R}$ and $a < b$, and suppose that for all $x, y, z \in (a, b]$,

$$f(x, f(y, z)) = f(f(x, y), z).$$

Then there exist a continuous, strictly increasing function g such that

$$g(f(x, y)) = g(x) + g(y)$$

for all $x, y \in (a, b]$.

Now set $f = F$, $a = S(T)$, $b = T$ and define a real-valued, continuous, increasing, and non-negative function $w = w(x)$ on $(S(T), T]$ such that

$$\ln w(x) = g(x), \quad x \in [S(T), T].$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

with $w(S(T))$ the limit as $x \to S(T)$. We then must have by continuity

$$w(F(x,y)) = w(x)w(y),$$

for $x, y \in [S(T), T]$. Thus, for every $A, B \in \mathcal{A}$ and consistent X, we have

$$w((A \wedge B|X)) = w((A|B,X))w((B|X)). \tag{3.7}$$

The function w thus suggests that conditional probabilities obey a 'product rule' similar to probabilities (*e.g.* $\mathbb{P}(A \cup B) = \mathbb{P}(A|B)\mathbb{P}(B)$, now conditioned on information X).

A final step for the product rule calls upon the fact that

$$w(S(T)) = 0, \quad w(T) = 1, \quad \text{and} \quad 0 < w(x) < 1 \quad \text{for } S(T) < x < T.$$

A proof of this assertion is given in Van Horn (2003). Briefly, it involves the observation that for any proposition A, consistent X and tautology D,

$$w((A|X)) = w((D \wedge A|X)) = w((D|A,X))w((A|X)) = w(T)w((A|X)),$$

so dividing by $w((A|X))$ gives $w(T) = 1$ and, therefore, $w(x) < 1$ for $x < T$. The proof that $w(S(T)) = 0$ follows from contradiction: assume otherwise and consider $A_1, A_2 \in \mathcal{A}$ and $x \in Y_\circ$. If $w(S(T)) = y > 0$, then $0 < z < 1$ and $z < \sqrt{z} < 1$, $w^{-1}(z) < x < w^{-1}(\sqrt{z})$. If $(A_1|X) = (A_1, A_2|X) = x$, then $w((A_1 \wedge A_2|X)) = w^2(x) < z = w(S(T))$, so $(A_1 \wedge A_2|X) < S(T)$, a contradiction.

So we can add to (3.7) the condition that the monotone continuous function satisfying (3.7) must also satisfy

$$0 \leq w(x) \leq 1, \tag{3.8}$$

with $w(x) = 0$ when x is impossible and $w(x) = 1$ when x is certain.

Disjunction (the sum rule for plausibilities). The aim of the sum or disjunction rule for plausibilities defined on pairs of propositions is to establish a relationship involving the plausibility of a proposition A conditioned on consistent X and its negation $\neg A$ conditioned on X. Full details can be found in Van Horn (2003) and Paris (1994). It suffices here to point out that a thread of logic similar to that in the conjunction relations leads again to a functional equation of the form

$$yS\left(\frac{x}{y}\right) = S(x)S(S(y)/S(x)) \tag{3.9}$$

for all $0 < x \leq y < 1$, where

$$\begin{aligned} yS\left(\frac{x}{y}\right) &= p(\neg A \wedge B|X), \\ S(y) &= p((A \vee \neg B) \wedge (\neg A \vee \neg B)|X), \\ S(x) &= p(\neg A \vee \neg B|X), \end{aligned}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

$p(x) = w(x)^m$, $m = -\log(2)/\log w(\alpha)$, and α is a fixed point of a map such that $S(A|X) = (\neg A|X)$ for all A and consistent X. From these relations, it possible to show that

$$w^m(\neg A|X) + w^m(A|X) = 1, \tag{3.10}$$

for positive m. This is known as the *sum rule*. We note that the product rule (3.7) can be equally well be written as

$$w^m((A \wedge B|X)) = w^m((A|BX))w^m((B|X)),$$

for positive m. For whatever m we choose, we can set

$$pl(x) = w^m(x).$$

Collecting all of the results laid down thus far, we arrive at the following version of Cox's theorem.

Theorem 3.2 (Cox's theorem via Van Horn). Given that rules R1–R5 hold, there exists a continuous, strictly increasing function $pl : \mathcal{A} \times \mathcal{A}\backslash\mathbf{0} \to \mathbb{R}^+$ such that for every A and $B \in \mathcal{A}$ and consistent X:

(i) $pl(A|X) = 0$ if and only if A is known to be false, given the information X;

(ii) $pl(A|X) = 1$ if and only if A is known to be true given the information X;

(iii) $0 \leq pl(A|X) \leq 1$;

(iv) $pl(A \wedge B|X) = pl(A|X)\ pl(B|A, X)$;

(v) $pl(\neg A|X) = 1 - pl(A|X)$.

Because of the commutative property, $A \wedge B = B \wedge A$, assertion (iv) in Cox's theorem implies that

$$pl(A|X)pl(B|AX) = pl(B|X)pl(A|BX), \tag{3.11}$$

which is recognized as Bayes' rule, provided we interpret the plausibilities as probabilities.

The five properties laid dawn in Cox's theorem are formally the rules of probability theory, stated in terms of plausibilities. As pointed out by Van Horn (2003, pp. 21–22), 'any system of plausible reasoning that is not isomorphic to probability theory must necessarily violate one of the requirements' (in the theorem). So 'set-theoretical probability theory then becomes the model theory for our logic'. We may thus interpret states of information to be probability distributions, with AX understood to be 'the probability distribution obtained from X by conditioning the set of values for which A is true'. Henceforth, we refer to the Cox–Jaynes plausibilities as probabilities in the spirit of those covered in the previous section.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

Property (iv) sets the Cox–Jaynes theory apart, in a sense, from frequentist theory, featuring conditional probability and Bayes' rule as a central property of the theory. The existence of the Kolmogorov set-theoretic framework of probability, which involves the same rules, proves that these rules are not contradictory, the primary difference being that in Cox theory, conditional probability is now the fundamental measure. In the Kolmogorov theory described earlier, all probabilities are conditioned on the choice of the sample set Ω.

3.4. The Terenin–Draper account of Cox's theorem and Cox–Jaynes probability

Terenin and Draper (2015) provide a detailed account of the structure and different philosophical issues underlying major probability theories, particularly Kolmogorov's set-theoretic theory based on the frequentist definition of probability and the Cox–Jaynes theory based on the notions of plausible reasoning. Some key aspects of their arguments are summarized in the following paragraphs.

(1) In his book *Probability Theory: The Logic of Science*, Jaynes (2003, p. 43) insists that the framework of his theory is established for only *finite sets of propositions* (generally meaning it can be applied to statistical models with only a finite number of parameters), with 'extension to infinite sets ... permitted only when this is the result of a well-defined and well-behaved limiting process from a finite set'. Terenin and Draper show this position is 'overly cautious'. The Van Horn proof also does not assume the sets of propositions are necessarily finite.

(2) Terenin and Draper claim that Van Horn's notion of *state of information* is defined at a non-rigorous level, making it hard to evaluate. They go on to 'rigorize' the statement and proof of Cox's theorem.

(3) They invoke *Stone's Representation Theorem* (Stone 1936, p. 8), which they interpret to establish that 'every Cox–Jaynes propositional Boolean algebra $(\mathcal{A}, \wedge, \vee, \neg, \mathbf{0}, \mathbf{1})$ is isomorphic to a Kolmogorov set-theoretic space $(\mathcal{F}, \cap, \cup, \cdot^c, \emptyset, \Omega)$'.

(4) They invoke the conditions of *sequential continuity* for non-empty (or true) B:

(a) if $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$ such that $A_i \nearrow A$, then $\mathbb{P}(A_i|B) \nearrow \mathbb{A}|\mathbb{B}$,

(b) if $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$ such that $A_i \searrow A$, then $\mathbb{P}(A_i|B) \searrow \mathbb{A}|\mathbb{B}$.

(5) They provide a extension of the σ-algebra to resolve the requirement of denseness in the range of the probability function $\mathbb{P}$; specifically, for any subset M of $\mathbb{R}$, they define $S = [\min M, \max M]$ and Ω_E (= the extension

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

of the sample set $\Omega) = \Omega \times S$, and $\mathcal{F}_E$ (= the extension of the σ-algebra $\mathcal{F}$) $= \mathcal{F} \otimes \mathcal{B}(S)$, $\mathcal{B}(S)$ being the Borel σ-algebra on S. One can then define the extended probability function $\mathbb{P}_E : \mathcal{F}_E \times \mathcal{F}_E \, \emptyset \to S$ that satisfies all of the axioms for probabilities (plausibility) including the product rule and the sum rule. They justify that this density property 'defeats' the counterexample of Halpern (1999*a*). Further details can be found in Terenin and Draper (2015).

3.5. Summary comments

Among conclusions that can be drawn from the developments outlined in this section are the following.

- The Cox–Jaynes theory of logical probability, based on the notions of plausible reasoning and on plausibilities defined on propositional Boolean algebras, is a self-consistent theory, equivalent to the Kolmogorov theory, but structured in a framework in which conditional probability is the primitive concept. In this sense, there is no $\mathcal{P}$ uncertainty if this theory of probability is employed to deal with uncertainties.
- The natural extension of Aristotelian logic with uncertainties is Bayesian (see Jaynes 2003).

We are thus on solid ground to study the major components of predictive science – at least many of them – within a Bayesian framework.

4. Bayesian framework for statistical calibration and validation of parametric models

4.1. Bayes' rule for parametric models

Recall that in Section 2.2 we considered the predictability of models defining a forward problem with a solution $u = u(\boldsymbol{\theta}, S)$ depending on parameters $\boldsymbol{\theta} \in \Theta$ and set in a scenario S for a problem $\mathcal{A}(\boldsymbol{\theta}, S; u(\boldsymbol{\theta}, \boldsymbol{S})) = 0$, and we considered the question of predictability of such models in the presence of uncertainty, addressing the principal sources of uncertainty described: $\mathcal{P}$, $\mathcal{M}$, $\mathcal{Y}$, $\boldsymbol{\theta}$ and h uncertainty. We will now explore these uncertainties in a Bayesian formalism.

As noted in Sections 3.3 and 3.4, we can dispense with $\mathcal{P}$ uncertainty immediately. We argued that within the framework of Cox–Jaynes probability theory (and equivalently, Kolmogorov theory), there is no $\mathcal{P}$ uncertainty, and every natural extension of Aristotelian logic in the presence of uncertainties is Bayesian (Terenin and Draper 2015). We will therefore employ tools and also concepts from the equivalent Kolmogorov theory when convenient. This leaves us with $\mathcal{M}$, $\mathcal{Y}$, $\boldsymbol{\theta}$ and h uncertainties.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

Terenin and Draper (2015, p. 18) state that 'All uncertainty in probabilistic models is therefore the result of model specification: either via the parameters (θ-uncertainty), or via the structure of the model itself ($\mathcal{M}$-uncertainty).' We add to this uncertainty in the data ($\mathcal{Y}$ uncertainty) and, since we are interested in computational models obtained through discretization of mathematical models, h uncertainty as well.

This latter area, h uncertainties, is in the realm of *model verification* and is either an issue of code verification, requiring such things as documentation, version control, benchmarks, *etc.*, or solution verification, which is a matter of mathematically based *a posteriori* error estimation.

We will ignore h uncertainty in this article as it is dealt with in detail in other accounts (*e.g.* Ainsworth and Oden 2000). That leaves us with $\mathcal{M}$, $\mathcal{Y}$ and $\boldsymbol{\theta}$ uncertainties. $\mathcal{M}$ uncertainty is a deep and challenging subject involving the study of model selection, model inadequacy, and the complicated interaction of uncertainties in the model with those in data and parameters. We take up this important subject after first looking at $\mathcal{Y}$ and $\boldsymbol{\theta}$ uncertainty in Section 8. First, we address $\mathcal{Y}$ and $\boldsymbol{\theta}$, beginning with their appearance in Bayes' rule, where one finds naturally the push and pull of the two classical pillars of science: *observation*, with uncertain data $\boldsymbol{y} \in Y$, and *theory*, represented by unvalidated models with uncertain parameters $\boldsymbol{\theta} \in \Theta$.

Thus, from Bayes' rule, we covert to probability densities π, and from the rule $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$, we let B represent the observational data $\boldsymbol{y}$ for scenario S and A represent the model parameters $\boldsymbol{\theta}$, leading to

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}, S) = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}, S)\pi(\boldsymbol{\theta}|S)}{\pi(\boldsymbol{y}|S)}, \tag{4.1}$$

where $\pi(\boldsymbol{\theta}|S)$ is the *prior* density of parameters $\boldsymbol{\theta} \in \Theta$ for scenario S, $\pi(\boldsymbol{y}|\boldsymbol{\theta}, S)$ is the *likelihood*, $\pi(\boldsymbol{\theta}|\boldsymbol{y}, S)$ is the *posterior* density, and $\pi(\boldsymbol{y}|S)$ is the *evidence*. Since $\int_\Theta \pi(\boldsymbol{\theta}|\boldsymbol{y}, S)\,\mathrm{d}\boldsymbol{\theta} = 1$ for all $\boldsymbol{y}$ in the space $\mathcal{Y}$ of observational data and for all scenarios S (S_c, the calibration scenarios; S_v, the validation scenarios; S_p, the prediction scenario), the evidence must be the marginalization of the numerator,

$$\pi(\boldsymbol{y}|S) = \int_\Theta \pi(\boldsymbol{y}|\boldsymbol{\theta}, S)\pi(\boldsymbol{\theta}|S)\,\mathrm{d}\boldsymbol{\theta}. \tag{4.2}$$

Here $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n) \in \mathcal{Y}_s \subset \mathbb{R}^n$, and it is understood that (4.2) holds for every scenario S, the data $\boldsymbol{y}$ being different in each.

4.2. Data uncertainty: experimental noise

Bayes' rule (4.1) embraces all of the pillars of science drawn upon to acquire (or increase) one's knowledge: *observations*, characterized by observational data $\boldsymbol{y} \in \mathcal{Y}$, *theory*, the mathematical modelling of hypotheses founded

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

on inductive logic, put forth to explain observed events, and *computations*, needed to evaluate the resulting distributions. To the classical statistician, models are represented by the likelihood $\pi(\boldsymbol{y}|\boldsymbol{\theta}, S)$, $\boldsymbol{\theta}$ representing specific features of the model and $\pi(\boldsymbol{y}|\boldsymbol{\theta}, S)$ the probability distribution for data given $\boldsymbol{\theta}$, for any scenario S; the outcome, $\pi(\boldsymbol{\theta}|\boldsymbol{y}, S)$, is the result of the information in the data and the predictability of the model. To the scientist or engineer, this statistical model is reached after a process of using the solution of the forward problem to evaluate the observations available to the models. These predictions of $\boldsymbol{y}$ of the models are quantities $\boldsymbol{d}(\boldsymbol{\theta}, S)$ which rarely coincide with the recorded observations $\boldsymbol{y}$.

Let us consider the problem of using model (4.1) for a given scenario S to predict the value of some physical entity. The target prediction could be, for example, the temperature at a point in a turbine engine, the volume fraction of cancer cells in a sample of human tissue, the maximum von Mises stress in a machine part, or the time-to-failure of a steel shaft in an aircraft engine. Such quantities are rarely deterministic. A fundamental assumption in classical mathematical statistics is that the physical entity we wish to predict with model (4.1) in any particular scenario S is determined by a stochastic law (the 'true' structure of the entity) characterized by a probability distribution $\boldsymbol{G}$ with density $\boldsymbol{g}$. The structure of $\boldsymbol{g}$ (or $\boldsymbol{G}$) is available to us only through samples taken from a set of experiments, the experimental observations $\boldsymbol{y}$, which are generally corrupted by experimental noise $\boldsymbol{\epsilon}$ (or experimental errors), since no measuring device is perfect. The observations $\boldsymbol{y}$ are realizations of a random variable Y.

But the observational data we record is not the truth; rather it is some function $\boldsymbol{f}$ of the true density $\boldsymbol{g}$ and of unknown experimental errors $\boldsymbol{\epsilon}$, often called the *experimental noise*:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{g}, \boldsymbol{\epsilon}) \tag{4.3}$$

(or the so-called stochastic extension $Y = f(\boldsymbol{G}, E)$). To go further, we must assume (or construct) an appropriate form of $\boldsymbol{f}$ and propose a probability distribution $p_{noise}(\boldsymbol{\epsilon})$ for $\boldsymbol{\epsilon}$. These constitute a *noise model*. The construction of an appropriate noise model for observational data is often a challenging endeavour requiring detailed information on the measurement processes and instrumentation. This may often mean we must construct a new parametric model of $\boldsymbol{g}$, knowing very little about the 'truth'. But there are approaches, some discussed later in this article, of constructing approximations of $\boldsymbol{g}$ that may improve in accuracy as new data are acquired. These approximations involve new parameters, not the parameters $\boldsymbol{\theta}$ of the model of the forward problem – sometimes called hyperparameters. Methods for estimating these parameters and enhancing our approximation of $\boldsymbol{g}$ are discussed later (*e.g.* Section 5.3).

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

In the majority of statistics applications, the most common noise model is the additive model

$$\boldsymbol{g} + \boldsymbol{\epsilon} = \boldsymbol{y}, \tag{4.4}$$

showing an additive correction to the truth is what is observed; the noise distribution $p_{noise}(\boldsymbol{\epsilon})$ is often assumed to be Gaussian. The extra parameters in the definition of $p_{noise}(\cdot)$ may be added to the model parameters to produce the total parameter vector $\boldsymbol{\theta}$. If we have n samples of $\boldsymbol{g}$, (4.4) is replaced by

$$\boldsymbol{g}_i + \boldsymbol{\epsilon}_i = \boldsymbol{y}_i, \quad 1 \leq i \geq n. \tag{4.5}$$

4.3. Model inadequacy

The experimental noise in general is a result of the methods and devices used to gather observational data and is independent of the model and of model parameters. Now let us consider how well the model predicts realities of various scenarios.

Having solved the forward problem for scenario S, use the solution $u(\boldsymbol{\theta}, S)$ to compute the model predictions of reality $\boldsymbol{g}_i$ for parameter $\boldsymbol{\theta}$. We denote such predictions by a parameter-to-observation map $B : U \to \mathcal{Y}$ such that $B_i(u(\boldsymbol{\theta}, S)) = d_i(\boldsymbol{\theta}, S)$. Since the model is imperfect, a modelling error $\gamma_i(\boldsymbol{\theta}, S)$ occurs, assumed for simplicity to be additive:

$$\boldsymbol{g}_i = \boldsymbol{d}_i(\boldsymbol{\theta}, S) + \boldsymbol{\gamma}_i(\boldsymbol{\theta}, S), \quad 1 \leq i \geq n. \tag{4.6}$$

where $\boldsymbol{\gamma}_i(\boldsymbol{\theta}, S)$ is defined as the *model inadequacy*, or *model form error*, or *model bias* (AIAA 1998, ASME 2006, Oberkampf and Roy 2010, Oden *et al.* 2018, Kennedy and O'Hagan 2001). This can be regarded as a component of model inadequacy, to be complemented by an assumed probability distribution $p_{model}(\boldsymbol{\theta}, S)$ for $\boldsymbol{\gamma}_i(\boldsymbol{\theta}, S)$ (see *e.g.* Bayarri *et al.* 2007, Jeffreys 1998, Kennedy and O'Hagan 2001).

Substituting (4.5) into (4.6) and rearranging terms gives

$$\boldsymbol{\epsilon}_i + \boldsymbol{\gamma}_i(\boldsymbol{\theta}, S) = \boldsymbol{y}_i - \boldsymbol{d}_i(\boldsymbol{\theta}, S), \quad 1 \leq i \geq n, \quad \boldsymbol{y}_i \in \mathcal{Y}_s, \quad \boldsymbol{\theta} \in \Theta. \tag{4.7}$$

Thus, *reality disappears from the equation! One can never compute the reality* $\boldsymbol{g}_i$. The closest we can arrive at the truth $\boldsymbol{g}_i$ is through imperfect observational data and through imperfect model predictions, as indicated in Figure 2.1 discussed earlier.

Given a noise model, such as (4.5), with a noise probability density $p_{noise}(\cdot)$, and given a model inadequacy model such as (4.6) with density $p_{model}(\cdot)$, one can conceive of a combined model $p_{com} \sim p_{noise+model}$. Then, given (4.7), the probability that the data $\boldsymbol{y}_i$ are observed for given $\boldsymbol{\theta}$ is the likelihood probability density

$$\pi(\boldsymbol{y}_i|\boldsymbol{\theta}, S) = p_{com}(\boldsymbol{y}_i - \boldsymbol{d}_i(\boldsymbol{\theta}, S)), \quad \text{for } 1 \leq i \leq n. \tag{4.8}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

Until recently, in the great majority of statistical applications, model inadequacy was ignored: $p_{com}(\cdot) = p_{noise}(\cdot)$, and $p_{noise}(\cdot)$ was assumed to be Gaussian.

4.4. Frequentist and Fisherian approaches

Statistics can be regarded as the study and interpretation of data in accordance with probability theory, and general approaches to statistical analysis in a broad sense, depend on the particular probability theory used. Two general approaches have emerged in this regard: (i) the *frequentist approach*, based primarily on the Kolmogorov theory of probability, and (ii) the *Bayesian approach*. There are several variants of these under each category.

To the frequentist, the probability of an event is considered to be one of an infinite number of possible repetitions of the same experiment, each producing independent results. The underlying model parameters remain constant during the repeatable experiments. In this setting, it makes sense to seek the parameters that best describe the statistics of the problem. This idea led to the fundamental frequentist approach, also referred to as the *Fisherian approach* in honour of statistician R. A. Fisher (1922), or as the *maximum likelihood approach*, in which the parameters maximize the likelihood probability for fixed data, to best characterize the true (reality) distribution.

According to Fisher (1922, p. 310), 'The likelihood that any parameter ... should have any assigned value ... is proportional to the probability that if this were so, the totality of observations should be that observed.' So if $\boldsymbol{\theta}$ is assigned a value, the probability that data $\boldsymbol{y}$ are observed is the conditional density (likelihood) $\pi(\boldsymbol{y}|\boldsymbol{\theta}, S)$.

In the case of independent and identically distributed (i.i.d.) samples of observational data $\boldsymbol{y}_i$, the joint distribution is $\prod_{i=1}^{n} \pi(\boldsymbol{y}_i|\boldsymbol{\theta})$ and is viewed as a function of $\boldsymbol{\theta}$ only, which is the *likelihood function*. For ease in computation, one is often concerned with the log-likelihood

$$\log \pi(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \pi(\boldsymbol{y}_i|\boldsymbol{\theta}), \tag{4.9}$$

with the latter holding for independent samples and for each relevant scenario S.

A common theme in classical statistics, particularly frequentist statistics, is that the 'true' distribution is contained in the probability distributions $p_{noise}(\boldsymbol{\epsilon})$ and $p_{model}(\boldsymbol{\theta}, S)$ accessible by the parametric model that defines the prediction $\boldsymbol{d}(\boldsymbol{\theta}, S)$. This is equivalent to assuming that there exists a (magic) parameter $\boldsymbol{\theta}^*$ such that $\boldsymbol{\gamma}(\boldsymbol{\theta}^*, S) = 0,\ 1 \leq i \leq n$; that is, the exact prediction can be derived by the model. Such models are said to be 'correct'

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

or 'well-specified'; otherwise, they are misspecified. It is no surprise that most models are indeed *misspecified*: no parameters exist that produce the truth. As we point out in Oden *et al.* (2018), 'Many fundamental questions arise about such misspecified model such as how close they can approximate the truth and if, while inadequate and misspecified, are they still sufficiently accurate to be used to make important predictions.' The famous statement of George E. P. Box (Box and Draper 1987, Box 1979) alluded to in Section 2.2 under model uncertainty – 'Essentially, all models are wrong, but some are useful' – comes to mind. The mission of model validation processes is to find such useful models given adequate observational data.

As noted earlier, the goal of the Fisherian approach is to find the parameter $\hat{\boldsymbol{\theta}} \in \Theta$ that maximizes the likelihood of a model. The *maximum likelihood estimator*, or MLE, is defined by

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \arg\max_{\boldsymbol{\theta}\in\Theta} [\log \pi(\boldsymbol{y}|\boldsymbol{\theta})], \quad \boldsymbol{y} \in \mathcal{Y}. \tag{4.10}$$

Under specific conditions on the smoothness of the likelihood function, sequences of MLEs obtained with increasing data samples $\boldsymbol{y}_i$ converge (in distribution) to a normal distribution with a covariance matrix equal to the inverse of the Fisher information matrix, $\boldsymbol{I}_1(\boldsymbol{\theta})$,

$$\begin{aligned} \boldsymbol{I}_1(\boldsymbol{\theta}) &= \int_{\mathcal{Y}} \pi(\boldsymbol{y}|\boldsymbol{\theta}) \frac{\partial \log \pi(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log \pi(\boldsymbol{y}|\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} \, \mathrm{d}\boldsymbol{y} \\ &= \operatorname{Var} \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial \log \pi(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] \\ &= -\mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial^2 \log \pi(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right], \end{aligned} \tag{4.11}$$

where $\operatorname{Var}(\cdot)$ is the variance and $\mathbb{E}_{\boldsymbol{\theta}}$ is the expected value with respect to $\pi(\boldsymbol{y}|\boldsymbol{\theta})$. This is conventionally written as

$$\frac{1}{\sqrt{n}}(\hat{\boldsymbol{\theta}}^n - \boldsymbol{\theta}) \to d \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_1^{-1}(\boldsymbol{\theta})), \tag{4.12}$$

where $\hat{\boldsymbol{\theta}}^n$ is the MLE for n i.i.d. observations. We write $\hat{\boldsymbol{\theta}}^n \to \hat{\boldsymbol{\theta}}$, and $\mathcal{N}(\mu, \Sigma)$ denotes the normal distribution with mean μ and inverse of the covariance matrix Σ.

Also of interest is the parameter $\boldsymbol{\theta}^*$ weighted by the truth $\boldsymbol{g}$ as the expected value, defined by

$$\boldsymbol{\theta}^*(\boldsymbol{y}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\boldsymbol{g}}[\log \pi(\boldsymbol{y}|\boldsymbol{\theta})]. \tag{4.13}$$

This is the parameter that for fixed data $\boldsymbol{y}$ maximizes the expected value with respect to the true distribution $\boldsymbol{g}$ of the log-likelihood. It has a special significance when viewed in a relative entropy setting. In particular,

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

the *Kullback–Leibler divergence* between two probability distributions with densities p and q is defined by

$$D_{KL}(p \parallel q) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x}. \tag{4.14}$$

The $D_{KL}(\cdot \parallel \cdot)$ defines the *relative entropy* between the two distributions, as

$$D_{KL}(p \parallel q) = -H(p) + H(p, q), \tag{4.15}$$

where $H(p)$ is the *Shannon information entropy* (modulo a reference density $m(\boldsymbol{x})$ to be discussed in Section 6.3)

$$H(p) = -\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{4.16}$$

and $H(p, q)$ is the cross-entropy

$$H(p, q) = -\int p(\boldsymbol{x}) \log q(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{4.17}$$

Suppose $\boldsymbol{g}$ is the 'true' distribution. Then, $D_{KL}(\boldsymbol{g} \parallel \pi(\cdot, \boldsymbol{\theta}))$ is a measure of the information lost by assuming the likelihood $\pi(\cdot|\boldsymbol{\theta})$ is the truth. Returning to (4.13), we see that

$$\begin{aligned} \boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta} \in \Theta} \ \mathbb{E}_{\boldsymbol{g}}[\log \pi(\boldsymbol{y}, \boldsymbol{\theta})] \\ &= \arg\min_{\boldsymbol{\theta} \in \Theta} \ H(\boldsymbol{g}, \pi(\boldsymbol{y}, \boldsymbol{\theta})) \\ &= \arg\min_{\boldsymbol{\theta} \in \Theta} \ D_{KL}(\boldsymbol{g} \parallel \pi(\boldsymbol{y}, \boldsymbol{\theta})). \end{aligned} \tag{4.18}$$

If there does not exist a parameter $\boldsymbol{\theta}$ such that $\boldsymbol{g}(\boldsymbol{y}) = \pi(\boldsymbol{y}|\boldsymbol{\theta})$, then the model is misspecified and the choice of $\boldsymbol{\theta}$ producing a likelihood closest to the truth is $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, with $\boldsymbol{\theta}^*$ given by (4.18), that is, the parameters defining the closest D_{KL} 'distance' from the model $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ to the truth $\boldsymbol{g}$.

A basic question often considered is the asymptotic behaviour of the MLE parameters $\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \hat{\boldsymbol{\theta}}^n$ given by (4.11) when $\boldsymbol{y} \in \mathbb{R}^n$ and $n \to \infty$. Under special conditions on the likelihood functions, it can be shown (see Kleijn and van der Vaart 2002, 2012) that the $\hat{\boldsymbol{\theta}}^n$ converge to $\hat{\boldsymbol{\theta}}$ in the sense of distributions and $n^{-1/2}(\hat{\boldsymbol{\theta}}^n - \hat{\boldsymbol{\theta}}) \to \mathcal{N}(0, V(\boldsymbol{\theta}^*))$, where V is called the sandwich matrix (Kleijn and van der Vaart 2002, Geyer 2003, Kleijn and van der Vaart 2012) depending on the $L_n(\boldsymbol{\theta})$ and, approximately, $V(\hat{\boldsymbol{\theta}}) = V(\boldsymbol{\theta}^*)$. For the well-behaved well-specified model, $V(\hat{\boldsymbol{\theta}}) = I_1^{-1}(\hat{\boldsymbol{\theta}})$, with $I_1(\boldsymbol{\theta})$ given by (4.11).

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

PART TWO
Principles of Bayesian model calibration and validation

5. The prediction pyramid: Bayesian statistical calibration and validation parametric models

The basic steps in implementing Bayesian inference of model parameters are clear. Beginning with a recognition of prior knowledge of the probability distribution of model parameters (the prior), developing the probability of observational data for given model parameters (the likelihood), we compute (by sampling methods) the parameter distribution to upgrade our knowledge of the key parameters of the system. We then use the best information at hand, the upgraded parameters, to solve the forward problem (*e.g.* (2.1)), and we then use the solution $u(\boldsymbol{\theta}, S)$ to compute the quantities of interest – the prediction desired features of a physical reality. *The central problem of predictive modelling is to determine the validity of a given model as a means for predicting an event, defined by one or more quantities of interest, given certain (possibly noisy) observational data and incomplete prior information on model parameters.* If a model is deemed invalid according to some criterion for acceptability, it must be discarded or modified to better fit validation data. Model selection and adaptivity are taken up in Part 3.

5.1. Rules and principles of model validation

The following rules and principles are put in play to produce a model validation process.

(1) *Quantities of interest.* A fundamental (but often ignored) principle is that it is necessary to specify precisely the intended purpose of the model prediction by specifying clearly the targeted outputs, that is, the *quantities of interest* (QoIs). A mathematical model may be acceptable as a means for predicting one QoI but not another. A model is very rarely useful in predicting all features of the behaviour of a physical system, but may be acceptable for predicting specific features. The target prediction will generally be a probability distribution delivered by the forward solution of a stochastic model, and it will be an event in the often complex scenario S_p, the *prediction scenario.*

The prediction scenario will only rarely pertain to observable events (a fact often misunderstood), since (i) once observed, it is no longer a prediction but is validation data, and (ii) S_p may involve predictions of systems that do not yet exist, as is the case in the design of engineered systems.

(2) *The prediction pyramid.* Symbolically, the hierarchy of scenarios and observational data can be represented in a 'prediction pyramid' in which

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

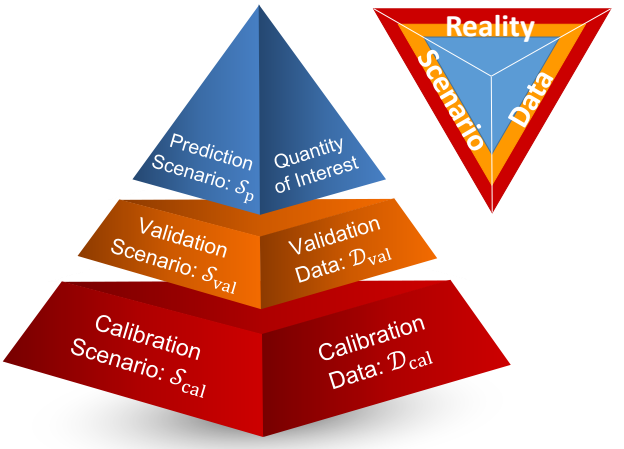


Figure 5.1. The prediction 'pyramid'. A symbolic representation of the hierarchy of scenarios and observational data, beginning with unit calibration scenario S_c delivering data $\boldsymbol{y}_c$, then validation scenario S_v with data $\boldsymbol{y}_v$, and finally the full prediction scenario S_p with the target QoI. The 'pyramid' is actually a tetrahedron with the hidden face representing the never-seen reality we attempt to predict.

the peak represents the full prediction scenario S_p, and the QoI is predicted (Figure 5.1). The 'pyramid' is actually a tetrahedron with a base and three faces, representing modelling scenarios, S_c, S_v, S_p, the corresponding data, $\boldsymbol{y}_c, \boldsymbol{y}_v$, and out-of-data predictions, the QoI, and the unseen and inaccessible face representing reality. The beginning lower levels of experimental observation define calibration scenario S_c and calibration data $\boldsymbol{y}_c$, and then, moving toward the peak, the validation scenario S_v and validation data $\boldsymbol{y}_v$ appear.

The notion of a sequence of observations and model parameter calibrations of increasing complexity, moving toward those with properties expected to indicate the model's ability to predict QoIs, is a richer and more focused version of what is called *cross-validation* in statistics, in which 'training points' are identified at which initial data are retrieved to fit parameters, and then data at points outside the training set are sampled to update parameters.

(3) *Calibration.* The 'tuning' of model parameters so that the model predictions agree with experimental data on simple component tests is the purpose of the calibration process and is performed in scenario S_c. In Oden *et al.* (2018), we recommend subscribing to the philosophy of E. T. Jaynes (2003, p. 86), who said 'The essence of "honesty" or "objectivity" demands that we take into account all the evidence we have, not just some arbitrarily chosen

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

subset of it.' The implementation of this philosophy, which we call 'Jaynes' principle' is a challenge in designing appropriate calibration scenarios.

(4) *Validation.* Model validation is defined as the process of estimating the accuracy with which a model can predict QoIs based on the accuracy with which it predicts actual experimental observations. These observations produce validation data $\boldsymbol{y}_v$ accessed in validation experiments performed in validation scenarios S_v. The major challenge in model validation is to design validation experiments that depict the influence of the QoIs on subsystems or surrogates of the full prediction scenario S_p. The word 'accuracy' here is of fundamental importance: one must choose a validation metric to measure the error in model predictions and experimental observations in S_v, and one must specify a tolerance that this measure of modelling error must meet if the model prediction is deemed sufficiently close to the experimental data. This is clearly a subjective process and depends on the choices of the models and on choices of the modeller. If the error tolerances are met, it is customary to declare that the model is 'valid', but technically one can only say that with this metric and with this tolerance, the model is 'not invalid', since additional information could always falsify a model. It is clear that the design of the validation experiments that adequately inform the model of the QoIs, the selection of a validation metric, and the selection of error tolerances are fundamentally important aspects of any validation process.

In summary, we follow these principles and rules.

(1) Identify QoIs, Q, in the full model scenario S_p and construct a parametric model $\mathcal{M} \sim A(\boldsymbol{\theta}, S_p; u(\boldsymbol{\theta}, S_p)) = 0$ of a forward problem on S_p with the goal of computing measure of distributions or densities $\pi(Q)$ and evaluating in some sense their uncertainty content (see Section 7). Of course, the density $\pi(Q)$ representing our approximation of the QoI must technically be conditional on many things: the model $\mathcal{M}$, the data, the parameter $\boldsymbol{\theta}$ used in the prediction scenario, *etc.*:

$$QoI \sim \pi(Q|\mathcal{M}, \boldsymbol{\theta}, S_p, S_v, \boldsymbol{y}_c, \ldots). \tag{5.1}$$

(2) Construct calibration scenarios S_c to calculate initial probability density functions for model parameters. This involves choosing a prior $\pi(\boldsymbol{\theta})$ for calibration scenario S_c.

(3) Collect calibration data $\boldsymbol{y}_c$ and update (or 'initialize') the parameters through Bayesian inversion:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}_c, S_c) = \frac{\pi(\boldsymbol{y}_c|\boldsymbol{\theta}, S_c)\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{y}_c|S_c)}. \tag{5.2}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

(4) Construct validation scenarios S_v, collect validation data $\boldsymbol{y}_v$, and use the calibration posterior as a prior for the validation update:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}_v, \boldsymbol{y}_c, S_v, S_c) = \frac{\pi(\boldsymbol{y}_v|\boldsymbol{\theta}, \boldsymbol{y}_c, S_c, S_v)\pi(\boldsymbol{\theta}|\boldsymbol{y}_c, S_c)}{\pi(\boldsymbol{y}_v|\boldsymbol{y}_c, S_v)}. \tag{5.3}$$

(5) Use the parameters determined in the validation update (5.3) to solve the validation forward problem:

$$A(\boldsymbol{\theta}, S_v; u(\boldsymbol{\theta}, S_v)) = 0. \tag{5.4}$$

(6) Compute the prediction $\boldsymbol{d}(\boldsymbol{\theta}, S_v)$ of $\boldsymbol{y}_v$ delivered by the solution $u(\boldsymbol{\theta}, S_v)$ and determine the accuracy with which the model predicts the validation observable using an appropriate metric $d(\cdot, \cdot)$. Given a tolerance γ_{tol}, if

$$d(\boldsymbol{d}(\boldsymbol{\theta}, S_v), \boldsymbol{y}_v) \leq \gamma_{tol}, \tag{5.5}$$

the model is deemed acceptable (not invalid).

(7) Using the parameter $\boldsymbol{\theta}$ in the validation model, solve the forward problem $A(\boldsymbol{\theta}, S_p; u(\boldsymbol{\theta}, S_p)) = 0$ on S_p and use the solution $u(\boldsymbol{\theta}, S_p)$ to compute the QoI and the probability density $\pi(Q)$ of (5.1). We may also *quantify the uncertainty* in the prediction $\pi(Q)$ by computing the mean and variances or higher moments of $\pi(Q)$ or measures of uncertainty content such as its information entropy,

$$H(\pi(Q)) = -\int \pi(Q(\boldsymbol{\theta})) \log \pi(Q(\boldsymbol{\theta}))\, \mathrm{d}\boldsymbol{\theta}.$$

5.2. Example: optimization and the MAP

A standard example that permits a comparison of certain aspects of the frequentist and Bayesian approaches involves the case in which the likelihood is given by

$$\pi(\boldsymbol{y}_s|\boldsymbol{\theta}, S) \propto \exp\left\{-\frac{1}{2}\|\boldsymbol{y}_s - \boldsymbol{d}(\boldsymbol{\theta}, S)\|^2_{\boldsymbol{\Sigma}}\right\}, \tag{5.6}$$

where $\|\cdot\|^2_{\boldsymbol{\Sigma}} = (\cdot)^T\boldsymbol{\Sigma}^{-1}(\cdot)$, $\boldsymbol{\Sigma} = [\Sigma_{ij}]$ being the $n \times n$ covariance matrix. In some cases, $\boldsymbol{\Sigma}$ is simply $\sigma^2\boldsymbol{I}$ ($\Sigma_{ij} = \sigma^2\delta_{ij}$), σ^2 being the variance, which can be added to $\boldsymbol{\theta}$ as an additional parameter. If $\boldsymbol{d}(\boldsymbol{\theta}, S)$ is linear in $\boldsymbol{\theta}$, the likelihood in (5.6) is Gaussian; otherwise, it is non-Gaussian.

Let $\boldsymbol{d}(\boldsymbol{\theta}, S)$ be linear: $\boldsymbol{d}(\boldsymbol{\theta}, S) = \boldsymbol{G}\boldsymbol{\theta}$, $\boldsymbol{G}$ being a $p \times p$ matrix. Then the MLE is

$$\boldsymbol{\theta}^* = (\boldsymbol{G}^T\sigma^{-1}\boldsymbol{G})^{-1}\boldsymbol{G}^T\sigma^{-1}\boldsymbol{y}. \tag{5.7}$$

If the prior $\pi(\boldsymbol{\theta})$ is also Gaussian, the posterior becomes

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}, S) \propto \exp\left\{-\frac{1}{2}\mathcal{S}(\boldsymbol{\theta})\right\}, \tag{5.8}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

where $\mathcal{S}(\boldsymbol{\theta})$ is the *misfit function*,

$$\mathcal{S}(\boldsymbol{\theta}) = \|\boldsymbol{y}_i - \boldsymbol{d}_i(\boldsymbol{\theta}, S)\|^2_{\boldsymbol{\Sigma}} + \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2_{\boldsymbol{\Sigma}_{prior}}, \tag{5.9}$$

$\boldsymbol{\Sigma}_{prior}$ being the covariance matrix of the prior density, and

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2_{\boldsymbol{\Sigma}_{prior}} = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}^{-1}_{prior}(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

$\boldsymbol{\theta}_0$ being a reference parameter vector. The prior is seen to provide a Tikhonov-type regularization of the misfit $\mathcal{S}(\boldsymbol{\theta})$. The vector

$$\boldsymbol{\theta}^{**} = \arg\max_{\Theta}\ \pi(\boldsymbol{\theta}|\boldsymbol{y}, S) \tag{5.10}$$

is called the maximum *a posteriori estimate* (MAP) for scenario S. We note that for uniform priors, MAP = MLE. We finally have

$$\begin{aligned}\pi(\boldsymbol{\theta}^{**}|\boldsymbol{y}, S) &= \pi_{posterior}(\boldsymbol{\theta}^{**}|\boldsymbol{y}, S)\\ &= \mathcal{N}(\boldsymbol{\theta}^{**}|\boldsymbol{\Gamma}^{-1}_{post}),\end{aligned} \tag{5.11}$$

where

$$\boldsymbol{\Gamma}_{post} = (\boldsymbol{G}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{G} + \boldsymbol{\Sigma}^{-1}_{prior})^{-1}. \tag{5.12}$$

5.3. Gaussian processes: tools to build models of noise and model inadequacies

At some steps in the processes of formulating the Bayesian inference problem we encountered the need to contemplate new parametric models, outside of models relevant to the principal forward problem, to approximate such entities as experimental noise or model inadequacy. The idea of Gaussian processes provides a popular and often powerful approach for constructing such models *if sufficient data are available*. In this section we present a brief introduction to this subject. A detailed account can be found in the popular monograph of Rasmussen and Williams (2006).

5.3.1. Stochastic or random process

A *stochastic* or *random process* is a collection of random variables $X = \{X(t); t \in I\}$ defined on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, I being an index set and the values of X lying in some 'state space' Y. In general, one selects a finite number n of *trial or training points* $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \in I^n$ and the process generates a probability distribution

$$\{X(\boldsymbol{x}_1), X(\boldsymbol{x}_2), \ldots, X(\boldsymbol{x}_n)\},$$

for which we compute probabilities $\mathbb{P}[X(\boldsymbol{x}_1), X(\boldsymbol{x}_2), \ldots, X(\boldsymbol{x}_n)]$. Thus, a stochastic process can be characterized by a function $f(X)$ of a random variable X and, for any set of sample or trial points $\{\boldsymbol{x}_i\}_{i=1}^n$, a corresponding set of probabilities

$$\{\mathbb{P}(X(\boldsymbol{x}_1)), \mathbb{P}(X(\boldsymbol{x}_2)), \ldots, \mathbb{P}(X(\boldsymbol{x}_n))\}.$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

5.3.2. Gaussian processes

A *Gaussian process* (GP) is a random process in which any finite number of random variables have a joint probability distribution that is Gaussian.

So here we think of a real Gaussian process as being defined by random variables represented by the value of a function $\boldsymbol{f}(\boldsymbol{x})$ at the location of a point $\boldsymbol{x}$ (*e.g.* $\boldsymbol{x} \in \mathbb{R}^m$, or $\boldsymbol{x} = t$, a particular time value). Since every Gaussian distribution is completely specified by its mean and covariance, a Gaussian process $\boldsymbol{f}(\boldsymbol{x})$ is defined by its *mean function* $\boldsymbol{m}(\boldsymbol{x})$ and its *covariance function* $\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}')$:

$$\begin{aligned} \boldsymbol{m}(x) &= \mathbb{E}[\boldsymbol{f}(\boldsymbol{x})], \\ \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}') &= \mathbb{E}[(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{m}(\boldsymbol{x}))(\boldsymbol{f}(\boldsymbol{x}') - \boldsymbol{m}(\boldsymbol{x}'))^T]. \end{aligned} \tag{5.13}$$

It is customary to use the following notation for such a process (Rasmussen and Williams 2006):

$$\boldsymbol{f}(\boldsymbol{x}) \sim GP(\boldsymbol{m}(\boldsymbol{x}), \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}')). \tag{5.14}$$

The great utility of Gaussian processes resides in their so-called *marginalization property*: since all distributions at any sample point are, by definition, Gaussian, any enlargement of sample points taking the process outside the initial sample must also be Gaussian, and any such larger sample will not change the distribution of the smaller sample set. This subtle property, also called the *consistency property*, becomes clear in the basic Gaussian extension described below.

5.3.3. The consistency or marginalization property of GPs

The fundamental consistency or marginalization property of GPs is embodied in the following theorem.

Theorem 5.1 (Rasmussen and Williams 2006). Let $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ be two random variables with joint Gaussian probability distribution

$$(\boldsymbol{y}_1, \boldsymbol{y}_2) \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}, \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^T & \boldsymbol{C} \end{bmatrix} \right). \tag{5.15}$$

Then

$$\boldsymbol{y}_1 \sim \mathcal{N}(\boldsymbol{a}, \boldsymbol{A}) \tag{5.16}$$

and

$$(\boldsymbol{y}_1 | \boldsymbol{y}_2) \sim \mathcal{N}(\boldsymbol{a} - \boldsymbol{B}\boldsymbol{C}^{-1}(\boldsymbol{y}_2 - \boldsymbol{b}), \boldsymbol{A} - \boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^T). \tag{5.17}$$

Proof. It is understood that $\boldsymbol{y}_1$ is associated with a density $\pi(\boldsymbol{y}_1)$ equal to the marginalization $\int \pi(\boldsymbol{y}_1, \boldsymbol{y}_2)\, \mathrm{d}\boldsymbol{y}_2$. The conditional probability (5.17) follows the evaluation of

$$\pi(\boldsymbol{y}_1 | \boldsymbol{y}_2) = \frac{\pi(\boldsymbol{y}_1, \boldsymbol{y}_2)}{\pi(\boldsymbol{y}_2)} \tag{5.18}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

being Gaussian, making use of the Gaussian distribution equation: if $\boldsymbol{f} = \{f_1, f_2, \ldots, f_n\}^T$, and $\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$, then

$$\pi(\boldsymbol{f}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{f}-\boldsymbol{m})\right\}. \tag{5.19}$$

□

The joint density is

$$\pi(\boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}\Gamma(y_1, y_2)\right),$$

where

$$\begin{aligned}\Gamma(y_1, y_2) &= \{(y_1-a)^T(y_2-b)^T\}\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{B}^T & \boldsymbol{C}\end{bmatrix}^{-1}\begin{bmatrix}y_1-a\\ y_2-b\end{bmatrix}\\ &= (y_1-a)^T\boldsymbol{A}'(y_1-a) + 2(y_1-a)^T\boldsymbol{B}'(y_2-b)\\ &\quad + (y_2-b)^T\boldsymbol{C}'(y_2-b)\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{A}' &= (\boldsymbol{A}-\boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^T)^{-1},\\ \boldsymbol{B}' &= -\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}-\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} = \overset{\sim}{\boldsymbol{B}}^T,\\ \boldsymbol{C}' &= (\boldsymbol{C}-\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}.\end{aligned}$$

Thus, with some algebra, one has

$$\Gamma(y_1, y_2) = \mathcal{N}(b, \boldsymbol{C})\mathcal{N}(\mu, \boldsymbol{\Sigma}'),$$

where

$$\begin{aligned}\mu &= a - \boldsymbol{B}\boldsymbol{C}^{-1}(y_2-b),\\ \boldsymbol{\Sigma}' &= \boldsymbol{A}-\boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^T.\end{aligned}$$

Then

$$\begin{aligned}\pi(y_1|y_2) &= \pi(y_1, y_2)/\pi(y_2)\\ &= \mathcal{N}(b, \boldsymbol{C})\mathcal{N}(\mu, \boldsymbol{\Sigma}')/\mathcal{N}(b, \boldsymbol{C})\\ &= \mathcal{N}(\mu, \boldsymbol{\Sigma}')\\ &= \mathcal{N}(a-\boldsymbol{B}\boldsymbol{C}^{-1}(y_2-b), \boldsymbol{A}-\boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^T),\end{aligned}$$

which was asserted. □

5.3.4. The Gaussian process response surface

We use the GP framework to construct *response surfaces*. The idea is as follows. Let $\boldsymbol{Y}(\boldsymbol{x})$ denote a random field for which realizations

$$\{\boldsymbol{y}(x_1), \boldsymbol{y}(x_2), \ldots, \boldsymbol{y}(x_m)\}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

are given, and let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ be *training points* ($\boldsymbol{x}_i \in \mathcal{R}^n$) (or 'input points' or 'control points'). Then the *training set* is

$$\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}(\boldsymbol{x}_i), i = 1, 2, \ldots, m\}. \tag{5.20}$$

We represent $\boldsymbol{y}$ by a GP

$$\boldsymbol{y} \sim GP(\boldsymbol{\psi}^T(\boldsymbol{x})\boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{5.21}$$

where

$$\boldsymbol{\psi}^T(\boldsymbol{x})\boldsymbol{\beta} = \sum_{i=1}^{q} \boldsymbol{\psi}_i(\boldsymbol{x})\beta_i = \boldsymbol{m}(\boldsymbol{x}) = \text{the mean function} \tag{5.22}$$

(also called the 'trend'), $\psi_i(\boldsymbol{x})$ are appropriate basis functions, $\boldsymbol{\beta}$ is a vector of coefficients, and $\boldsymbol{\Sigma}$ is the *covariance matrix*, which is any symmetric, positive semidefinite matrix. Commonly, we take

$$\begin{aligned} \Sigma_{ij} &= k(x_i, x_j) \\ &= \lambda_0 \exp\left\{-\frac{1}{2}\sum_{k=1}^{d} \frac{(x_i^k - x_j^k)^2}{l_k}\right\} + \lambda_1 + \lambda_2\delta_{ij}. \end{aligned} \tag{5.23}$$

Here λ_0, l_k, λ_1 and λ_2 are parameters, with

- λ_0 = a measure of precision of variance (an amplitude of a covariance envelope),
- l_k = characterization of flatness of covariance,
- λ_1 = a bias 'latent value', and
- λ_2 = a regularization.

By adjusting values of the parameters, we can conceivably manipulate the shape and flatness of the covariance between sample points in $\mathcal{D}$.

Given the GP (5.21), the *predicted output* $\boldsymbol{y}^* = \boldsymbol{y}(\boldsymbol{x}^*)$ at a point outside the training set $\mathcal{D}$, $(\boldsymbol{x}^*, \boldsymbol{y}^*) \notin \mathcal{D}$, is, according to (5.17),

$$(\boldsymbol{y}^*|\boldsymbol{y}) = GP(\mathbb{E}((\boldsymbol{y}^*|\boldsymbol{y})), \boldsymbol{\Sigma}^*), \tag{5.24}$$

where

$$\begin{aligned} \mathbb{E}(\boldsymbol{y}^*|\boldsymbol{y}) &= \boldsymbol{\psi}^T(\boldsymbol{x})\boldsymbol{\beta} - \boldsymbol{k}(\boldsymbol{x}^*, \boldsymbol{x})\boldsymbol{K}^{-1}(\boldsymbol{y} - \boldsymbol{m}), \\ \boldsymbol{\Sigma}^* &= \boldsymbol{k}(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{k}(\boldsymbol{x}^*, \boldsymbol{x})\boldsymbol{K}^{-1}\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}^*), \end{aligned} \tag{5.25}$$

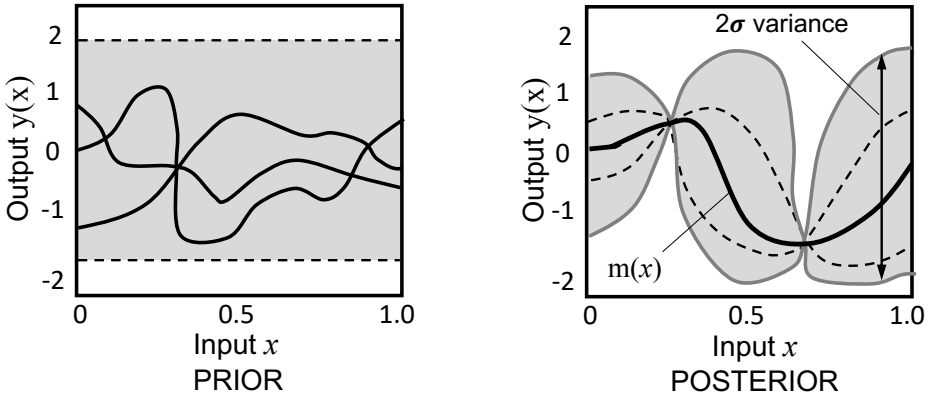
and

$$\boldsymbol{K} = \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}). \tag{5.26}$$

Clearly, $\boldsymbol{m}(\boldsymbol{x}) = \boldsymbol{\psi}^T(\boldsymbol{x})\boldsymbol{\beta}$. The new parameters ω defining the GP are called hyperparameters, for example,

$$\boldsymbol{\omega} = (\boldsymbol{\beta}, \lambda_0, \{l_k\}_{k=1}^d, \lambda_1, \lambda_2). \tag{5.27}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press



(a) (b)

Figure 5.2. Illustration of a dataset fit by a GP with noise-free data point: (a) three functions drawn at random from a GP prior, and (b) three functions drawn from the posterior, the shaded area representing twice the standard deviation. This illustration is a 'pictorial introduction to Bayesian modelling' inspired by a similar figure in Rasmussen and Williams (2006, p. 3).

Figure 5.2 shows a conceptualization of a GP representation over a one-dimensional input set. On the left, we see a collection of prior probability density functions. The use of five input or training points *anchor* the output function $\boldsymbol{y}(\boldsymbol{x})$ at points $\boldsymbol{y}(\boldsymbol{x}_i)$ on the mean curve, $\boldsymbol{m}(\boldsymbol{x})$, as shown. The shape of the covariance between these points is fashioned by adjusting the covariance parameters. The factorization rule can be used to update function values

$$\pi(\boldsymbol{y}(\boldsymbol{x}_1), \boldsymbol{y}(\boldsymbol{x}_2), \ldots, \boldsymbol{y}(\boldsymbol{x}_m)) = \prod_i^m \pi(\boldsymbol{y}(\boldsymbol{x}_1))\pi(\boldsymbol{y}(\boldsymbol{x}_2)), \ldots, \pi(\boldsymbol{y}(\boldsymbol{x}_m)).$$

5.3.5. Model inadequacy and validation using Gaussian processes

We close this section with an example of the use of GPs to represent model inadequacy and also to take account of experimental noise, following Bayarri *et al.* (2007).

Let $\boldsymbol{g}(\boldsymbol{x})$ denote the truth or reality of a physical entity at point $\boldsymbol{x} \in \mathbb{R}^n$ (denoted $\boldsymbol{g}$ earlier), and let $\boldsymbol{d}(\boldsymbol{x}, \boldsymbol{\theta})$ denote the model prediction of this field at $\boldsymbol{x}$ (denoted $d(\boldsymbol{\theta})$ earlier). Then the model inadequacy $\boldsymbol{\gamma}$ $(= \gamma_i(\boldsymbol{\theta}, S)$ is

$$\boldsymbol{\gamma}(\boldsymbol{\theta}, \boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{d}(\boldsymbol{x}, \boldsymbol{\theta}). \tag{5.28}$$

Here $\boldsymbol{\theta}$ is the parameter vector of the forward prediction problem. Let $\boldsymbol{g}(\boldsymbol{x}_i)$ denote the 'field data' at $\boldsymbol{x}_i$ in a set of test points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ (denoted $\boldsymbol{y}_i$ earlier). Thus, for a additive model,

$$\boldsymbol{g}(\boldsymbol{x}_i) = \boldsymbol{g}(\boldsymbol{x}_i) + \epsilon_i, \tag{5.29}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

where ϵ_i is experimental noise. To complete the noise model, we assume ϵ_i is i.i.d. and

$$\epsilon_i \sim \mathcal{N}(0, 1/\lambda). \tag{5.30}$$

Thus

$$\boldsymbol{g}(\boldsymbol{x}_i) - \boldsymbol{d}(\boldsymbol{x}_i, \boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}, \boldsymbol{x}_i) + \epsilon_i. \tag{5.31}$$

We next introduce priors for the key probability density functions:

$$\pi(\boldsymbol{\theta}), \quad \pi(\lambda) \quad \text{and} \quad \pi(\gamma) \sim GP(\boldsymbol{\psi}^T(\boldsymbol{x})\boldsymbol{\beta}, \boldsymbol{\Sigma}_\gamma), \tag{5.32}$$

with hyperparameters $\boldsymbol{\omega}_j = (\beta_\gamma, \lambda_0^\gamma, \{l_k^\gamma\}_{k=1}^{d_\gamma}, \lambda_1^\gamma, \lambda_2^\gamma)$. The posterior update is

$$\pi(\boldsymbol{\theta}, \lambda, \gamma | \boldsymbol{y}) \propto \boldsymbol{f}(\boldsymbol{y}^F | \boldsymbol{\theta}, \lambda, \gamma)\pi(\theta, \lambda, \gamma), \tag{5.33}$$

with $\boldsymbol{f}$ being a multivariate normal distribution. The model output is

$$\boldsymbol{d} \sim GP(\boldsymbol{\psi}^T(\boldsymbol{x})\boldsymbol{\beta}, \boldsymbol{\Sigma}_\gamma). \tag{5.34}$$

5.3.6. Computing the hyperparameters

One popular method to estimate the hyperparameters $\boldsymbol{\omega}$ is to use the MLE

$$\boldsymbol{\omega}_{MLE} = \arg\max_{\boldsymbol{\omega}} \pi(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\omega}), \tag{5.35}$$

where

$$\pi(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\omega}) \propto \prod_k \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{d}^m(\boldsymbol{\omega}, \boldsymbol{x}))^2 / \sigma^2 \right\}, \tag{5.36}$$

with $\boldsymbol{y} = \boldsymbol{d}^m + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. A common approach is to use the marginal likelihood

$$\frac{\partial}{\partial \omega_j} \log \pi(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\omega}) = 0 = \frac{1}{2}\boldsymbol{y}^T \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{\omega}_j} \boldsymbol{\Sigma}^{-1} \boldsymbol{y} - \frac{1}{2} \text{ trace } \left(\Sigma^{-1} \frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\omega}_j} \right). \tag{5.37}$$

The full Bayesian uses the posterior

$$\pi(\boldsymbol{\theta}, \boldsymbol{\omega} | \boldsymbol{y}) \propto \pi(\boldsymbol{y} | \boldsymbol{\theta}, \boldsymbol{\omega})\pi(\boldsymbol{\theta}, \boldsymbol{\omega}). \tag{5.38}$$

In implementing this GP update, Bayarri *et al.* (2007) describe a Markov chain Monte Carlo process starting with the posterior

$$\begin{aligned} \pi(\boldsymbol{\theta}, \lambda, \boldsymbol{\gamma}(\boldsymbol{\omega}_\gamma) | \boldsymbol{y}) &\propto f(\boldsymbol{y} | \boldsymbol{\theta}, \lambda, \boldsymbol{\gamma}(\boldsymbol{\omega}))\pi(\boldsymbol{\theta}, \lambda, \boldsymbol{\gamma}) \\ &\approx \boldsymbol{y} - \boldsymbol{d}(\boldsymbol{x}_i, \boldsymbol{\theta}) - \epsilon_i, \end{aligned} \tag{5.39}$$

where the outputs $\{\boldsymbol{\theta}^{(i)}, \lambda, \boldsymbol{d}(\boldsymbol{x}, \boldsymbol{\theta}^{(i)}), \boldsymbol{\gamma}^{(i)}(\boldsymbol{x})\}$ are sampled and all quantities of interest can be estimated from these samples.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

The prediction of reality $\boldsymbol{y}(\boldsymbol{x})$ at a new set of training points $\mathcal{D}_{new}$ is drawn from the predictive posterior distribution $\pi(\boldsymbol{g}_{new}|\boldsymbol{y}, \boldsymbol{d})$, where $\boldsymbol{g} = \boldsymbol{d} + \boldsymbol{\gamma}$, determined from

$$\boldsymbol{g}_{new} \sim \frac{1}{N} \sum_{k=1}^{N} (\boldsymbol{y}_{new}^{M(k)} + \boldsymbol{\gamma}_{new}^{(k)}), \tag{5.40}$$

where $\boldsymbol{d}_{new}^{(k)}$ and $\boldsymbol{\gamma}_{new}^{(k)}$ are drawn from the joint posterior. Then, $\boldsymbol{g}_{new}$ is the 'new' prediction of the reality. Variance and tolerance bounds (credible intervals) can be computed approximately from these calculations.

6. Information theory and the principle of maximum entropy

6.1. Introduction

Apart from Bayesian approaches, but very much complementary to them, are the methods and concepts underlying information theory and the concept of information entropy. Introduced in the landmark paper of Claude Shannon (1948), information theory provides a rational approach to quantifying and understanding the information content in a probability distribution (or a digital signal). When enriched by the observations and works of E. T. Jaynes (2003), well-defined measures of uncertainty in information or in probability emerge that are fundamentally important in many applications, including the rational construction of priors and in Bayesian approaches to inverse problems, such as image reconstruction and pattern recognition. We take a brief journey into the subject in the present section. More detailed accounts can be found in the original paper of Shannon (1948) and the books of Jaynes (2003), Gregory (2005), Klir (2005) and Cover and Thomas (2006), and the notes of Carter (2007).

6.2. An elementary look at information

Let us first address the problem of defining a meaningful measure of the information content in a probability density p. We begin by considering a random experiment with n sampled outputs, with probabilities $p_1, p_2, \ldots, p_n$, respectively, and with each $p_i > 0$. We wish to construct a measure of the information obtained by observing a particular p_i. Such an information measure should also be related to the uncertainty in p_i which we also want to quantify. We propose that the information I has the following commonsense properties.

- $I : (0, 1] \to [0, \infty)$ $(I(p) \geq 0)$.
- $I(p_i = 1) = 0$ (there is no information in a certain probability).

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

- $I(p)$ should behave properly for joint, independent outcomes. In particular, if the outcomes of n experiments are denoted by p_i ($\in \mathbb{N}^n$) and the outcomes of another m independent experiments by q_i ($\in \mathbb{N}^m$), then the joint outcome should be $p_i \cdot q_j$ for all $i \in \mathbb{N}^n$ and $j \in \mathbb{N}^m$. The 'anticipatory' information (Klir 2005, Carter 2007) of the joint outcome will therefore be the sum of the individual outcomes:

$$I(p_i \cdot q_j) = I(p_i) + I(q_j). \tag{6.1}$$

Equation (6.1) is referred to by some authors (*e.g.* Klir 2005) as a *Cauchy equation*, the solution of which is a class of functions I defined for all $\alpha \in (0, 1]$ by

$$I(a) = c \log_b a,$$

where c is an arbitrary constant and b is a distinct constant $\neq 1$. Since $I(\cdot)$ must be decreasing on $(0, 1]$ and $\log(\cdot)$ is increasing, c must be negative. By scaling so that $I(1/2) = 1$ and choosing $c = -1$, we get

$$I(a) = -\log_b a, \tag{6.2}$$

or, for probability p_i of outcome i,

$$I(p_i) = -\log_b p_i = \log_b \left(\frac{1}{p_i} \right). \tag{6.3}$$

We remark that the property (6.2), and hence (6.3), follows from (6.1), or more generally from properties of continuous real-valued functions $g : (0, \infty) \to \mathbb{R}$ with the property that $g(x_1 x_2) = g(x_1) + g(x_2)$. Indeed, from this one shows that $g(x^n) = ng(x)$ and, by continuity, $g(x^\alpha) = \alpha g(x)$, $\alpha > 0$. Setting $x = a$ reveals that $g(a^\alpha) = \alpha g(a)$ or $g(a)^{-1} g(a^\alpha) = \alpha$. This shows that $g(a)^{-1} g(x)$ is a left inverse of a^x. A straightforward argument shows that $g(0)$ is also injective and, therefore, that $g(a)^{-1} g(x) =$ the inverse of $a^x = \log_a x$, so that

$$g(x) = g(a) \log_a x.$$

Setting $I(\cdot) = g(\cdot)$, a, $g(a) = 1$ and $a = b$ yields (6.2).

Example. If we take $b = 2$, the units of information are called *bits*. Carter (2007) notes the example of flipping a fair coin to give heads H or tails T with probability $p = 1/2$. The information is $I(p) = -\log_2(1/2) = 1$ bit. For n flips $I(p^n) = -\log((1/2)^n) = n \log(2) = n$ bits.

6.3. A first look at information entropy

These properties will prove useful in determining the connection between information content in a probability density function and its 'entropy'. For a discrete probability distribution $p = \{p_1, p_2, \ldots, p_n\}$, we will define the

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

information entropy $H(p)$ as the real number,

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i, \tag{6.4}$$

where the logarithm can be selected for any base b. We generally take $b = 2$. Exactly why (6.4) is a meaningful definition of information entropy will be taken up in Section 6.5, where we present Shannon's theorem and arguments as to why this property is a fundamental measure of uncertainty content in a probability distribution.

Consider a situation in which we are given n quantities $\{a_1, a_2, \ldots, a_n\}$ with corresponding probabilities $\{p_1, p_2, \ldots, p_n\}$, each a_i sampled independently of the others. It is natural to ask: What is the average amount of information we get from each a_i?

From entity a_i, we obtain the information

$$I(p_i) = -\log(p_i) = \log(1/p_i).$$

In a run of N observations, we see around Np_i occurrences of a_i. Thus, in the N independent observations, we find the total information

$$I = \sum_{i=1}^{n} Np_i \log\left(\frac{1}{p_i}\right).$$

Then, the expectation (mean) value of the information is I/N, or

$$\mathbb{E}[I(p)] = \langle I(p) \rangle = -\sum_{i=1}^{n} p_i \log(p_i) = H(p).$$

We have shown that the *entropy of a probability distribution is the expectation value of the information of the distribution:*

$$\mathbb{E}[I(p)] = \langle I(p) \rangle = H(p). \tag{6.5}$$

6.4. The Maximum Entropy Principle

A simple inequality attributed to J. Willard Gibbs provides an upper bound on $\ln(x)$.

Gibbs' inequality. We have

$$\ln(x) \leq x - 1, \quad \text{for } x > 0, \tag{6.6}$$

with equality if and only if $x = 1$.

Proof. Since $(\ln x)' = 1/x$, the log function is tangent to $y(x) = x - 1$ at $x = 1$. Since $\ln x$ is concave for $x > 0$, it lies below each of its tangents, and, in particular, it lies below the tangent to $x = 1$. Hence

$$\ln x \leq x - 1,$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

with equality if and only if $x = 1$. □

A classical application of Gibbs' inequality arises when we consider the relative entropy of two probability distributions, say $p = \{p_1, p_2, \ldots, p_n\}$ and $p = \{q_1, q_2, \ldots, q_n\}$, where $p_i, q_i \geq 0$, and $\sum_i p_i = \sum_i q_i = 1$. Their Kullback–Leibler divergence (or *relative entropy*) is

$$D_{KL}(p||q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}.$$

Thus

$$\begin{aligned} -D_{KL}(p||q) &= -\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \\ &\leq \sum_{i=1}^{n} p_i \left(1 - \frac{q_i}{p_i}\right) \\ &= \sum_{i=1}^{n} (q_i - p_i) \\ &= 0, \end{aligned} \tag{6.7}$$

from Gibbs' inequality. Therefore,

$$D_{KL}(p||q) = -H(p) + H(p,q) \geq 0. \tag{6.8}$$

where $H(p,q)$ is the cross-entropy

$$H(p,q) = -\sum_{i=1}^{n} p_i \log q_i.$$

Next, given a probability distribution $\{p_1, p_2, \ldots, p_n\}$, we compute the difference

$$\begin{aligned} H(p) - \log(n) &= -\sum_{i=1}^{n} p_i \log p_i - \log(n) \\ &= \sum_{i=1}^{n} p_i \log\left(\frac{1}{p_i}\right) - \log(n) \sum_{i=1}^{n} p_i \\ &= \sum_{i=1}^{n} p_i \left(\log\left(\frac{1}{p_i}\right) + \log\left(\frac{1}{n}\right)\right) \\ &= \sum_{i=1}^{n} p_i \log\left(\frac{1/n}{p_i}\right). \end{aligned} \tag{6.9}$$

But, from Gibbs' inequality (6.6), this latter sum is less than or equal to

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

zero, the equality holding only when $p_i = 1/n$. Thus,

$$0 \leq H(p) \leq \log(n). \tag{6.10}$$

The entropy $H(p)$ is zero only when one of the p_i is one and all others are zero (then the average information $\langle I(p) \rangle = 0$) and $H(p) = \log(n)$; it is maximum only when all of the probabilities have the same value, $1/n$, that is, all events are equally likely.

We are thus led to a fundamental property of the entropy of a probability.

Maximum Entropy Principle. Let $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ be a set of points in $\mathbb{R}^m$ and let $p_i = p(\boldsymbol{x}_i)$ denote the probability of an event at $\boldsymbol{x}_i$. Let

$$\mathcal{P}_n = \left\{ (p_1, p_2, \ldots, p_n) | p_i \in [0,1], \ \sum_i p_i = 1 \right\}$$

be the set of all probability distributions on X with n components. Then of all $p \in \mathcal{P}_n$, the one that maximizes the entropy defines the information content $\mathbb{E}[I(p)]$ and quantifies the uncertainty in the information. A more general treatment is given in the next subsection.

6.5. Shannon's theorem

It is often asserted that the information entropy $H(p)$ is the only meaningful measure of uncertainty and information in probability theory. More rigorous developments that support this assertion can be found in the original paper of Shannon (1948), in the treatise of Jaynes (2003), the descriptive exposition of Klir (2005), and other works. Without providing full details, we summarize the basic axioms underlying the theorem, following Klir (2005).

Our goal is to develop a more rigorous justification of (6.4) as an example of uncertainty content in a probability distribution and to provide a more general basis for the Maximum Entropy Principle. Given again $\mathcal{P}_n$, we seek a function $H(p_1, \ldots, p_n)$ satisfying the following conditions.

Axiom 1. When a zero component is added to a probability distribution, the quantity of uncertainty should not change: $H(p_1, p_2, \ldots, p_n) = H(p_1, p_2, \ldots, p_n, 0)$.

Axiom 2. $H(\cdot)$ is a continuous function of all of its arguments, $H(p) = H(p_1, p_2, \ldots, p_n)$.

Axiom 3. Uncertainty is invariant under permutation π of the probability components: $H(p_1, p_2, \ldots, p_n) = H(\pi(p_1, p_2, \ldots, p_n))$, $\pi(p)$ being a permutation of p.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

Axiom 4. The maximum uncertainty is obtained when all probability components are equal to $1/n$:

$$H(p_1, p_2, \ldots, p_n) \leq H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

Axiom 5. The uncertainty of any joint probability must not exceed the sum of the uncertainties of the corresponding marginal distributions (Klir 2005, p. 73): that is, given a joint probability distribution p_{ij}, $1 \leq i \leq n$, $1 \leq j \leq m$,

$$H(p_{11}, p_{12}, \ldots, p_{1m}, p_{21}, p_{22}, \ldots, p_{2m}, \ldots, p_{n1}, p_{n2}, \ldots, p_{nm}) \leq H\left(\sum_{j=1}^{m} p_{1j}, \sum_{j=1}^{m} p_{2j}, \ldots, \sum_{j=1}^{m} p_{mj}\right) + H\left(\sum_{i=1}^{n} p_{i1}, \sum_{i=1}^{n} p_{i2}, \ldots, \sum_{i=1}^{n} p_{im}\right).$$

Axiom 6. We have for joint probability $p_i q_j$, $1 \leq i \leq n$, $1 \leq j \leq m$,

$$H(p_1 q_1, p_1 q_2, \ldots, p_1 q_m, p_2 q_1, p_2 q_2, \ldots, p_2 q_m, \ldots, p_n q_1, p_n q_2, \ldots, p_n q_m) = H(p_1, p_2, \ldots, p_n) + H(q_1, q_2, \ldots, q_m).$$

Thus, the uncertainty of a joint probability should equal the sum of uncertainties of the corresponding marginal distributions.

Axiom 7. The uncertainties of probabilities with components equal to $1/n$ (as in Axiom 4) should increase with increasing n.

Axiom 8. A probability distribution must have the *branching property*. Given $A = \{x_1, x_2, \ldots, x_s\}$ and $B = \{x_{s+1}, x_{s+2}, \ldots, x_n\}$, then

$$H(p_1, p_2, \ldots, p_n) = H(p_A, p_B) + p_A H\left(\frac{p_1}{p_A}, \frac{p_2}{p_A}, \ldots, \frac{p_s}{p_A}\right) + p_B H\left(\frac{p_{s+1}}{p_B}, \frac{p_{s+2}}{p_B}, \ldots, \frac{p_n}{p_B}\right).$$

Axiom 9. $H(1/2, 1/2) = 1$ if the units of H are to be bits.

Shannon's theorem is then as follows.

Theorem 6.1. The only functional H on $\mathcal{P}^n$ that satisfies the axioms listed above is the Shannon entropy $H(p) = -\sum_{i=1}^{n} p_i \log p_i$.

Details of the proof can be found in Klir (2005) or, in slightly different forms of the axioms, in Jaynes (2003) and of course Shannon (1948). As noted in (4.16), for the continuous densities $p = p(\boldsymbol{x})$, we write

$$H(p) = -\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{6.11}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

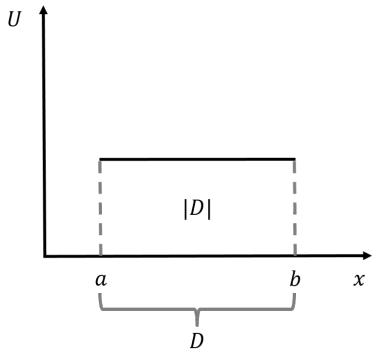


Figure 6.1. An m-dimensional example of uniform distribution over a domain D.

While in much of the literature on information entropy, (6.11) is given as the definition of entropy for the case of continuous $p = p(\boldsymbol{x})$, it is not technically correct as it is not invariant under a change of variables, *e.g.* under a translation. It can be corrected by the introduction of a function $m(x)$, so that

$$H(p) = -\int p \log \frac{p}{m} \, \mathrm{d}\boldsymbol{x},$$

where m is an invariant measure proportional to a limiting density of a discrete set of sample points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ (see *e.g.* Jaynes 2003, p. 375).

6.6. Example: entropy of the uniform distribution

Let U be a *uniform random variable*, that is, U takes on some constant value in a bounded domain $D \subset \mathbb{R}^n$, with volume (or measure) $|D| < \infty$ (see Figure 6.1). In one dimension, the corresponding density would be

$$\pi(x) = \frac{1}{|D|}, \quad a \leq x \leq b,$$

meaning that for any x in $[a, b]$, the same probability $1/(b-a)$ occurs. The entropy of U is

$$H(U) = \log(|D|).$$

The question arises, how uncertain is the random variable U? The answer is that the entropy of any other random variable X (call it $H(X)$ for now) will always be smaller than $H(U)$:

$$H(X) \leq H(U).$$

The uniform distribution (or, for instance, the uniform prior $\pi_{prior}(x) =$ const.) is thus the distribution of maximum uncertainty (or, as some say, maximum ignorance).

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

Considerably more information is contained in the case in which the mean μ and the variance σ^2 of a distribution are known. Then, the correct distribution $\pi(x)$ must maximize the entropy subject to appropriate constraints, that is,

$$\begin{aligned}\pi(x) &= \arg\max_{\pi} H(x) \\ &= \arg\max_{\pi}\left\{\int_{\mathbb{R}} \pi(x)\log\pi(x)\,\mathrm{d}x\right\},\end{aligned}$$

subject to the constraints

$$\begin{aligned}\int_{\mathbb{R}} \pi(x)\,\mathrm{d}x &= 1,\\ \int_{\mathbb{R}} x\pi(x)\,\mathrm{d}x &= \mu,\\ \int_{\mathbb{R}} (x-\mu)^2\pi(x)\,\mathrm{d}x &= \sigma^2.\end{aligned}$$

Constructing a Lagrangian functional Υ defined by the entropy and Lagrange multipliers for these constraints, we easily solve the constrained maximization problem and establish that the solution is

$$\pi(x) = \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\},$$

a Gaussian, as was probably expected.

This process of computing prior distributions by maximizing the entropy subject to constraints was introduced by Jaynes (1968) and used by him to generate priors for very general situations in which some *a priori* information is known about the variable x (generally, in our applications $x = \theta$, the parameters). We examine Jaynes' approach more thoroughly in the next section.

6.7. Maximum entropy priors: the general solution for maximum entropy with linear constraints

An important application of the Maximum Entropy Principle for constructing prior distributions when some statistical information is known, was developed by Jaynes (1968) and described later in his book *Probability Theory: The Logic of Science* (Jaynes 2003). This section reviews the basic ideas.

Consider a real-valued variable x about which we have n propositions $\{A_1, A_2, \ldots, A_n\}$. We have prior information I about x that takes on values $\{x_1, x_2, \ldots, x_n\}$ corresponding to the respective propositions. The probabilities p_i that we assign to the possibilities x_i form the set $\{p_1, p_2, \ldots, p_n\}$.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

Next, suppose we have a set of m functions of x, $f_k(x)$, $1 \leq k \leq m < n$. We want to prescribe the expectations of the f_k relative to the probability distribution $\{p_1, p_2, \ldots, p_n\}$:

$$\langle f_k(x) \rangle = \sum_{i=1}^{n} p_i f_k(x_i) = F_k, \tag{6.12}$$

where F_k are given numbers. *Thus, given the F_k and given the x_i, we wish to find the probabilities p_i which yield maximum entropy, subject to the relevant constraints, including $\sum_i p_i = 1$.* This leads to the problem of finding critical points of the functional

$$\Upsilon(p, \lambda_0, \lambda_1, \ldots, \lambda_m) = H(p) - \lambda_0' \left(1 - \sum_{i=1}^{n} p_i\right) - \sum_{k=1}^{m} \lambda_k \left(F_k - \sum_{i}^{n} p_i f_k\right), \tag{6.13}$$

where $\lambda_0', \lambda_1, \ldots, \lambda_m$ are Lagrange multipliers and $f_k = f_k(x_i)$. We have

$$\sum_{i=1}^{n} \left[\frac{\partial H(p)}{\partial p_i} \delta p_i - \lambda_0' \delta p_i + \sum_{k=1}^{m} \lambda_k f_k \delta p_i \right] = 0, \quad \text{for all } \delta p_i.$$

From this one can show that

$$p_i^{\max} = p_i = \exp\left(\lambda_0 - \sum_{k=1}^{m} \lambda_k f_k\right), \tag{6.14}$$

where $\lambda_0 = \lambda_0' + 1$. Returning to the constraints, we have

$$1 = \sum_{i=1}^{n} p_i = \exp\left[-\lambda_0 \sum_{i}^{n} \exp\left(-\sum_{k=1}^{m} \lambda_k f_k\right)\right].$$

At this point, we introduce the *partition function*

$$Z(\lambda_1, \lambda_2, \ldots, \lambda_m) = \sum_{i}^{n} \exp\left(-\sum_{k=1}^{m} \lambda_k f_k\right), \tag{6.15}$$

so

$$\lambda_0 = \log Z. \tag{6.16}$$

Thus, we have

$$F_k = \sum_{i}^{n} \exp\left[\log Z - \sum_{k=1}^{m} \lambda_k f_k(x_i)\right],$$

which we recognize as

$$F_k = -\frac{\partial}{\partial \lambda_k} (\log Z(\lambda_1, \lambda_2, \ldots, \lambda_m)), \tag{6.17}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

which determines the λ_k as functions of the given F_k. The maximum entropy is

$$H_{\max} = -\sum_{i=1}^{n} p_i^{\max} \log p_i^{\max},$$

where $p_i^{\max}$ is given by (6.14). Thus,

$$H_{\max} = \lambda_0 + \sum_{k=1}^{m} \lambda_k F_k. \tag{6.18}$$

We can now assert that H attains its maximum if and only if $p_i = p_i^{\max}$, that is,

$$p_i = \frac{1}{Z} \exp\Big[-\sum_{k=1}^{m} \lambda_k f_k(x_i)\Big]. \tag{6.19}$$

One significance of problem (6.12) is this: if we know certain properties of a distribution, such as its mean, moments, or other features F_k, then we construct the prior which fits these constraints and which maximizes the entropy. Further details can be found in Jaynes (2003).

7. Monte Carlo methods: MCMC and sensitivity analysis

7.1. Introduction

Some basic issues are, as yet, unaddressed in our introduction to Bayesian inference and statistical calibration and validation. First and foremost, how do we numerically evaluate the influence of the posterior distributions arising from Bayes' rule on key quantities of interest? This issue of numerically evaluating Bayes' rule was a principal contributing factor to the long-dormant period in which the importance of Bayesian approaches was largely ignored by the statistics community. Secondly, there is the question of what model parameters appreciably affect the quantities of interest in a computer simulation. Specifically, in almost every application of interest, the *quantities of interest* (QoIs) which are the target centrepiece of the prediction are not influenced by the choice of some, often many, values of the model parameters. This raises the question: Of all the parameters in a given model, and given a QoI, which parameters can be eliminated, because variations in them have negligible effect on the values of the QoIs?

Remarkably, both of these questions can be addressed with great efficiency using *Monte Carlo* (MC) methodologies. Monte Carlo methods represent a broad class of computational algorithms that rely on repeated random samples of a function to obtain numerical representations of values of the function over its domain. In recent years, there have been many hundreds, perhaps thousands of papers on various versions of MC methods applied to

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

problems in UQ (uncertainty quantification). A comprehensive review of the mathematical structures of many of these methods can be found in the notes of Johansen and Evers (2007). Our goal here is to give a very brief review of the classical ideas for reference in later sections.

The term 'Monte Carlo' was used by the earlier inventors and users of the idea, in analogy with gambling with random bets at the Monte Carlo Casino in Monaco. It was introduced by Manhattan Project scientists Nicholas Metropolis and Stanislaw Ulam (see Metropolis *et al.* 1953), and with further improvements by Wilfred Hastings (1970). The basic idea involves solving deterministic problems using a probabilistic analogy.

The most common Monte Carlo algorithm is that associated with integration of a function $f(x)$ over a domain Ω, $\int_\Omega f(x)\,\mathrm{d}x$. There the idea is to introduce a set X of randomly selected points $X = \{x_1, x_2, \ldots, x_n\}$ in Ω and approximate the integral by the sum

$$\int_\Omega f(x)\,\mathrm{d}x \approx |\Omega|\frac{1}{n}\sum_{i=1}^{n} f(x_i), \tag{7.1}$$

$|\Omega|$ being the volume of Ω. In actual applications, it is sufficient to employ a pseudorandom set of sample points. The error is $O(1/\sqrt{n})$:

$$\mathrm{Var}\left(\int_\Omega f(x)\,\mathrm{d}x - |\Omega|\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) = \frac{1}{n}\left\{\sum_{i=1}^{n} f^2(x_i) - \left[\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right]^2\right\}. \tag{7.2}$$

In the applications of Bayes' rule of interest here, the goal is to use Monte Carlo sampling of a distribution to trace out its variation over its domain, as opposed to integrating a posterior over that domain. Thus, we have, for example, the posterior probability density $\pi_{post}(\boldsymbol{\theta}|\boldsymbol{y}, S)$ – denoted $\pi_{post}(\boldsymbol{\theta})$ for simplicity – and we generate a sequence of samples from $\pi_{post}(\boldsymbol{\theta})$ by randomly selecting parameter value $\boldsymbol{\theta}_i$ in the domain $\mathrm{dom}\,\pi_{post}$. The challenge is to develop a systematic strategy to sample all (or a sufficient number) of values to fully represent the structure of π_{post}. This we accomplish by following a random walk over the contours of the surface which is defined by a Markov chain, described in Section 6.2. An example of such an approximation of the surface is given in Figure 7.1.

As to sensitivity analysis, the idea is to find out how the QoI, or an appropriate surrogate value Y, varies when specific parameters of a model take on random incremental variations. To quote the famous Bayesian proponent Sir Harold Jeffreys (1998, p. 414), 'Parameters in a law that make no contribution to the results of any observation can be eliminated mathematically, leaving the observations to be described only in terms of the relevant parameters.' Our only reservation concerning Jeffreys' suggestion

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

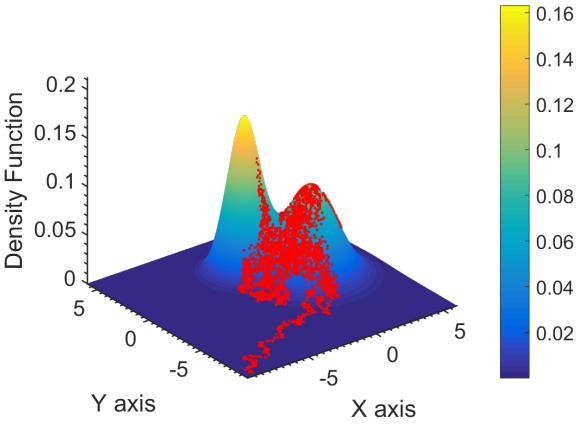


Figure 7.1. The structure of a posterior distribution $\pi_{post}(\boldsymbol{\theta})$ depending on the parameters $\boldsymbol{\theta} = (x, y)$ constructed by computing samples along a random walk, starting at $\boldsymbol{\theta} = (-9, -9)$.

is that setting a parameter θ to zero must be consistent with the definition of its parameter space, that is, we must have $0 \in \Theta$.

Ideally, we want to know how large are the variations in the QoI due to variations in the parameters. The parameter sensitivities are thus related to the numbers

$$\left|\frac{\partial Q(u(\boldsymbol{\theta}, S))}{\partial \boldsymbol{\theta}_i}\right|, \quad 1 \leq i \leq n, \ \boldsymbol{\theta}_i \in \theta \in \Theta. \tag{7.3}$$

Unfortunately, such sensitivity calculations are often very computationally intensive and do not provide a useful measure of parameter sensitivity. Instead, we are led to surrogate measures of output dependency on parameters. Thus if $Y = Y(\boldsymbol{\theta})$ is an output functional (say Y is a QoI), then we seek numerical methods which compute the sensitivity of $Y(\boldsymbol{\theta})$ to perturbations in model parameter $\boldsymbol{\theta}$. We describe such sensitivity measures and calculations advocated by Saltelli, Chan and Scott (2000) below. As mentioned earlier, these sensitivity calculations employ Monte Carlo type methodologies.

7.2. Markov chains

In this section we write out for future reference a very brief introduction to Monte Carlo sampling methods underlying the famous and powerful Markov chain Monte Carlo methods and later variance-based Monte Carlo sampling methods of sensitivity analysis for model sensitivities. We must build a representation of the structure of the surface of probability density such as $\pi_{post}(\boldsymbol{\theta})$ by using a Monte Carlo sampling of a process that traces out a path connecting the random sample points $\boldsymbol{x}$ in a way that can cover

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

the domain of the function as the number of samples is increased. Such a sampling can be performed by a *Markov chain*, which is a sequence of random variables generated by sampling along paths in the function domain called a chain. A Markov chain is a random process that traverses from one 'state' to another in such a way that it creates a 'chain' through a state space in which the predicted distribution on the next state in a chain of probabilities depends only on the current state and not on the events or the probabilities that produced them. Such a 'memoryless-ness' property is called a *Markov property*. In applications of interest here, we pick an initial point in, say, $\mathbb{R}^n$, and create a *random walk*, a path through the space, guided by a set of decision rules.

In a more formal setting, a Markov chain can be considered to be a sequence of random variables X, Y, Z accessed by a process in a certain order, $X \to Y \to Z$, such that the conditional probability of Z depends only on Y and is independent of X. The joint probability (or mass) can thus be written as

$$\begin{aligned} \pi(x, y, z) &= \pi(z|y, x)\pi(y|x)\pi(x) \\ &= \pi(z|y)\pi(y|x)\pi(x) \\ &= \pi(z|x, y)\pi(x|y) \\ \Rightarrow \pi(x, z|y) &= \pi(x|y)\pi(z|y). \end{aligned} \tag{7.4}$$

The memoryless property shows that no processing of information about step Y can increase the information at Y about information at X (written in terms of mutual information as $I(X, Y) \geq I(X, Z)$). See Cover and Thomas (2006).

7.3. Markov chain Monte Carlo

The challenge in applying Monte Carlo sampling is to limit the number of samples in regions of the domain of the integrand where the integrand is very small. A process for tracing out the random walk along a chain that selects directions in which the highest probability of capturing meaningful values of the integrand is the goal of the *Markov chain Monte Carlo* (MCMC) strategy.

Suppose that we want to map out and explore the structure of a posterior density $\pi_{post}(\boldsymbol{\theta})$ $(= \pi(\boldsymbol{\theta}|\boldsymbol{y}, S))$, through a series of trial points $\boldsymbol{\theta}^k$ on a Markov chain starting at an initial value $\boldsymbol{\theta}^0$ of the parameters $\boldsymbol{\theta}$, and designed to have the Markovian property: that each new sample $\boldsymbol{\theta}^{(k+1)}$ does not depend on the earlier samples, beyond $\boldsymbol{\theta}^{(k-1)}$. The new value $\boldsymbol{\theta}^{(k+1)}$ depends on the previous sample $\boldsymbol{\theta}^k$ through a *transition probability* $p(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)})$ that is designed to make the path and samples generate all of the posterior.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

A first step is to pick a *proposal (conditional) distribution* $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k)})$ that gives the value $\boldsymbol{\theta}^*$ of the $k+1$ step, given the kth value $\boldsymbol{\theta}^{(k)}$. The density q can be of any form, such as multivariate Gaussian with mean equal to the current sample $\boldsymbol{\theta}^{(k)}$. The density then decreases with distance away from $\boldsymbol{\theta}^{(k)}$, the current sample.

Secondly, we must decide if we accept the candidate sample $\boldsymbol{\theta}^*$. We do this on the basis of the *Metropolis ratio*

$$r = \frac{\pi_{post}(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{(k)})} \times \frac{q(\boldsymbol{\theta}^k|\boldsymbol{\theta}^{(*)})}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k)})}. \tag{7.5}$$

If the proposal distribution is symmetric, then the second factor is 1. If not, and $r \geq 1$, we set $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^*$. If $r < 1$, we accept $\boldsymbol{\theta}^*$ with probability r. This latter step is done by sampling a random variable U from the uniform distribution $U(0,1)$. If $U \leq r$, we set $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^*$; otherwise, $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)}$.

This second step is captured by the *acceptance probability* $\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(k)})$, defined by

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(k)}) = \min(1, r) = \min\left(1, \frac{\pi_{post}(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^*)}{\pi_{post}(\boldsymbol{\theta}^k)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k)})}\right). \tag{7.6}$$

Thus,

$$\boldsymbol{\theta}^{(k+1)} = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \min\{1, \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(k)})\}, \\ \boldsymbol{\theta}^{(k)} & \text{with probability } 1 - \min\{1, \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(k)})\}\ . \end{cases} \tag{7.7}$$

This version of MCMC is called the *Metropolis–Hastings algorithm.* It is used as the basis for many popular statistical sampling programs, and, in particular, the code QUESO (Prudencio and Schulz 2011) is employed in calculations in later sections.

7.4. Parameter sensitivity analysis

As noted earlier, it is frequently the case that QoIs or experimental observations are not sensitive to changes in certain model parameters. Since computational cost and complexity may strongly depend on the number of parameters appearing in a model, any rational process that leads to a reduction in the number of relevant parameters is regarded as very valuable. This fact has led to the development of several effective methods for parameter sensitivity analysis, represented in Sobol' (1990, 1993, 2007), Homma and Saltelli (1996) and Saltelli (2002), as well as Saltelli *et al.* (2000, 2008, 2009, 2010) and Saltelli and Sobol' (1995). It must be appreciated that parameter sensitivity depends on the particular QoI; parameter changes may strongly affect some QoI values but not affect other choices of QoIs at all.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

The setting for parameter sensitivity analysis is as follows. For a given parametric model involving a set of parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$, select an *output function* $Y(\boldsymbol{\theta})$ that represent a principal output of interest. The output function is thus a surrogate for the QoI, and indeed, when feasible, one can select $Y(\boldsymbol{\theta})$ to be the QoI: $Y(\boldsymbol{\theta}) = \tilde{Q}(\theta) = Q(u(\boldsymbol{\theta}, S_p))$, S_p being the prediction scenario. The goal is to measure the sensitivity of $Y(\boldsymbol{\theta})$ to changes in components of $\boldsymbol{\theta}$.

7.4.1. Morris's method of elementary effects

A popular method of parameter sensitivity analysis is that of Morris (1991) as the method of *elementary effects*. The principal ideas behind the method are listed as follows.

(1) We identify an output function $Y(\boldsymbol{\theta}) = Y([\theta_1, \theta_2, \ldots, \theta_n])$ which is an output model of the QoI and is a function of the n-dimensional vector of parameters $\boldsymbol{\theta}$.

(2) We rescale each parameter θ_i to the unit interval $[0, 1]$.

(3) We consider an n-dimensional hypercube, $\Omega^n = [0, 1]^n$, which we partition into p^n cells, p being an integer number > 2, such that each parameter domain $[0, 1]$ is partitioned into bins

$$\theta_i \sim [0, 1] \sim \left[0, \frac{1}{p-1}, \frac{1}{p-2}, \ldots, 1\right].$$

These bins $[1/(p-i), 1/(p-i-1)]$ contain sample points from which Monte Carlo samples of θ_i will be selected. The bins are called the *levels* of the cube.

(4) We introduce a sampling distribution π_i for each θ_i, π_i possibly denoting a prior density, or simply a uniform distribution assigned to all θ_i, $i = 1, 2, \ldots, n$.

(5) We next choose a small number δ, for instance $\delta = p/(2(p-1))$, $p \geq 2$, and compute the elementary effect EE_i by the difference approximations

$$EE_i = \frac{Y(\theta_1, \ldots, \theta_{i-1}, \theta_i + \delta, \theta_{i+1}, \ldots, \theta_n) - Y(\boldsymbol{\theta})}{\delta}, \quad 1 \leq i \leq n. \tag{7.8}$$

The EE_i thus represent the change in the output due to a δ-perturbation in each parameter θ_i.

(6) We now compute EE_i^j of the output for random (Monte Carlo) samples taken from each bin. If the parameter distribution is not uniform, we sample quantiles in $[0, 1]$ deriving input values from the inverse of the cumulative distribution function.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

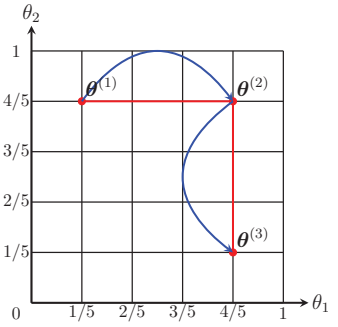


Figure 7.2. An illustration of a two-dimensional hypercube ($n = 2$) with $p = 6$ levels, $\delta = 3/5$, and a trajectory r along $n+1 = 3$ sampling points: $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, $\boldsymbol{\theta}^{(3)}$.

(7) Now we must traverse through the hypercube doing sample paths, called *trajectories*, consisting of paths of $n+1$ orthogonal steps through Ω^n defined as follows.

(a) Select a random starting point $\boldsymbol{\theta}^{(1)}$.

(b) If $\boldsymbol{\theta}_i$ denotes a set of unit coordinate vectors on the axes of the hypercube, select a new vector $\boldsymbol{\theta}^{(2)}$ differing from $\boldsymbol{\theta}^{(1)}$ in its ith component by δ,

$$\boldsymbol{\theta}^{(2)} = \boldsymbol{\theta}^{(1)} \pm \boldsymbol{e}_i\delta,$$

with $+$ or $-$ taken such that $\boldsymbol{\theta}^{(2)} \in \Omega$.

(c) Continuing in this manner,

$$\boldsymbol{\theta}^{(3)} = \boldsymbol{\theta}^{(2)} \pm \boldsymbol{e}_j\delta,$$

we finally get

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} \pm \boldsymbol{e}_k\delta.$$

This defines a trajectory r through Ω^n. An example is provided in Figure 7.2.

(8) Next, we compute the mean and variance of the elementary effects over r trajectories, that is,

$$\mu_i = \frac{1}{r}\sum_{j=1}^{r} EE_i^j, \tag{7.9}$$

$$\sigma_i^2 = \frac{1}{r-1}\sum_{j=1}^{r}(EE_i^j - \mu_i)^2. \tag{7.10}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

Thus μ_i is the mean sensitivity and σ_i^2 is the variance of the ith parameter sensitivity over r trajectories. A common measure of the sensitivity assigned to parameter i is

$$\mu_i^* = \frac{1}{r}\sum_{j=1}^{r} |EE_i^j|. \tag{7.11}$$

By laying out the numbers $\mu_1^*, \mu_2^*, \ldots, \mu_n^*$, we are able to estimate the relative sensitivity of the output to changes in the parameters. When a measure μ_k^* is regarded as very 'small' according to some preset tolerance, that parameter can be eliminated or set equal to a deterministic constant, this latter choice sometimes being made when sensitivities of other parameters are correlated with those of θ_i. We note that the sensitivity measures μ_i, σ_i^2 and μ_i^* depend on the number r of trajectories. Presumably, these measures should converge to constants as $r \to \infty$.

7.4.2. Variance-based sensitivities

The so-called variance-based sensitivity measure is advocated in several papers and books (*e.g.* Cukier *et al.* 1973, Sobol' 1993, Saltelli *et al.* 2000). As in the case of the elementary effects measures, we construct an n-dimensional unit hypercube $\Omega^n = \{\theta,\ 0 \le x_i \le 1,\ \ i = 1, 2, \ldots, n\}$ and generate random (Monte Carlo) samples of the uncertain inputs $\theta_1, \theta_2, \ldots, \theta_n$ from some prescribed distribution (generally uniform or Gaussian). Let $\boldsymbol{\theta}_{\sim i}$ denote the parameter vector which excludes the ith component of $\boldsymbol{\theta}$,

$$\boldsymbol{\theta}_{\sim i} = (\theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)^T$$

We let $V_i(y|\boldsymbol{\theta}_{\sim i})$ denote the variance of the Monte Carlo sample of Y evaluated at $\boldsymbol{\theta}_{\sim i}$ and let $\mathbb{E}_{\boldsymbol{\theta}_{\sim i}}[V_i(y|\boldsymbol{\theta}_{\sim i})]$ be the average over all such variances. Then the *variance-based sensitivity* of θ_i is defined by

$$S_{Ti} = \frac{\mathbb{E}_{\boldsymbol{\theta}_{\sim i}}[V_i(y|\boldsymbol{\theta}_{\sim i})]}{V(Y)}, \tag{7.12}$$

where $V(Y)$ is the variance of the output function.

Generally, it is convenient to regard the output as being defined by a square-integrable function $f(\theta)$ defined on the n-dimensional hypercube Ω^n. Then, to derive (7.12) we introduce the so-called Hoeffding decomposition of f (Sobol' 1990, 1993) and the conditional expectations $\mathbb{E}(Y|\theta_i)$ to arrive at the following decomposition of the output variance:

$$V(Y) = \sum_i V_i + \sum_j \sum_{j>i} V_{ij} + \cdots + V_{12\cdots m}, \tag{7.13}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

where

$$\begin{aligned} V_i &= V(f_i(\theta_i)) = V_{\theta_i}(\mathbb{E}_{\boldsymbol{\theta}_{\sim i}}(Y|\theta_i)), \\ V_{ij} &= V(f_{ij}(\theta_i,\theta_j)) \\ &= V_{\theta_i\theta_j}(\mathbb{E}_{\boldsymbol{\theta}_{\sim i}}(Y|\theta_i,\theta_j)) - V_{\theta_i}(\mathbb{E}_{\boldsymbol{\theta}_{\sim i}}(Y|\theta_i)) - V_{\theta_j}(\mathbb{E}_{\boldsymbol{\theta}_{\sim j}}(Y|\theta_j)). \end{aligned} \tag{7.14}$$

Hence $V_{\theta_i}(\mathbb{E}_{\boldsymbol{\theta}_{\sim i}}(Y|\theta_i))$ and $\mathbb{E}_{\boldsymbol{\theta}_{\sim i}}(Y|\theta_i,\theta_j)$ are called first-order effects, and $V_{\theta_i\theta_j}(\mathbb{E}_{\boldsymbol{\theta}_{\sim i}}(Y|\theta_i,\theta_j))$ are the effects of the pair (θ_i,θ_j) on V, *etc.* Details can be found in Homma and Saltelli (1996); see also the appendix of Farrell, Oden and Faghihi (2015).

For certain cases, Y depends on the interaction between different parameters θ_i and θ_j $(i \neq j)$. These are non-additive models. Then we divide (7.14) by $V(Y)$ to obtain sensitivity indices $S_i, S_{ij}, \ldots, S_{1,2,\ldots,m}$, yielding

$$\sum_i S_i + \sum_i \sum_{j>i} S_{ij} + \cdots + S_{1,2,\ldots,m} = 1. \tag{7.15}$$

The *total effective index* can be derived from this property and (7.12), that is,

$$\begin{aligned} S_{T_i} &= \frac{\mathbb{E}_{\theta_{\sim i}}(V_{\theta_i}(Y|\theta_i))}{V(Y)} \\ &= 1 - \frac{V_{\theta_i}(\mathbb{E}_{\theta_{\sim i}}(Y|\theta_i))}{V(Y)}. \end{aligned} \tag{7.16}$$

S_{T_i} measures the total contributions to the output variation in parameter θ_i. Algorithms based on Monte Carlo sampling for computing such indices are discussed in Saltelli (2002), Saltelli *et al.* (2008), Saltelli, Chan and Scott (2009) and Saltelli *et al.* (2010); a slightly amended version is given in Farrell *et al.* (2015).

7.4.3. Scatter plots

A generally convenient and often inexpensive qualitative way to visualize the sensitivity of outputs to parameter variations is through scatter plots, which involve clouds of points showing an output $Y(\boldsymbol{\theta})$ plotted against randomly chosen values of each parameter in a model. Examples are shown in Figure 7.3. The idea is to take random (Monte Carlo) samples of a uniform distribution of a parameter $\boldsymbol{\theta}_i$ and plot the corresponding $Y(\theta_i)$ values over a two-dimensional domain. A parameter is considered to have a possibly important influence on $Y(\theta)$ if the cloud exhibits a distinct pattern. A uniform cloud, or a cloud close to zero, suggests that the sensitivity of the output to changes in that particular parameter is small.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

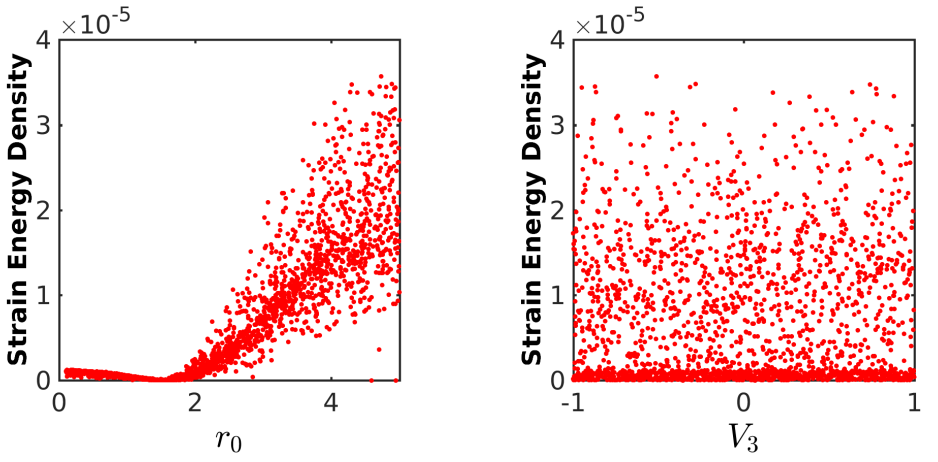


Figure 7.3. Parameters may have a significant influence on the output. The uniform distribution in (b) suggests an insensitivity of the outputs to the parameters. (Reprinted from Farrell *et al.* 2015 with permission from Elsevier.)

7.5. Design of validation experiments

The calculation of parameter sensitivities can provide an important tool in assessing the effectiveness of the design of the validation experimental scenario S_v. When feasible, one computes parameter sensitivities for the output functions $Y(\boldsymbol{\theta})$ set equal to the QoI, $Q(\boldsymbol{\theta})$, in the full prediction scenario S_p. One then recomputes the sensitivity for a proposed validation scenario with $Y(\boldsymbol{\theta})$ set equal to the validation observables. If the validation sensitivities are not in good agreement with those for the full prediction scenario, then the validation experiments are ineffective and cannot determine parameters that inform the prediction of the QoI. Such comparisons of S_p and S_v sensitivities provide a useful indication of the effectiveness of the design of the validation experiments. In the same vein, comparing S_c and S_p sensitivities can also provide insight in designing calibration experiments.

7.6. An algorithm for sensitivity indices

Returning to (7.12), a Monte Carlo scheme to compute the indices with a reduced computational cost of evaluating multi-dimensional integrals is presented here, due to Faghihi in Farrell *et al.* (2015) and inspired by Saltelli *et al.* (2000). The principal steps are as follows.

(1) Generate two $N \times k$ independent sampling matrices, $\boldsymbol{A}$ and $\boldsymbol{B}$, where each row is a sample point in the hyperspace of k dimensions. N is called a base sample with respect to the probability distributions of the input variables.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

(2) Define a matrix $\boldsymbol{D}_i$ formed by all columns of $\boldsymbol{A}$ except the ith column, which is from $\boldsymbol{B}$.

(3) Compute the model outputs for all the matrices (*i.e.* $N \times 1$ vectors of model output):

$$\mathcal{Y}_A = f(\boldsymbol{A}), \quad \mathcal{Y}_B = f(\boldsymbol{B}), \quad \mathcal{Y}_{D_i} = f(\boldsymbol{D}_i)$$

(4) The total effect of indices can be estimated by

$$S_{T_i} \approx 1 - \frac{\frac{1}{N}\sum_{j=1}^{N} \mathcal{Y}_A^{(j)} \mathcal{Y}_{D_i}^{(j)} - \left(\frac{1}{N}\sum_{j=1}^{N} \mathcal{Y}_A^{(j)}\right)^2}{\frac{1}{N}\sum_{j=1}^{N} \mathcal{Y}_A^{(j)} \mathcal{Y}_A^{(j)} - \left(\frac{1}{N}\sum_{j=1}^{N} \mathcal{Y}_A^{(j)}\right)^2}. \tag{7.17}$$

Relation (7.17) can be obtained by substituting the Monte Carlo estimators

$$V_{\boldsymbol{\theta}_{\sim i}}\left[\mathbb{E}_{\theta_i}(Y|\theta_{\sim i})\right] \approx \frac{1}{N}\sum_{j=1}^{N} \mathcal{Y}_A^{(j)} \mathcal{Y}_{D_i}^{(j)} - \mathbb{E}^2(Y), \tag{7.18}$$

along with

$$V_Y = \mathbb{E}(Y^2) - \mathbb{E}^2(Y) = \frac{1}{N}\sum_{j=1}^{N} \mathcal{Y}_A^{(j)} \mathcal{Y}_A^{(j)} - \left(\frac{1}{N}\sum_{j=1}^{N} \mathcal{Y}_A^{(j)}\right)^2, \tag{7.19}$$

into (7.12).

7.7. An example: coarse-grained models of atomistic systems

An example of the use of variance-based sensitivity analyses is found in comparison of parameter sensitivities for coarse-grained molecular models of atomistic systems in thermodynamic equilibrium. We know that all matter in our universe is made up of atoms. Atoms are characterized according to their number of electrons, atomic mass, number of neutrons, *etc.*, and are tabulated in the well-known Periodic Table of modern chemistry and physics, there being some 118 distinct atomic structures currently known. In classical models of atomic systems, atoms are considered to be point masses in motion due to the action of forces, and their locations with respect to a fixed origin are given by the generalized coordinate vectors $\boldsymbol{r}$, $i = 1, 2, \ldots, n$, with n being the number of atoms. We let $\boldsymbol{p}_i$ denote the momentum vector associated with the ith atom. The motion of such an atomistic system is governed by Hamilton's equations,

$$\dot{\boldsymbol{r}}_i = +\frac{\partial H}{\partial \boldsymbol{p}_i}, \quad \dot{\boldsymbol{p}}_i = -\frac{\partial H}{\partial \boldsymbol{r}_i}, \tag{7.20}$$

with $1 \leq i \leq n$, and the Hamiltonian H is given by

$$H(\boldsymbol{r}, \boldsymbol{p}) = \sum_i \frac{\boldsymbol{p}_i \cdot \boldsymbol{p}_i}{2m_i} + V(\boldsymbol{r}), \tag{7.21}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

where m_i is the mass of particle i and $V(\boldsymbol{r}) = V(\boldsymbol{r}_1, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$ is the force potential. A typical model for the potential $V(\cdot)$ (employed, for example, in the molecular dynamics code LAMMPS: Plimpton 1995) is of the form

$$V(\boldsymbol{r}) = V_{bond}(\boldsymbol{r}) + V_{angle}(\boldsymbol{r}) + V_{dihedral}(\boldsymbol{r}) + V_{non\text{-}local}(\boldsymbol{r}) + V_{coulomb}(\boldsymbol{r}), \quad (7.22)$$

where

$$V_{bond}(\boldsymbol{r}) = \sum_{i=1}^{N_b} \frac{1}{2} k_{ri} (\boldsymbol{r}_i - \boldsymbol{r}_{0i})^2, \quad (7.23)$$

$$V_{angle}(\boldsymbol{r}) = \sum_{i=1}^{N_a} \frac{1}{2} k_{\theta_i} (\theta_i - \theta_{0i})^2, \quad (7.24)$$

$$V_{dihedral}(\boldsymbol{r}) = \sum_{i=1}^{N_d} \sum_{n=1}^{4} \frac{V_n i}{2} [1 + (-1)^{n-1} \cos(n\phi_i)], \quad (7.25)$$

$$V_{non\text{-}local}(\boldsymbol{r}) = \sum_{i=1}^{N_{nl}} \sum_{j>i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{\alpha} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{\beta} \right], \quad r_{ij} \le r_c, \quad (7.26)$$

$$V_{coulomb}(\boldsymbol{r}) = \sum_{i=1}^{N_q - 1} \sum_{j>i}^{N_q} 4\epsilon_0 \left[\frac{q_i q_j}{r_{ij}} \right]. \quad (7.27)$$

Here V_{bond} represents, in general, covalent bonds between atoms with bonding constants k_{ri}, V_{angle} is the potential energy due to change in angles between successive bonding lengths between atoms, the dihedral angle ϕ_i between consecutive links, $V_{non\text{-}local}$ denotes Lennard-Jones-type potentials, generally with $\alpha = 6$, $\beta = 12$ or perhaps $\alpha = 6$, $\beta = 9$, and $V_{coulomb}$ is the coulomb potential for discrete charges q_i and q_j at coordinates $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$, respectively, with $r_{ij} = \|\boldsymbol{r}_j - \boldsymbol{r}_i\|$. Here in (7.23) and (7.24), $\boldsymbol{r}_{0i}$ and θ_{0i} are the original coordinate parameters assumed to be given in the initial configuration of the system, ϕ_i in (7.25) is a torsional degree of freedom, and $\|\cdot\|$ is the Euclidean norm.

In most molecular dynamics applications, the momentum vectors are equated to the Newtonian momentum, $\boldsymbol{p}_i = m_i \dot{\boldsymbol{r}}_i$, and (7.20) reduces to (generally huge) systems of nonlinear second-order differential equations describing a Newtonian model for the *all-atom* (AA) system. Because of the enormous size and complexity of such systems, it is standard practice to construct *coarse-grained* (CG) approximations of the system, obtained by aggregating groups of atoms into N super atoms or molecules, thereby dramatically reducing the size of the model. This results in a large class $\mathcal{M}$ of models with the atom coordinates replaced by molecular coordinates $\boldsymbol{R}_I$, $I = 1, 2, \ldots, N \ll n$, and with a force potential of the same form, but with

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

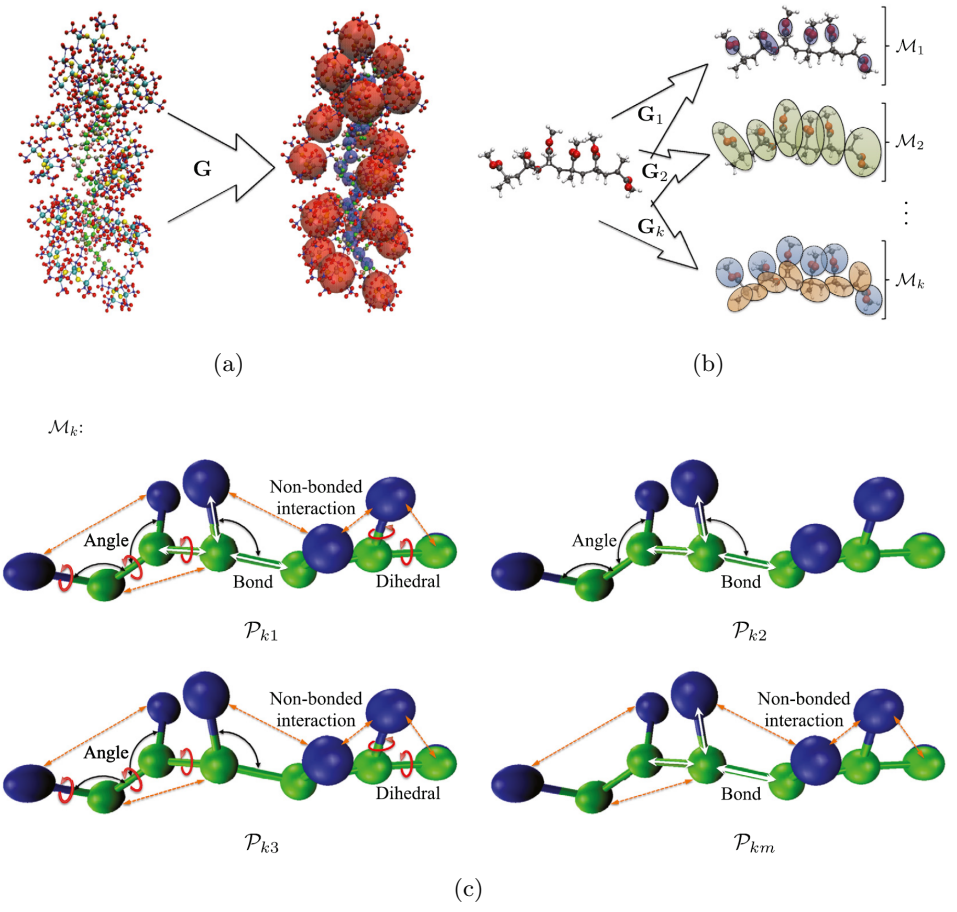


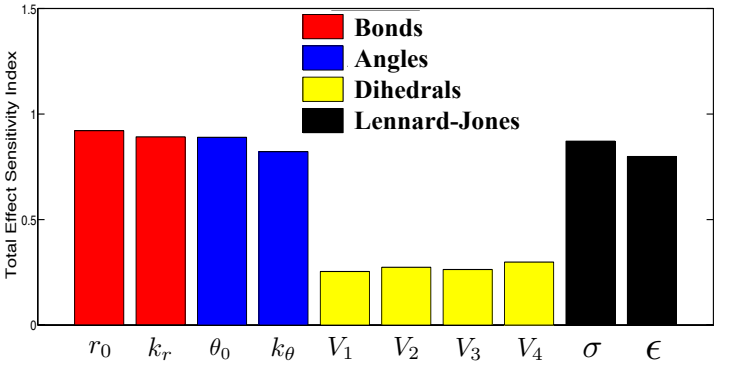
Figure 7.4. (a) Coarse-graining of an all-atom system. (b) Different choices of AA-to-CG map G. (c) Different choices of intermolecular force fields V, for each map G. (Reprinted from Farrell *et al.* 2015 with permission from Elsevier.)

parameters

$$\boldsymbol{\theta} = \{K_{Ri}, R_{0i}, K_{\theta i}, V_{Ni}, \epsilon_{ii}, \sigma_{ij}\}, \quad 1 \leq i \leq N, \tag{7.28}$$

wherein we ignore coulomb-charge effects. A major problem with this very classical and often used process of coarse-graining is that a given AA configuration, which we assume is the *truth* to simplify the analysis, can be segregated into many possible CG configurations, as indicated in the map $G : AA \to CG$ illustrated in Figure 7.4, and for each choice of a CG model, then are numerous choices of force potential, each choice with different sets of parameters.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press



Parameters	S_T
r_0	0.92
k_r	0.89
θ_0	0.89
k_θ	0.82
V_1	0.25
V_2	0.27
V_3	0.26
V_4	0.30
σ	0.87
ε	0.80

Figure 7.5. Bar charts illustrating the results of the sensitivity analysis using total sensitivity indices for the deformation of the polyethylene chain using 20 000 Monte Carlo samples of the parameter space. (Reprinted from Farrell *et al.* 2015 with permission from Elsevier.)

In Farrell *et al.* (2015), this problem was taken up for specific applications involving modelling the stretching of a polyethylene chain, in constant volume and temperature ($T = 100$ K), consisting of 12 'beads' (molecules) with two carbon atoms per bead. The model involved 12 parameters, each assigned a uniform distribution:

$$\begin{aligned} R_0 &\sim \mathcal{U}(0.1, 5), & K_R &\sim \mathcal{U}(0.1, 90), \\ \theta_0 &\sim \mathcal{U}(80, 100), & K_\theta &\sim \mathcal{U}(0, 15), \\ V_1 &\sim \mathcal{U}(-1, 1), & V_2 &\sim \mathcal{U}(-1, 1), \\ V_1 &\sim \mathcal{U}(-1, 1), & V_2 &\sim \mathcal{U}(-1, 1), \\ \sigma &\sim \mathcal{U}(0.05, 7), & \epsilon &\sim \mathcal{U}(0.01, 5). \end{aligned}$$

The goal is to estimate the total sensitivity indices for each parameter with respect to the output function

$$Y(\boldsymbol{\theta}) = \langle V(\boldsymbol{\theta}, G_1) \rangle - \langle V(\boldsymbol{\theta}, G_2) \rangle,$$

where G_0 denotes the initial configuration of the chain and G is the map defining the configurations attained by applying a force to one end of the chain while keeping the other end constrained.

We use (7.12) to compute the total sensitivities of each of these 12 parameters. Results are presented in Figure 7.5, which is reproduced from that in Farrell *et al.* (2015). Scatter plots of some parameters are given in Figure 7.6. From these results we observe that the sensitivities of the dihedral parameters are much smaller than those of other parameters, and we conjecture that by omitting them altogether, the values of the output $Y(\theta)$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

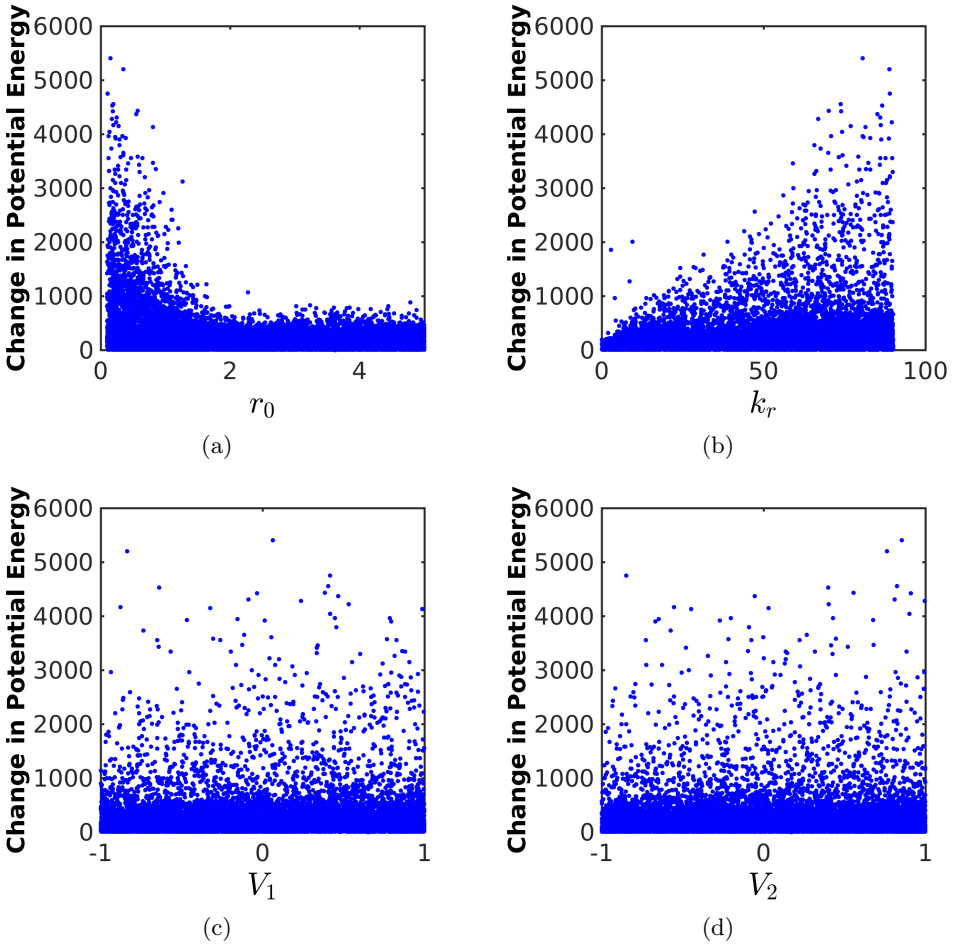


Figure 7.6. Scatter plots illustrating a quantitative sensitivity analysis for the polyethylene chain using 40 000 Monte Carlo samples of the parameter space. The observable is the change in potential energy as the chain is deformed. (Reprinted from Farrell *et al.* 2015 with permission from Elsevier.)

will not change appreciably. To confirm the conjecture, we compute the same output functions except that the dihedral parameters are omitted. Figure 7.7 shows the results of these different models: (i) the full model with covalent bonds, angles, dihedrals and Lennard-Jones terms, (ii) the model without dihedrals (*i.e.* only with bonds, angles and Lennard-Jones terms), and (iii) the model with bonds, dihedral and Lennard-Jones terms. We observe that by eliminating the parameters with lowest sensitivities (the dihedrals), the distributions of the output is essentially unchanged. Hence, we have confirmed that Jeffreys' advice pays off excellently.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

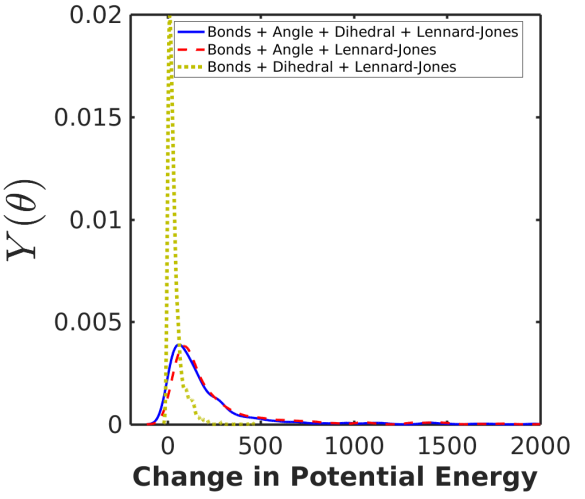


Figure 7.7. Comparison between the kernel density estimates of the change in potential energy resulting from the deformation of a polyethylene chain for the cases when dihedral and angle parameters are excluded from calculations. Results show 500 Monte Carlo samples of the parameter space. (Reprinted from Farrell *et al.* 2015 with permission from Elsevier.)

PART THREE
Adaptive modelling: model selection, OPAL and adaptive control of multiscale models

8. $\mathcal{M}$ uncertainty

8.1. Model selection

There are several Bayesian motivated procedures in the literature for comparing the suitability of models (*e.g.* Bayes factors, Occam factors, BIC = Bayesian Information Criterion), but the 'gold standard' approach focuses on the posterior model evidence or, equivalently, weighted evidence called posterior model plausibilities. Chow (1981) attributes the ideas to the works of Jeffreys (1939, 1961, 1998). The idea of computing model plausibilities appears to have been abandoned for many years owing to the complexity of computing integrals of the distributions in marginalization of the product of the likelihood and prior, but these concerns appear to have been set aside with the advent of modern parallel computers and efficient sampling algorithms. The notion of plausibilities (or the building blocks, model evidences) has been revived in the work by Beck and co-workers (Beck and Yuen 2004, Beck 2010, Beck and Taflanidis 2013) and others (Farrell *et al.* 2015, Prudencio and Schulz 2011).

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

The frequentist work in this area follows the work of Akaike (1974, 1976, 1977, 1998) and its various generalizations (Hurvich and Tsai 1989, Lebreton, Burnham, Clobert and Anderson 1992, Schwarz 1978). A readable account of these latter approaches is found in Burnham and Anderson (2002) and in Konishi and Kitagawa (2008).

In the sections that follow, we introduce the theory of posterior model plausibilities and we outline an introduction to the information-theoretic and frequentist methods of Akaike and others. We then describe a general framework for model selection and model validation embodied in OPAL – the Occam Plausibility Algorithm – introduced and implemented in Farrell *et al.* (2015) and Oden *et al.* (2018), which brings together most of the tools and concepts presented thus far: model calibration, validation, parameter sensitivity, model plausibility, model inadequacy and prediction of the QoIs.

At this point, we have considered Bayesian and entropy-information theoretic approaches to address several of the major sources of uncertainty, including $\mathcal{P}$ uncertainty and $\mathcal{Y}$ and θ uncertainties. What remains is one of the most important problems in predictive science: $\mathcal{M}$ uncertainty, uncertainty in the selection of the centrepiece of prediction, the model used to make the prediction.

For many years, the community of scientists and engineers were taught that model selection was the province of experienced practitioners who had earned the privilege of selecting mathematical models based on divine-given judgement, discounting that the models selected often delivered poor predictions of the observed events. Or more commonly, very classical models of mechanics and physics were morphed and distorted and ultimately used to model events far outside the original domains of the textbook applications. In the meantime, the statistics world was concerned about model selection and methods to guide selection, but most of the time only in a limited sense of calculating likelihoods for regression calculations or low-dimensional approximations.

In more recent times, the theory and methodology of model selection has taken strides forward with the advent of modern computing capabilities and modern algorithms for implementing methods based on Bayesian inference and information theory. In this section, two classes of methods for model selection are discussed, one based on Bayesian ideas involving the notion of posterior model plausibilities, and the other based on frequentist notions, information theory and relative entropy.

8.2. Bayesian model plausibilities

The idea of Bayesian *a posteriori* model plausibilities embodies a rigorous and systematic basis for model selection, given fixed observational data. The setting is as follows: for a given scenario $S \in \{S_c, S_v, S_p\}$ and given

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

data $\boldsymbol{y} \in \mathcal{Y}_S$, we have a collection of m possible models that are proposed to predict $\boldsymbol{y}$, each characterized by a forward problem $A_j(\boldsymbol{\theta}_j, S, u(\boldsymbol{\theta}_j, S)) = 0$, $j = 1, 2, \ldots, m$, each model having its own parameters and parameter space, $\boldsymbol{\theta}_j \in \Theta_j$. We simplify notation by writing simply $\mathcal{P}_j$ or $\mathcal{P}_j(\boldsymbol{\theta}_j)$ for model j, and we denote by $\mathcal{M}$ the set of m such models: $\mathcal{M} = \{\mathcal{P}_1(\boldsymbol{\theta}_1), \ldots, \mathcal{P}_m(\boldsymbol{\theta}_m)\}$. For each model $\mathcal{P}_j$ a prior $\pi(\boldsymbol{\theta}_j|S)$ is assumed to be specified.

Next, we simply write down Bayes' rule, now acknowledging that we have additional information, namely that the models $\mathcal{P}_j(\boldsymbol{\theta}_j)$ are known and that they belong to the set $\mathcal{M}$. Thus, invoking Jaynes' principle, that to be honest we must use 'all of the evidence we have, not just some arbitrarily chosen subset of it', we rewrite Bayes' rule as

$$\pi(\boldsymbol{\theta}_j|\boldsymbol{y}, \mathcal{P}_j, \mathcal{M}, S) = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}_j, \mathcal{P}_j, \mathcal{M}, S)}{\pi(\boldsymbol{y}|\mathcal{P}_j, \mathcal{M}, S)} \pi(\boldsymbol{\theta}_j|\mathcal{P}_j, \mathcal{M}, S), \quad 1 \leq j \leq m. \tag{8.1}$$

The denominator is the model evidence, and it is now much more than just a normalization term, playing a key role in the analyses. It must be the marginalization of the numerator,

$$\pi(\boldsymbol{y}|\mathcal{P}_j, \mathcal{M}, S) = \int_{\Theta_j} \pi(\boldsymbol{y}|\boldsymbol{\theta}_j, \mathcal{P}_j, \mathcal{M}, S) \pi(\boldsymbol{\theta}_j|\mathcal{P}_j, \mathcal{M}, S) \, \mathrm{d}\boldsymbol{\theta}_j, \quad 1 \leq j \leq m. \tag{8.2}$$

A key observation at this point is that the evidence in (8.2) has the form of a likelihood for a second-level Bayesian inference for the finite-dimensional problems of selection models from the set $\mathcal{M}$. In particular, we write

$$\rho_j = \pi(\boldsymbol{y}|\mathcal{P}_j, \mathcal{M}, S) = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}_j, \mathcal{P}_j, \mathcal{M}, S) \pi(\boldsymbol{\theta}_j|\mathcal{P}_j, \mathcal{M}, S)}{\pi(\boldsymbol{y}|\mathcal{M}, S)}, \tag{8.3}$$

where $\pi(\boldsymbol{y}|\mathcal{M}, S)$ is a normalizing factor chosen so that

$$\sum_{j=1}^{m} \rho_j = 1. \tag{8.4}$$

The model (or models) in $\mathcal{M}$ with plausibilities closest to unity are regarded as the most plausible in $\mathcal{M}$ for given $\boldsymbol{y}$ and S.

The plausibilities are seen to be weighted model evidences, with weights equal to the prior $\pi(P_j|\mathcal{M}, S)$. In the case of m models thought to be equally plausible prior to the calculation of evidence, they may simply be set equal to $1/m$.

We remark that a model $\mathcal{P}_j$ may be the most plausible in a set $\mathcal{M}$ given $\boldsymbol{y}$, but it may be invalid according to specific validation metric and tolerances: model validity and model plausibility are thus two different characteristics of parametric models.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

8.3. Frequentist approaches: the Akaike Information Criterion and others

Methods for model selection based on frequentist arguments or more appropriately, on information theories, are often derived on the notion of information lost when a model $\mathcal{P}_j(\boldsymbol{\theta}_j) \in \mathcal{M}$ deviates from the 'true' distribution ($\boldsymbol{G}$ or g): the best models deliver parameters that define likelihoods closest to the truth. The Akaike methods or their generalizations, introduced in Akaike (1974) and expanded in Akaike (1976, 1977, 1998), are based on this concept.

Once again, given a set $\mathcal{M}$ of models $\mathcal{P}_1(\boldsymbol{\theta}_1), \ldots, \mathcal{P}_m(\boldsymbol{\theta}_m)$ and likelihood densities $\pi(\boldsymbol{y}|\boldsymbol{\theta}_j, S)$, $1 \le j \le m$, the information lost, $I_j(g, \pi)$, when model $\mathcal{P}_j$ is used to approximate the truth $\boldsymbol{g}$ is given by the D_{KL}-divergence,

$$I_j(g,\pi) = D_{KL}(g \parallel \pi(\cdot|\boldsymbol{\theta}_j, S)) \tag{8.5}$$
$$= \begin{cases} \displaystyle\int_{\mathcal{Y}_S} g(\boldsymbol{y}) \log \frac{g(\boldsymbol{y})}{\pi(\boldsymbol{y}|\boldsymbol{\theta}_j, S)} \, \mathrm{d}\boldsymbol{y} & \text{for the continuous case,} \\ \displaystyle\sum_{i=1}^{n} g_i \log g_i/\pi_i & \text{for the discrete case,} \end{cases}$$

for $1 \le j \le m$.

The loss is minimized for each model $\mathcal{P}_j$ by choosing parameters $\boldsymbol{\theta}_j$ to be the MLE: $\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \arg\max_{\boldsymbol{\theta} \in \Theta} \log \pi(\boldsymbol{y}|\boldsymbol{\theta}, S)$. The number A_j representing the information lost when model $\mathcal{P}_j$ is averaged over samples $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$A_j = \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{y}} \big[\log(\pi(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j(\boldsymbol{y}), S)) \big]$$
$$= \int_{\mathcal{Y}_S} \int_{\mathcal{Y}_S} g(\boldsymbol{x}) g(\boldsymbol{y}) \log \pi(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j(\boldsymbol{y}), S) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y}, \quad 1 \le j \le m. \tag{8.6}$$

Akaike (1974) derived approximations of A_j to reduce the complexity of computations, in which it is assumed that the log-likelihood is twice differentiable with respect to the parameters, that the sample $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ is large enough that asymptotic properties valid as $n \to \infty$ hold, and that the sequence of MLEs $\hat{\boldsymbol{\theta}}_j(\boldsymbol{n})$ taken for sample size n converge to the D_{KL}-minimizer $\boldsymbol{\theta}_j^*$. Under such assumptions, the number A_j can be replaced by the computable *Akaike Information Criterion*

$$AIC_j(\boldsymbol{y}) = -2 \log(\pi(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_j, S)) + 2k_j, \quad 1 \le j \le m, \quad \boldsymbol{y} \in \mathcal{Y}_S, \tag{8.7}$$

where k_j is the number of parameters in model $\mathcal{P}_j$, the best model in $\mathcal{M}$ being that with the smallest $AIC_j(\boldsymbol{y})$.

8.4. Akaike weights

In applications of the AIC, the actual AIC_j values are of little use: what one wishes to determine are the relative merits of various models as measured

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

by the relative criteria

$$\Delta_j = AIC_j - AIC_{\min}, \quad 1 \le j \le m. \tag{8.8}$$

The best model among the set of all models has a relative value $\Delta^* = \Delta_{\min} = \min\{\Delta_j\}_{j=1}^m$. It is useful to transform the AIC relative values Δ_j into a model probability by setting them proportional to (minus one-half of) a log-likelihood: $\log \pi_{\text{like}_j} \propto (-\Delta_j/2)$. Then we can define the model probabilities by the Akaike weights:

$$W_j = \frac{\exp\left(-\frac{1}{2}\Delta_j\right)}{\sum_{k=1}^m \exp\left(-\frac{1}{2}\Delta_k\right)}. \tag{8.9}$$

Clearly,

$$\sum_{k=1}^m W_k = 1. \tag{8.10}$$

The weight W_j is clearly the weight of evidence in favour of $\mathcal{P}_j$, and the highest corresponds to the best model.

Several investigators have proposed generalizations of the AIC. For example, the 'second-order' AIC for small sample size has been proposed by Hurvich and Tsai (1989):

$$\begin{aligned} AIC_{cj}(\boldsymbol{y}) &= -2\log(\pi(\boldsymbol{y}|\boldsymbol{\theta}_j, S)) \\ &\quad + 2k_j + \frac{2k_j(1+k_j)}{n-k_j-1} \\ &= AIC_j(\boldsymbol{y}) + \frac{2k_j(1+k_j)}{n-k_j+1}, \end{aligned} \tag{8.11}$$

n being the sample size. Another measure has been proposed in Lebreton *et al.* (1992) when data are 'overdispersed'. Then, instead of $AIC_j(\boldsymbol{y})$, one uses

$$QAIC_j(\boldsymbol{y}) = -2\log\left[\pi(\boldsymbol{y}|\hat{\boldsymbol{\theta}}, S)/\hat{C}\right] + 2k_j \tag{8.12}$$

where $\hat{C}$ is the ratio of the chi-squared statistic χ^2 and its number of degrees of freedom. A small-sample version of this estimator, $QAIC_{cj}(\boldsymbol{y})$, analogous to (8.11), can be obtained by replacing $AIC_j(\boldsymbol{y})$ in (8.11) with $QAIC_j(\boldsymbol{y})$. See Burnham and Anderson (2002) for additional details.

8.5. OPAL: the Occam Plausibility Algorithm

The various procedures and principles for model selection, calibration and validation presented up to this point can be implemented sequentially in a systematic process we refer to as OPAL, the Occam Plausibility Algorithm. It was introduced in Farrell *et al.* (2015) and has been discussed and implemented in, for example, Oden *et al.* (2018). The label 'Occam' derives

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

from Occam's razor, the 'principle' that one should use the 'simplest' theory or model among competing models that lead to essentially the same prediction. Since 'simplest' is a rather subjective label, various measures of simplicity could be conceived. Influenced by model selection criteria such as AIC, we chose as starting point the number k_j of parameters of a model $\mathcal{P}_j$ as a rough measure of its simplicity.

Let us assume that we wish to predict a target QoI representing a specific physical reality, and that we have access to experimental facilities that could provide calibration data. Among many possible computational models we might consider for the prediction, which are valid and most plausible given data are unknown. The steps underlying OPAL are as follows.

(1) *Initialization.* As discussed in the preceding sections, a set $\mathcal{M}$ of m possible models is selected for predicting the QoI:

$$\mathcal{M} = \{\mathcal{P}_1(\boldsymbol{\theta}_1), \mathcal{P}_2(\boldsymbol{\theta}_2), \ldots, \mathcal{P}_m(\boldsymbol{\theta}_m)\}, \tag{8.13}$$

each model having its own parameter space ($\boldsymbol{\theta}_k \in \Theta_k$). The models in $\mathcal{M}$ may be diverse or closely related, some sharing the same types of parameter.

(2) *Sensitivity analysis.* For each model class $\mathcal{P}(\boldsymbol{\theta}_j)$ we compute parameter sensitivities using techniques described in Section 6.3 for the prediction scenario S_p. We follow the advice of Jeffreys (1998, p. 414): 'Parameters in a law that make no contribution to the results of any observation can be eliminated mathematically, leaving the observations to be described only in terms of the relevant parameters.' We make a small amendment to this idea: if $\boldsymbol{\theta}_k$ is the vector of insensitive parameters, we set it equal to zero only if $\mathbf{0} \in \Theta_k$. Otherwise it is assigned a constant (deterministic) value.

Using an appropriate tolerance, only those model classes are retained that contain parameters to which the output is sensitive, as indicated by the relative value of the sensitivity measures deduced in Section 6.3 (and those eliminated for which $\mathbf{0} \notin \Theta_k$). This results in a reduced set $\bar{\mathcal{M}}$ of model classes:

$$\bar{\mathcal{M}} = \{\bar{\mathcal{P}}_1(\boldsymbol{\theta}_1), \bar{\mathcal{P}}_2(\boldsymbol{\theta}_2), \ldots, \bar{\mathcal{P}}_l(\boldsymbol{\theta}_l)\}, \quad l \leq m. \tag{8.14}$$

It is noted that the choice of the number of parameters as the basis for identifying model categories implicitly assumed that the parameters are uncorrelated and independent, which may not always be the case. Furthermore, the implementation of a parameter sensitivity analysis, with scatter plots, can often reveal such dependencies. Also, parameter correlations can be exposed through calculation of auto-correlation functions for distributions of parameter values (Lima *et al.* 2017).

(3) *Occam categories.* The reduced set $\bar{\mathcal{M}}$ is partitioned according to the number of parameters in each. These partitions are called *Occam categories*,

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

with the models containing the smallest number of parameters designated category 1, those in the next higher number category 2, and so forth. This produces a collection of subsets, $\bar{\mathcal{M}}_1, \bar{\mathcal{M}}_2, \ldots, \bar{\mathcal{M}}_N$, with models in set $\bar{\mathcal{M}}_k$, $1 \leq k \leq N$. By our interpretation of Occam's razor, models of category k are simpler or less complex than those of category $l > k$.

(4) *Calibration.* Beginning with models in the category 1 set $\bar{\mathcal{M}}_1$, all models in $\bar{\mathcal{M}}_1$ are calibrated statistically for a specified calibration scenario S_c. The calibrated category 1 models for a new set are then given by

$$\bar{\mathcal{M}}_1^* \sim \{\mathcal{P}_1^*(\boldsymbol{\theta}_1^*), \mathcal{P}_2^*(\boldsymbol{\theta}_2^*), \ldots, \mathcal{P}_k^*(\boldsymbol{\theta}_k^*)\}, \quad k < l. \tag{8.15}$$

It remains to be determined which of these models is most plausible, given calibration data $\boldsymbol{y}_c$, and which among them will prove to be valid, given data $\boldsymbol{y}_v$.

(5) *Plausibility.* Using (8.15) we next calculate the Bayesian posterior plausibilities $\rho_j^*, 1 \leq j \leq k$, of the category 1 model. The most plausible model (or models), say $\mathcal{P}_i^*(\boldsymbol{\theta}_i^*)$, is that with plausibility ρ_i^* such that

$$\rho_i^* \geq \rho_i^*, \quad 1 \leq j \leq k. \tag{8.16}$$

(6) *Validation.* At this point we have identified the most plausible and simplest model in category 1. Is it valid? We now move to the validation scenario S_v, update the calibrated parameters of the most plausible category 1 model using Bayesian statistical inversion as discussed earlier, and solve the forward problem in S_v to establish model predictions on the S_v experimental observational data $\boldsymbol{y}_v$. We must now determine whether the accuracy with which the model prediction agrees with data is sufficient to declare the model 'valid' (*i.e.* not invalid).

Again this step requires that we have selected a suitable metric for measuring error and a suitable tolerance for measuring validity. This step also underscores the fundamental importance of designing validation scenarios that reflect the ability of the model to predict the QoI.

For specificity, let $q_v(\boldsymbol{y})$ denote the probability density of the validation observational data $\boldsymbol{y} = \boldsymbol{y}_v$ and let $\pi_v(\boldsymbol{y}, \boldsymbol{\theta}^{**})$ denote the model-generated density approximating q_v obtained using the updated posterior distributions of S_v. Let $d(\cdot, \cdot)$ be an appropriate metric, and let γ_{tol} be our validation tolerance. Then, for the model to be valid, we must have

$$d(q_v, \pi_v(\boldsymbol{y}, \boldsymbol{\theta}^{**})) \leq \gamma_{tol}, \tag{8.17}$$

One example of such a metric, employed in Farrell *et al.* (2015), is

$$D_{KL}(q_v \parallel \pi_v(\boldsymbol{y}, \boldsymbol{\theta}^{**})) = \int_{S_v} q_v(\boldsymbol{y}) \log \frac{q_v(\boldsymbol{y})}{\pi(\boldsymbol{y}, \boldsymbol{\theta}^{**})} \, \mathrm{d}\boldsymbol{y}. \tag{8.18}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

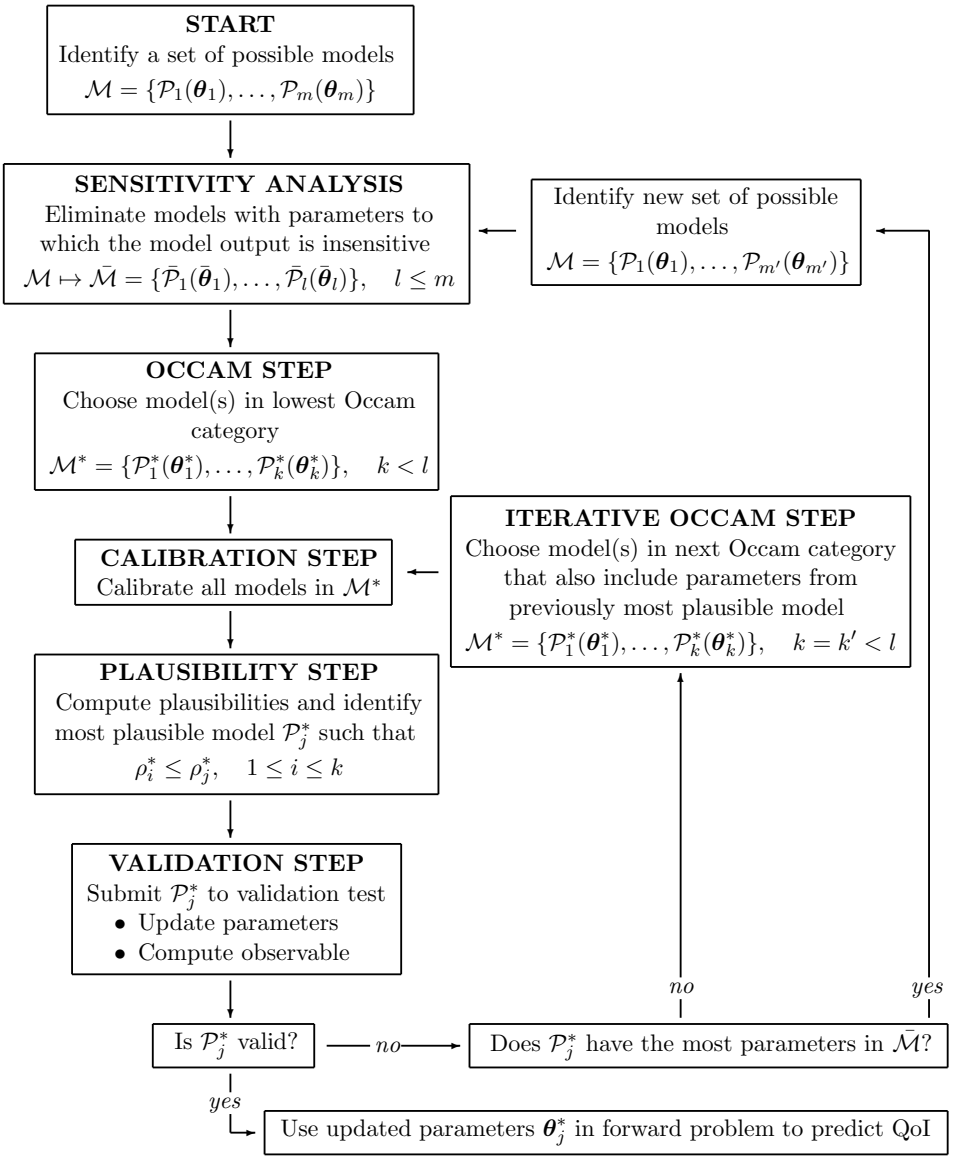


Figure 8.1. Flowchart for OPAL, the Occam Plausibility Algorithm (Farrell *et al.* 2015, Oden *et al.* 2018).

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

(7) *Iteration or solution.* If the model is deemed valid, for example, if (8.17) holds, then we use the parameters $\boldsymbol{\theta}^{**}$ of the validated model in the forward problem for the prediction scenario S_p to obtain the forward solution $(u(\boldsymbol{\theta}^{**}, S))$ and then compute the prediction of the QoI. If no model in category 1 is valid, we return to step (3) and set $\mathcal{M} = \bar{\mathcal{M}}_2$, the models of category 2, and repeat the steps outlined above until a valid model is established.

(8) The process is terminated when one model is identified that passes the validation criterion, even though other untested models may lead to a smaller error. If no models in the original set $\mathcal{M}$ are valid, the set of candidate models must be enlarged until (if possible) a set containing a valid model is obtained.

A flowchart presenting the steps in implementing OPAL is given in Figure 8.1.

Several remarks are in order. First, instead of the number of parameters, we could use other measures of 'simplicity' as a basic for establishing Occam categories, such as estimated computational complexity, operation counts, number of dependent variables, *etc.* Second, as a matter of computational efficiency, the calculation of plausibilities could be done for more complex validation scenarios when computationally feasible. Finally, criteria for model selection other than Bayesian plausibility could be used, such as the Akaike Information Criterion in a frequentist version of OPAL. Such alternatives are discussed in Farrell-Maupin and Oden (2017).

9. Adaptive control of multiscale models

9.1. Adaptive control of multiscale models

Multiscale modelling – the use of mathematical and computational models to simulate physical events that occur over several spatial and temporal scales – has been among topics of central importance in computational science for nearly two decades. There are many reasons for the emergence of this field. Among them are the remarkable advances in high-performance computing and parallel algorithms, making the use of complex multiscale models a possible undertaking. But equally important are the manifold applications in recent years that demanded the study and analysis of events at many scales. These include nanomanufacturing, semiconductor design and fabrication, analysis of numerous biomedical systems, drug design, geophysical systems, climate modelling, and many other areas. More recently the study of multiscale phenomena in the presence of uncertainties has emerged as a challenging area of research. Some approaches to stochastic multiscale models are taken up in the present article.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

There is a large body of literature on multiscale methods. We have mentioned the survey article of Liu, Karpov, Zhang and Park (2004), Ming and Zhang (2007), Curtin and Miller (2003), and the monograph of E (2011). A verity of different approaches to capture multiscale events has been proposed. Many fell into the category of *hierarchical methods*, in which the results of a finer-scale simulation are used as data for solving a coarser-scale model. The so-called *concurrent methods* may simultaneously solve a finer-scale model in some local region of the solution domain containing features of interest, and resolve coarser-scale models in complementary regions of the domain. Among methods in these category, we mention the heterogeneous multiscale method (HMM: E and Engquist 2003), the bridging-scale methods (BSM: Wagner and Liu 2003, Park, Karpov, Klein and Liu 2005), the quasi-continuum method (QCM: Tadmor, Ortiz and Phillips 1996), the coupled atomistic and discrete dislocation method (CADD: Shilkrot, Miller and Curtin 2004) and the goal-oriented adaptive control methods (Oden, Prudhomme, Romkes and Bauman 2006, Oden and Prudhomme 2001, Oden and Prudhomme 2002, Oden, Prudhomme and Bauman 2005, Oden and Vemaganti 2000, Oden, Vemaganti and Moës 1999, Vemaganti and Oden 2001), which belong to the general category of concurrent methods. Still other work on these subjects can be found in the articles mentioned here.

In this part of the exposition, we describe a general approach to the adaptive control of multiscale models built around *a posteriori* estimates of modelling error in key quantities of interest. The account builds on the developments given in Vemaganti and Oden (2001) and Oden and Vemaganti (2000) but extends them to highly complex stochastic systems.

9.2. A posteriori estimation of modelling error

One approach to multiscale modelling is to estimate and control modelling error in key quantities of interest – the error being a measure of the difference between QoIs at coarser scales and those delivered at the finer scale. The presumption is that a general mathematical model of a physical system can be defined, representing events at the finest (smallest) spatial and temporal scales, and that the model can serve as a datum with respect to which other models of events at coarser (larger) scales can be compared. The fine-scale or base model can be highly complex, intractable or very expensive to resolve, and even involve incomplete data, and it is never (or rarely) 'solved'. Contemporary literature refers to such reference models as 'high-fidelity' models.

Two aspects of this approach should be noted. First, it is clear that for the high-fidelity base model to be regarded a valid model for the QoIs, its parameters must be known, at least statistically. At first glance, this is paradoxical, as the forward problem for the base model cannot be solved. We

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

confront this problem via the analysis of high-dimensional inverse problems, assuming observational data are available, and we reduce many complications through the use of common assumptions and approximations in use for such analyses (*e.g.* Gaussian likelihood, *etc.*). This follows almost exactly the example of Bayesian inversion described in Section 5.2. Second, the theory for treating multiple scales when QoIs at one scale are correlated with those of another, is, we believe, incomplete and its development remains a challenging area of research. We discuss one such modelling paradigm involving multiscale models of cancer growth in Section 10.5.

While the idea of estimating modelling error and adaptive modelling appeared in the late 1990s in the papers of Oden *et al.* (1999) and Oden and Vemaganti (2000), a more general approach to estimation of modelling error for nonlinear problems in mechanics was presented by Oden and Prudhomme (2002) employing a framework similar to that of Becker and Rannacher (2001) in their study of optimal control to *a posteriori* error estimation in finite element methods. Generalizations of these methods were discussed in Oden *et al.* (2006) and applied to broad classes of problems in molecular dynamics and composite materials. Braack and Ern (2003) presented methods for estimation and control of both modelling and discretization errors. In more recent work, Maier and Rannacher (2016, 2018) described a post-processing procedure for selecting optimal models for estimating error in quantities of interest. Extensions of the adaptive modelling approaches in Oden and Vemaganti (2000) and Vemaganti and Oden (2001) to stochastic systems encountered in the modelling of random heterogeneous materials are discussed in Section 9.5.

Returning to the original abstract problem (2.1), we drop references to the scenario S and reframe it in a function-space setting. Beginning with a high level of abstraction and generality, we consider the forward problem, now written compactly as follows:

$$\begin{aligned} &\text{Find } u \in U \text{ such that} \\ &\mathcal{A}(\boldsymbol{\theta})u = F \quad \text{in } V', \end{aligned} \tag{9.1}$$

where $\mathcal{A}(\boldsymbol{\theta})$ is a parametric family of maps from a topological vector space (TVS) U into the dual space V' of a TVS V, and F is given in V'; $\boldsymbol{\theta} \in \Theta$ is a finite vector of parameters. Let $\langle \cdot, \cdot \rangle \equiv \langle \cdot, \cdot \rangle_{V' \times V}$ denote duality pairing on $V' \times V$ and let $\mathcal{B}(\cdot; \cdot)$ denote the semilinear form

$$\mathcal{B}(\boldsymbol{\theta}, u; v) = \langle \mathcal{A}(\boldsymbol{\theta})u, v \rangle, \tag{9.2}$$

the semicolon indicating that $\mathcal{B}(\boldsymbol{\theta}; \cdot; \cdot)$ may be nonlinear in arguments u to the left of the semicolon, but linear in v, and the linear form F is given by

$$F(v) = \langle F, v \rangle, \quad \text{for } v \in V. \tag{9.3}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

Then problem (9.1) assumes the following variational form:

$$\begin{aligned} &\text{For any } \boldsymbol{\theta} \in \Theta, \text{ find } u = u(\boldsymbol{\theta}) \in U \text{ such that} \\ &\mathcal{B}(\boldsymbol{\theta}, u; v) = F(v) \quad \text{for all } v \in V. \end{aligned} \tag{9.4}$$

The goal of the model problem is to use the solution u of the forward problem to compute the target QoIs, defined by functionals Q on U:

$$Q : U \to \mathbb{R}, \quad \text{QoI} = Q(u(\boldsymbol{\theta})). \tag{9.5}$$

If (9.1) defines an intractable high-fidelity model, so must (9.5) be intractable. Thus, we must seek a computable approximation to $Q(u)$.

Before introducing such an approximation, we list a few fundamental points derived in Oden and Prudhomme (2002) and Oden *et al.* (2006).

(1) *Optimal control of models to deliver a given QoI.* Given $Q : U \to \mathbb{R}$, $\boldsymbol{\theta} \in \Theta$ and the general problem (9.4),

$$Q(u) = \inf_{v \in M} Q(v); M = \{z \in V : \mathcal{B}(\boldsymbol{\theta}, z; q) = F(q) \text{ for all } q \in V\}. \tag{9.6}$$

This is the optimal control problem for goal-oriented estimation described in Oden and Prudhomme (2002) and Becker and Rannacher (2001).

(2) *Solution.* Using standard Lagrangian multipliers for the constant $v \in M$, we arrive at the following pair of problems:

$$\begin{aligned} &\text{Find } (u, w) \in U \times U \text{ such that} \\ &\mathcal{B}(\boldsymbol{\theta}, u; v) = F(v) \qquad \text{for all } v \in V \quad \text{(the primal problem)}, \\ &\mathcal{B}'(\boldsymbol{\theta}, u; w, v) = Q'(u; v) \quad \text{for all } v \in V \quad \text{(the adjoint problem)}, \end{aligned} \tag{9.7}$$

where primes ($'$) indicate Gâteaux derivatives, that is,

$$\mathcal{B}'(\boldsymbol{\theta}, u; w, v) = \lim_{\epsilon \to 0} \quad \varepsilon^{-1}[\mathcal{B}(\boldsymbol{\theta}, u + \varepsilon w, v) - \mathcal{B}(\boldsymbol{\theta}, u; v)], \tag{9.8}$$

$$Q'(u; v) = \lim_{\epsilon \to 0} \quad \varepsilon^{-1}[Q(u + \varepsilon v) - Q(u)]. \tag{9.9}$$

(3) *General form of the error in Q.* The following is proved in Oden and Prudhomme (2002).

Theorem 9.1. Let the semilinear form $\mathcal{B}(\boldsymbol{\theta}, \cdot, \cdot)$ of (9.2) and the QoI $Q(\cdot)$ have functional (Gâteaux) derivatives of order three. Let $(u, w) \in U \times U$ be a solution of (9.7). Let $(\bar{u}, \bar{w})$ be an arbitrary pair in $U \times U$. Then the error produced by $\bar{u}$ in the QoI is given by

$$Q(u) - Q(\bar{u}) = R(\bar{u}; \bar{w}) + R(\bar{u}, e_0) + \Delta(\bar{u}, \bar{w}, e_0, \varepsilon_0), \tag{9.10}$$

where $R(\bar{u}, v)$ is the residual functional

$$R(\bar{u}, v) = F(v) - \mathcal{B}(\boldsymbol{\theta}, \bar{u}; v), \quad v \in V, \tag{9.11}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

e_0 and ε_0 are the errors

$$e_0 = u - \bar{u}, \quad \varepsilon_0 = w - \bar{w}, \tag{9.12}$$

and $\Delta(\bar{u}, \bar{w}, e_0, \varepsilon_0)$ is a remainder depending on second and third Gâteaux derivatives of $\mathcal{B}(\boldsymbol{\theta}, \cdot; \cdot)$ and $Q(\cdot)$:

$$\begin{aligned}\Delta = \frac{1}{2}\int_0^1 [&\mathcal{B}''(\boldsymbol{\theta}, \bar{u} + se_0; e_0, e_0, \bar{w} + S\varepsilon_0) \\ &- Q''(\bar{u} + se_0; e_0, e_0) + \{Q'''(\bar{u} + se_0; e_0, e_0) \\ &- 3\mathcal{B}''(\boldsymbol{\theta}, \bar{u} + se_0; e_0, e_0, \varepsilon_0) \\ &- \mathcal{B}'''(\boldsymbol{\theta}, \bar{u} + se_0; e_0, e_0, e_0, \bar{w} + s\varepsilon_0)\}(1 - s)]\, \mathrm{d}s,\end{aligned} \tag{9.13}$$

the primes ($''$,$'''$) denoting second and third Gâteaux derivatives. See Oden and Prudhomme (2002) and Oden *et al.* (2006) for details, and the related work by Becker and Rannacher (2001).

Remarkably, $(\bar{u}, \bar{w})$ is an arbitrary pair in $U \times U$. Thus Δ can conceivably be very large. If $\mathcal{B}(\boldsymbol{\theta}, \cdot; \cdot)$ is bilinear for all $\boldsymbol{\theta}$, and $Q(\cdot)$ is linear, Δ is zero and (9.11) collapses to

$$Q(u) - Q(\bar{u}) = R(\bar{u}; \bar{p}) + \mathcal{B}(\boldsymbol{\theta}, e_0; \varepsilon_0). \tag{9.14}$$

(4) *Surrogate problem.* Since $(\bar{u}, \bar{w})$ in (9.10)–(9.14) are completely arbitrary vector pair in $V \times V$, a natural choice is to structure approximations to $\mathcal{B}(\boldsymbol{\theta}, \cdot; \cdot)$, $F(\cdot)$, and $Q(\cdot)$ to minimize the remainder Δ and to replace (9.2) (and (9.7_2)) by tractable problems. If $\mathcal{B}_0(\boldsymbol{\theta}, \cdot; \cdot)$, $F_0(\cdot)$ and $Q_0(\cdot)$ denote their respective approximations, then the surrogate problem takes the following form:

$$\begin{aligned}&\text{Find } (u_0, w_0) \in U \times U \text{ such that} \\ &\mathcal{B}_0(\boldsymbol{\theta}, u_0; v) = F_0(v) \qquad \text{for all } v \in V, \\ &\mathcal{B}_0'(\boldsymbol{\theta}, u; w, v) = Q_0(u; v) \quad \text{for all } v \in V.\end{aligned} \tag{9.15}$$

Clearly the surrogate forms in (9.15) are expected to be drawn from a coarser-scale problem than (9.7) (or (9.1)), and are generally constructed through homogenization or averaging processes. If (9.15) is 'close' to (9.7) or if $\mathcal{B}(\boldsymbol{\theta}, \cdot, \cdot)$ is bilinear and $Q(\cdot)$ is linear, the remainder Δ can be neglected (or is zero), and the error in the QoI is

$$Q(u) - Q(\bar{u}) = R(u_0; w_0) + R(u_0; \varepsilon_0) = R(u_0; w). \tag{9.16}$$

The Goals algorithm, described in, for example, Vemaganti and Oden (2001) or Oden *et al.* (2006), naturally involves a sequence of surrogate models of the form (9.15) designed to reduce the error to acceptable tolerance sequentially as more information from the high-fidelity model is accessed.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

9.3. Adaptive control of models: the Goals algorithm

Once an estimate of error in QoI inherent in each member of a sequence of surrogate models is in hand, several algorithms can be implemented to systematically reduce modelling error until a predefined error tolerance is reached. One such algorithm is the straightforward scheme embedded in the Goals algorithm, introduced in Oden and Vemaganti (2000) and described in more detail in Oden *et al.* (2006). Basic steps are listed as follows.

(1) For specificity, let $D \subset \mathbb{R}^d$ be an open bounded regime which is the domain of the high-fidelity solution $(u, w) \in U \times U$, and let $\{G_j\}_{j=1}^K$ be a partition of

$$D : \bar{D} = \bigcup_j^K G_j, \quad G_i \cap G_j = \emptyset, i \neq j, \quad 1 \leq i, j \leq K.$$

The semilinear and linear forms in (9.15) can be broken into components over the partition so that

$$\mathcal{B}(\boldsymbol{\theta}, u; v) = \sum_{j=1}^K \mathcal{B}_{G_j}(\boldsymbol{\theta}, u; v), \quad \mathcal{F}(v) = \sum_{j=1}^K F_{G_j}(v) \tag{9.17}$$

for $u, v \in V$, $\mathcal{B}_{G_j}(\boldsymbol{\theta}, \cdot; \cdot)$ and $F_{G_j}(\cdot)$ being restrictions to subdomain G_j.

We suppose that the QoI is confined to a subdomain $D_Q \subset D$, which is to be called the *domain of influence*. Generally D_Q can be the subdomains G_1 of the partitions or a union of subdomains entering the QoI.

(2) Through an averaging or homogenization process, we construct the surrogate problem (9.15) and solve it for the surrogate solution $(u_0^{(1)}, w_0^{(1)}) \in V \times V$ defined over all D.

(3) The modelling error in the surrogate is estimated using (9.16):

$$e_0^{(1)} \approx Q(u) - Q(u_0^{(1)}) \approx R(u_0^{(1)}; w_0^{(1)}) + R(u_0^{(1)}; w - w_0). \tag{9.18}$$

The first term in (9.18) is computable, the second is not because w is unknown and must be approximated. Such approximations are usually problem-dependent; examples are given in a wide range of applications in Oden and Vemaganti (2000), Oden *et al.* (2005) and Oden *et al.* (2006). Alternatively, for bilinear $\mathcal{B}(\boldsymbol{\theta}, \cdot; \cdot)$ and linear $Q(\cdot)$, rigorous two-sided bounds of *a posteriori* estimates of error in QoIs can be established. These are reviewed in Section 9.4.

(4) An error tolerance γ_{tol} is introduced to define the maximum error admitted in the adaptive process. Obviously, if $\|e_0^{(1)}\| \leq \gamma_{tol}$, the procedure is terminated and the QoI is predicted with accepted accuracy. If $\|e_0^{(1)}\| > \gamma_{tol}$,

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

the accuracy of the surrogate model must be improved. The corrected surrogate ($k = 2$) involves the form

$$\begin{aligned}\mathcal{B}_0^{(2)}(\boldsymbol{\theta}, u_0^{(2)}, v) &= \mathcal{B}_0^{(2)}(\boldsymbol{\theta}, u_{0,D_Q}^{(2)}; v_{D_Q}) + \mathcal{B}_{D\setminus D_Q}(\boldsymbol{\theta}, u_{0,D\setminus D_Q}^{(2)}; v_{D\setminus D_Q}) \\ &= F(v) \quad \text{for all } v \in V, \\ \mathcal{B}_0^{(2)'}(\boldsymbol{\theta}, u_0^{(2)}; w_0^{(2)}, v) &= Q'(u_0^{(2)}; v), \end{aligned} \tag{9.19}$$

where $u_{0,D_Q}^{(2)}$ is the restriction of the second iterate solution $u_0^{(2)}$ to the domain of influence, and $\mathcal{B}_{D\setminus D_Q}(\boldsymbol{\theta}, \cdot; \cdot)$ is the semilinear form defined on restrictions of functions in V to $D\setminus D_Q$. In other words, fine-scale features of the high-fidelity model are included in D_Q while coarse-scale (homogenized) features exist over $D \setminus D_Q$.

(5) Compute $\|e_0^{(2)}\| \approx \|Q(u) - Q(u_0^{(2)})\|$. If $\|e_0^{(2)}\| \le \gamma_{tol}$, stop; if not, redo (9.19) to compute $(u_0^{(3)}, w_0^{(3)})$ and continue adding fine-scale features until the tolerance is met.

One algorithm for the sequential addition of subdomains G_j containing fine-scale information is the 'greedy algorithm' described in Vemaganti and Oden (2001). Given an approximation of the adjoint solution w, the residual $R^{(k)}$ at iterate k is given by (9.18) as $R^{(k)} = R(u_0^{(k)}, w)$. The contribution to $R^{(k)}$ from subdomain G_j is denoted $R_j^{(k)}$, that is,

$$R^{(k)} = \sum_{j=1}^{(k)} R_j^{(k)}.$$

Let $R_{\max}^{(k)} = \max_{1\le j\le K} R_j^{(k)}$. Then if $R_j^{(k)} \ge \alpha R_{\max}^{(k)}$, $\alpha \in (0,1)$, we add fine-scale features of G_j to the fine-scale subdomain and compute $R^{(k+1)}$. This process is continued until the tolerance γ_{tol} is reached.

9.4. Two-sided bounds of errors in QoIs

The two-sided (upper and lower) bounds on estimated error in quantities of interest described in Oden and Vemaganti (2000) can be extended to linear stochastic elliptic boundary-value problems modelling diffusion-type phenomena in random heterogeneous media. There the problem is as follows:

$$\begin{aligned}&\text{Find } u \in H = L^2_{\mathbb{P}}(\Omega) \otimes H^1_{\Gamma_D} \text{ such that} \\ &\mathbb{E}\left[\int_D \boldsymbol{A}\boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\boldsymbol{x}\right] = \mathbb{E}\left[\int_D f v \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_N} \sigma v \,\mathrm{d}s\right], \end{aligned} \tag{9.20}$$

for all $v \in H$, where Ω is the sample set of outcomes in a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{F}$ being a σ-algebra of subsets of Ω and $\mathbb{P}$ the probability

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

measure, H_{Γ_D} is the Hilbert space

$$H_{\Gamma_D} = \{v \in H^1(D) : \text{trace(v)}|_{\Gamma_D} = 0\},$$

$\boldsymbol{A}$ is a $d \times d$ symmetric positive-definite matrix

$$\boldsymbol{A} \in (L^\infty(D))^{d\times d}, \quad f^D \in L^2(\Omega, 0), \quad \sigma \in H^{-1/2}(\Omega, \Gamma_N),$$

and $\mathbb{E}[\cdot]$ is the expected value

$$\mathbb{E}[Y] = \int_\Omega Y(\omega)\, \mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{R}^N} y\pi(y)\, \mathrm{d}y,$$

$\pi(y)$ being the probability density associated with $\mathbb{P}$. Denoting the left-hand side of (9.20) by $\mathcal{B}(u, v)$ and the right-hand side by $F(v)$, the problem assumes the following classical form:

$$\begin{aligned} &\text{Find } u \in H \text{ such that} \\ &\mathcal{B}(u, v) = F(v) \quad \text{for all } v \in H. \end{aligned} \tag{9.21}$$

For each QoI $Q \in \mathcal{L}(H, \mathbb{R})$ we also have the following adjoint problem:

$$\begin{aligned} &\text{Given } Q : H \to \mathbb{R}, \text{ find } w \in H \text{ such that} \\ &\mathcal{B}(u, w) = Q(v) \quad \text{for all } v \in H. \end{aligned} \tag{9.22}$$

Problems (9.20) and (9.21) are equivalent to the stochastic boundary-value problem of finding a scalar field $u : \Omega \times D \to \mathbb{R}$ such that, almost surely for all $\omega \in \Omega$,

$$\begin{aligned} \boldsymbol{\nabla} \cdot \boldsymbol{A}\boldsymbol{\nabla} u(\omega, \boldsymbol{x}) &= f(\omega, \boldsymbol{x}), \quad \boldsymbol{x} \in D, \\ \boldsymbol{A}\boldsymbol{\nabla} u(\omega, \boldsymbol{x}) &= \sigma(\omega, \boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma_N, \\ u(\omega, \boldsymbol{x}) &= 0, \qquad \boldsymbol{x} \in \Gamma_D, \end{aligned} \tag{9.23}$$

where $\boldsymbol{\nabla}$ is the spatial gradient, f and σ are prescribed data, and $\partial D = \Gamma_N \cup \Gamma_D$, Neumann data σ being prescribed on $\partial D \setminus \Gamma_D$. Assuming there exist positive constants $\alpha_{\min}$, $\alpha_{\max}$ such that

$$\mathbb{P}\{\omega \in \Omega : \alpha_{\min}\boldsymbol{a}^T\boldsymbol{a} \le \boldsymbol{a}^T\boldsymbol{A}(\omega, \boldsymbol{x})\boldsymbol{a} \le \alpha_{\max}\boldsymbol{a}^T\boldsymbol{a}, \text{ for all } \boldsymbol{a} \in \boldsymbol{R}^d\} = 1, \tag{9.24}$$

problem (9.20) has a unique solution $u \in H$ by the Lax–Milgram theorem.

Following the arguments laid down earlier for models such as (9.21), the high-fidelity problem (9.20) (and (9.21)) is generally intractable, and a tractable sequence of surrogates is needed. We thus follow the plan described earlier and introduce a sequence of hybrid models in which subdomain G_j of D contains fine-scale features while homogenized (possibly deterministic) properties exist in the complementary domain in D. For the kth surrogate, we have

$$\boldsymbol{A}_0^{(k)}(\omega, \boldsymbol{x}) = \begin{cases} \boldsymbol{A}(\omega, \boldsymbol{x}), & \boldsymbol{x} \in D_k, \\ \boldsymbol{A}_{00}(\boldsymbol{x}), & \boldsymbol{x} \in D \setminus D_k, \ k = 1, 2, \ldots, K, \end{cases} \tag{9.25}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

D_k being the union of fine-scale subdomains. The sequence of surrogate forward and adjoint problems is then characterized as follows:

$$\begin{aligned} &\text{Find } (u_0^{(k)}, w_0^{(k)}) \in H \times H \text{ such that} \\ &\mathcal{B}_0(u_0^{(k)}; v) = F_0(v) \quad \text{for all } v \in H, \\ &\mathcal{B}_0'(v, w_0^{(k)}) = Q_0(v) \quad \text{for all } v \in H, \end{aligned} \tag{9.26}$$

$k = 1, 2, \ldots, K$, where

$$\mathcal{B}_0(u_0^{(k)}; v) = \mathbb{E}\left[\int_D \boldsymbol{A}_0^k \boldsymbol{\nabla} u_o^{(k)} \cdot \boldsymbol{\nabla} v \, \mathrm{d}\boldsymbol{x}\right]. \tag{9.27}$$

A posteriori estimates of errors in $Q(u_0^{(k)})$ are given in the following theorem.

Theorem 9.2. Let u_0 and w_0 be solutions of the surrogate problem (9.15) (suppressing index (k)). Then the modelling error in the QoI is bounded above and below according to

$$\eta_{low} \leq Q(u) - Q(u_0) \leq \eta_{upp}, \tag{9.28}$$

where u is the solution of the forward problem (9.21),

$$\begin{aligned} \eta_{low} &= \frac{1}{4}(\eta_{low}^+)^2 - \frac{1}{4}(\eta_{upp}^-)^2 + R_{u_0}(w_0), \\ \eta_{upp} &= \frac{1}{4}(\eta_{upp}^+)^2 - \frac{1}{4}(\eta_{low}^-)^2 + R_{u_0}(w_0), \end{aligned} \tag{9.29}$$

with

$$(\eta_{upp}^{\pm})^2 = s^2 \|u_0\|_{\boldsymbol{I}_0\mathcal{B}}^2 + s^{-2}\|w_0\|_{\boldsymbol{I}_0\mathcal{B}}^2 \pm \mathbb{E}[2(u_0, w_0)_{\boldsymbol{I}_0\mathcal{B}}], \tag{9.30}$$

$$(\eta_{low}^{\pm})^2 = \frac{|R_{su_0 \pm s^{-1}w_0}(u_0 + \gamma^{\pm} w_0)|}{\|u_0 + \gamma^{\pm} w_0\|_{\mathcal{B}}}, \tag{9.31}$$

$$R_g(v) = -\mathbb{E}\left[\int_D \boldsymbol{A}\boldsymbol{I}_0 \boldsymbol{\nabla} g \cdot \boldsymbol{\nabla} v \, \mathrm{d}\boldsymbol{x}\right], \tag{9.32}$$

and the inner product and norms are given for $h, g \in H$, by

$$(h, g)_{\boldsymbol{I}_0\mathcal{B}} = \int_D \boldsymbol{A}\boldsymbol{I}_0 \boldsymbol{\nabla} h \cdot \boldsymbol{\nabla} g \, \mathrm{d}\boldsymbol{x}, \tag{9.33}$$

$$\|h\|_{\boldsymbol{I}_0\mathcal{B}}^2 = (h, h)_{\boldsymbol{I}_0\mathcal{B}}, \tag{9.34}$$

$\gamma^{\pm}$ are real numbers and $\boldsymbol{I}_0$ is the deviation matrix

$$\boldsymbol{I}_0 = \boldsymbol{I}_0(\omega, \boldsymbol{x}) = \boldsymbol{I} - \boldsymbol{A}^{-1}(\omega, \boldsymbol{x})\boldsymbol{A}(\omega, \boldsymbol{x}), \tag{9.35}$$

$\boldsymbol{I}$ being the $d \times d$ identity matrix, and $s \in \mathbb{R}^+$is a scaling factor (optimal if $s^2 = \|e_0\|_{\mathcal{B}}/\|\varepsilon_0\|_{\mathcal{B}}$).

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

9.5. Example: adaptive modelling of diffusion in a random heterogeneous medium

An example of adaptive modelling of a stochastic boundary-value problem using a version of the Goals algorithm described earlier is outlined below. It concerns heat conduction in a random two-phase media described by (9.21) for the case

$$\boldsymbol{A}(\omega, \boldsymbol{x}) = \kappa(\omega, \boldsymbol{x})\boldsymbol{I}, \quad \text{for } \omega \in \Omega,\ \boldsymbol{x} \in D, \tag{9.36}$$

where κ is the heat conduction, Ω is the set of two-phase material realizations. Generally, κ is of the form

$$\kappa(\omega, \boldsymbol{x}) = \kappa_I \mathcal{X}_I(\omega, \boldsymbol{x}) + \kappa_M \mathcal{X}_M(\omega, \boldsymbol{x}), \tag{9.37}$$

where κ_I and κ_M are conductivities of inclusions and matrix materials and $\mathcal{X}_I$, $\mathcal{X}_M$ are characteristic functions for subsets of the domain of volume $\mathcal{V}_I$ and $\mathcal{V}_M$ ($\mathcal{V} = |D|; \mathcal{X}_i(\omega, \boldsymbol{x}) = 1$ if $\boldsymbol{x} \in \mathcal{V}_i = 0$ if otherwise, $i = I, M.$). The volume fractions are $\phi_I = \mathcal{V}_I/\mathcal{V}$, $\phi_M = \mathcal{V}_M/\mathcal{V}$.

For the homogenized material, we take for the effective conductivity κ_0 given by

$$\kappa_0 = \frac{1}{2}(\kappa_+^e + \kappa_-^e), \tag{9.38}$$

where $\kappa_\pm^e$ are the Hashin–Shtrikman bounds (Hashin and Shtrikman 1963)

$$\begin{aligned} \kappa_e^- &= \kappa_2\phi_2 + \kappa_1\phi_1 - \frac{(\kappa_2 - \kappa_1)^2\phi_2\phi_1}{\kappa_2 + \kappa_2\phi_1 + \kappa_1\phi_2}, \\ \kappa_e^+ &= \kappa_1\phi_1 + \kappa_2\phi_2 - \frac{(\kappa_1 - \kappa_2)^2\phi_1\phi_2}{\kappa_1 + \kappa_1\phi_2 + \kappa_2\phi_1}, \end{aligned} \tag{9.39}$$

and the effective conductivity κ_e satisfies

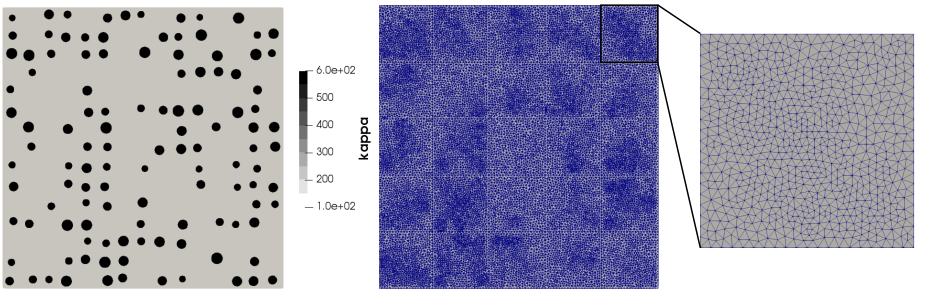
$$\kappa_-^e \le \kappa^e \le \kappa_+^e. \tag{9.40}$$

An example (deterministic) representing a single realization $\omega \in \Omega$ concerns a 1 m $\times$ 1 m square domain D with $u = 0$ on boundaries $x_1 = 0, x_1 = 1.0$ and heat flux $\sigma = 100$ W m^{-2} applied on $x_2 = 1.0$ and $\sigma = 0$ on $x_2 = 0$. The QoI is taken to be the average temperature over a circular domain D_q of radius 0.05 m centred at point $(0.9, 0.9)$:

$$Q(v) = |D_q|^{-1} \int_{D_q} v \, \mathrm{d}x, \quad v \in H^1(D).$$

The conductivities of the phases are $\kappa_I = 600$ W (mK)$^{-1}$ and $\kappa_M = 100$ W (mK)$^{-1}$. The domain D is partitioned with 25 regular square subdomains. One realization of the model is shown in Figure 9.1(a) with randomly distributed circular inclusions. The full problem is solved on a fine element

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press



(a) (b)

Figure 9.1. One realization of the randomly distributed circular inclusions in matrix form: (a) conductivity, (b) mesh.

Table 9.1. Results of error estimation in each adaptive steps in constructing the surrogate models. In all the levels, $\phi_I = 0.1020$ for the volume fraction of the inclusions.

k	$Q(u_0^{(k)})$	$\eta_{exact}^{(k)}$	$\eta_{low}^{(k)}$	$\eta_{upp}^{(k)}$	$\mathcal{E}_\eta$
0	0.0205	0.0572	−1.673	1.761	0.7687
1	0.038	0.0391	−3.52	3.61	1.15
2	0.0450	0.0328	−3.280	3.34	0.964
3	0.0500	0.02778	−2.006	2.059	0.963
4	0.062	0.016	−1.99	2.031	1.23
5	0.0642	0.0135	−1.75	1.786	1.38
6	0.0705	0.0070	−0.656	0.673	1.17

mesh of over 33 000 elements, giving a good approximation of the 'exact' QoI of $Q_{exact}(u) \approx 0.0778$.

Seven steps in the Goals algorithm are used to reduce the modelling error to below a tolerance of $\gamma_{tol} = 0.02 Q_{exact}$. Results are collected in Table 9.1. In this table, $\eta_{exact}^{(k)} = 0.0778 - Q(u_0^{(k)})$, and $\mathcal{E}_\eta$ is the effectivity index,

$$\mathcal{E}_\eta = \frac{\frac{1}{2}(\eta_{upp}^{(k)} + \eta_{low}^{(k)})}{\eta_{exact}^{(k)}}. \tag{9.41}$$

The sequence of adaptive error indicators is shown in Figure 9.2.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

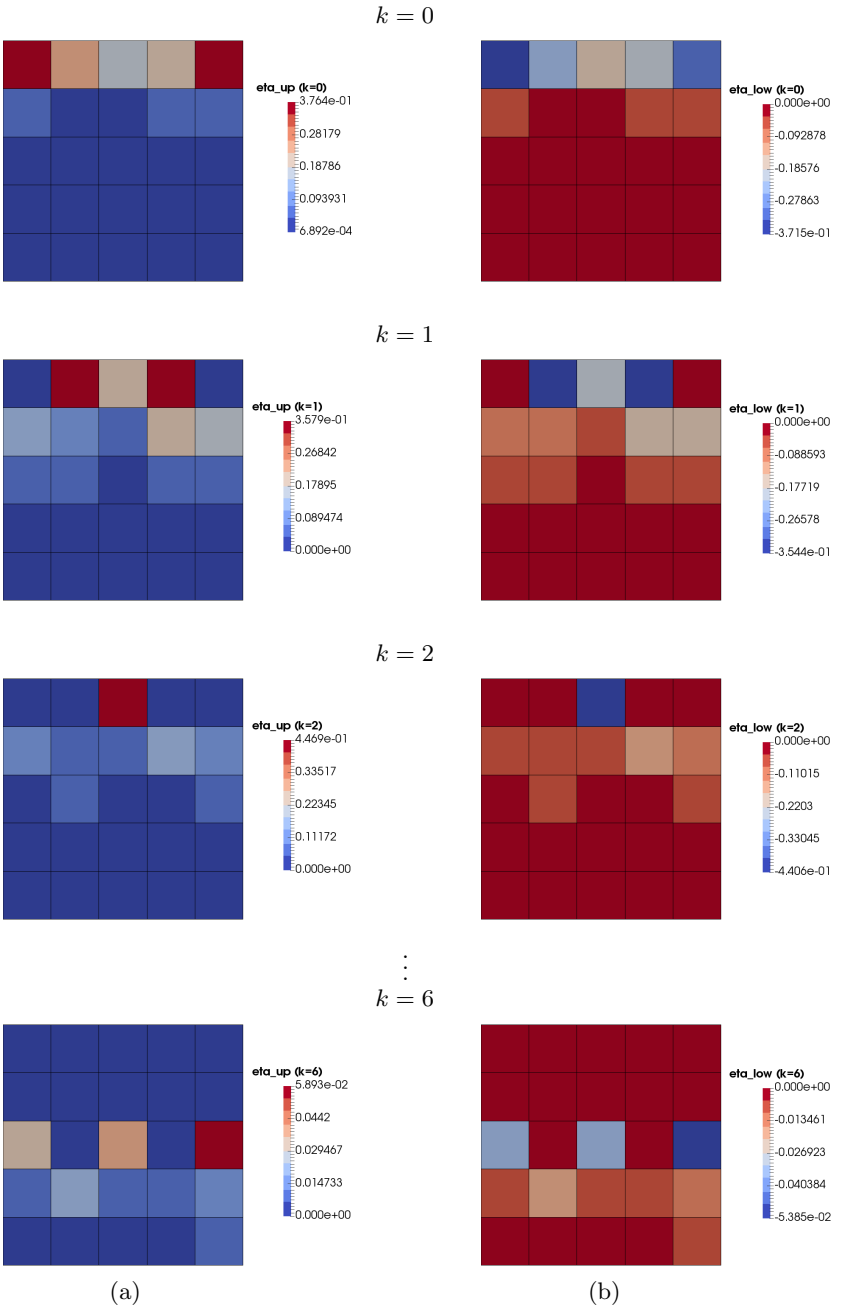


Figure 9.2. Patterns of error indicators over a 25-block partition of D and the finite element meshes employed in computing the surrogate solutions for levels $k = 0, 1, \ldots, 6$: (a) $\eta_{est}^{up(k)}$ at each block, (b) $\eta_{est}^{low(k)}$ at each block.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

10. Selection, calibration, and validation of families of stochastic parametric models: prediction of tumour growth

10.1. Selection versus adaptation

While the notion of adaptive modelling, particularly as described in the preceding sections, generally means the adaptation of the structure of a model to meet certain goals (such as accuracy in QoIs), it could also be interpreted as the adaptation of one's choice of a model to be consistent with observational data. While not a common idea in works on adaptive modelling, the selection and validation of models are perhaps the most important components in all of computational science. Both adaptivity and model selection and validation share two traits – the subjective choice of criteria and tolerances for accepting one model over another – and they both aim at defining acceptable mathematical models for scientific predictions.

The algorithm OPAL, described in Section 8.5, is designed to provide one framework for selecting, calibrating and validating sets of parametric model classes. We describe an implementation of OPAL in this section drawn from work on modelling tumour growth in living tissue (*e.g.* Lima, Oden and Almeida 2014, Lima *et al.* 2016, Lima *et al.* 2017).

10.2. A canonical example: classes of phenomenological models of tumour growth

One of the most challenging areas in contemporary predictive computational science is the modelling and prediction of the growth of tumours in living organisms. It is field in which there is no general agreement on which models of this complex biophysical phenomena are adequate for prediction, in which reliable observational data are sparse and incomplete, and in which little data are available on the reliability of proposed models. Beyond these obstacles, the numerical solution and analysis of most tumour models is formidable and in early stages of development. We will, nevertheless, summarize an application of OPAL and the methods of model selection to one class of tumour growth models, following, in particular, Lima *et al.* (2016, 2017).

10.2.1. The heterogeneous micro-cellular environment of tumours

An abstraction of the complex microenvironment of tumour cells is depicted in Figure 10.1, where we observe a colony of tumour cells in various stages of growth or decline, healthy cells, and other relevant constituents. The structural framework supporting this micro-cellular environment is called the *stroma*, and it is made up of collagen together with a collection of extra-cellular molecules called the *extra-cellular matrix* (ECM). Cells called

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

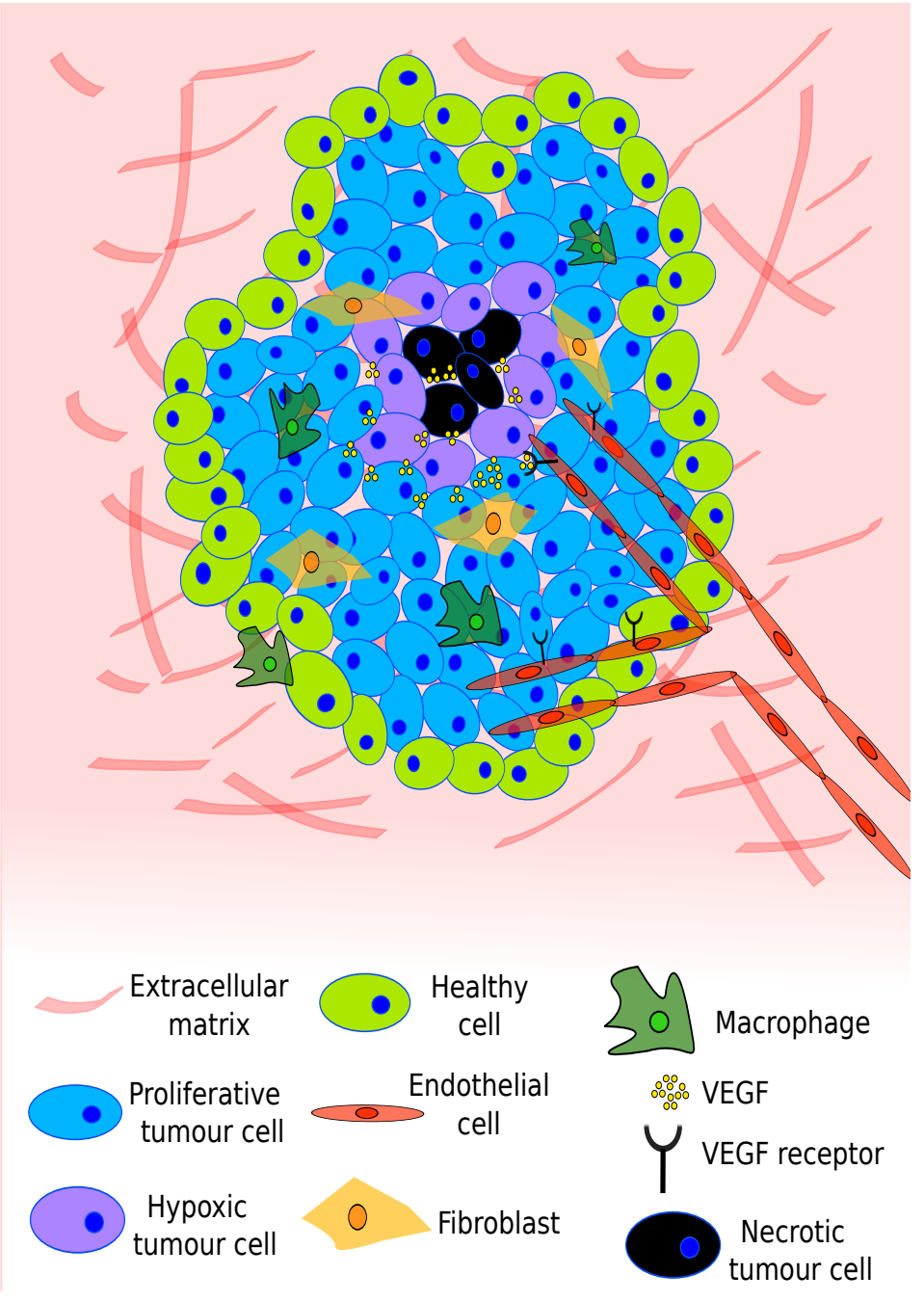


Figure 10.1. A schematic figure of tumour environment and different involved in modelling tumour growth. (From Oden 2017.)

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

fibroblasts synthesize the collagen and ECM and make up the main connective tissue cells present in the organism. Also present are *macrophages*, large white blood cells that engulf and digest cellular debris. *Endothelial cells* form the lining of blood vessels and lymphatic vessels.

We recall that cancer manifests itself in a series of stages in which genetic mutations of cells causes over-expression of oncogenes, the genes causing proliferation of tumour cells, or under-expression of mechanisms that control or suppress cell growth. Thereby, these genes alter homeostasis (the processes that preserve the architecture of the body) and lead to the creation of colony of aberrant cells. This mutation process may take years to progress, or can be accelerated by other factors in the microenvironment. Oncogenes are activated by subcell growth signals that promote cell proliferation or halt the cell cycle.

At the tissue scale, tumour cells that depend on oxygen and other nutrients to survive may enter the so-called *hypoxic state* between a state of proliferation, in which the possibility of mitosis occurs, and a state of unplanned cell death, called *necrosis*. In the necrotic state, water enters the cell, increasing pressure until it bursts, and the cell undergoes calcification going into a dormant state of micro-calcification. This cell death is different from apoptosis, programmed cell death, that occurs when the cells are no longer needed to preserve the latent architecture of the tissue or when they pose a threat to the organism. Thus, a colony of tumour cells can contain proliferative cells which can grow in number due to mitosis, hypoxic cells which are in essentially a suspended state that depends on the local nutrient and oxygen supply, and the necrotic cells.

The general tumour growth scenario is as follows. A mass of tumour cells establishes itself in its host tissue surrounded by its microenvironment, beginning with no vascular structure of its own. The survival of this collection of cells depends on its neighbouring cell structure, the *stroma*, for a supply of nutrient and growth factors to stimulate its growth. Without sufficient nutrients, an inner core of hypoxic cells forms and eventually some die, creating a necrotic core of dead cells. The outer cells, those with sufficient nutrients and growth factor concentrations, proliferate and cause an outward flux of cells that expands the tumour, causing mechanical compression of surrounding tissue and the base membrane that supports the tumour mass. The proliferating cells also absorb interstitial fluid to promote mitosis and growth of the mass of the tumour. The growing tumour interacts with the ECM, the surrounding environment, which is degraded and modified by the secretion of special enzymes. The ECM, in turn, releases growth factors that refuel tumour growth.

If insufficient nutrients are available to promote proliferation, hypoxia occurs which, remarkably, results in the release of proteins collectively called *vascular endothelial growth factors* (VEGF), which are among a group of

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

proteins labelled *tumour angiogenesis factors* (TAFs). These TAFs diffuse outward from the nutrient-deprived hypoxic cells and invade regions outside the tumour, eventually reaching nearby blood vessels, coming into contact with endothelial cells that form the walls of the blood vessels. When these endothelial cells detect a gradient in the TAF they secrete special enzymes such as *matrix metalloproteinases* (MMPs) that degrade the ECM and the base membrane allowing the endothelial cells to migrate away from the parent blood vessel toward the TAF source through *sprouts* – fingers of blood vessels that gradually grow toward the hypoxic cell mass in microvascular tubes that, on contact with the hypoxic cells, supply the needed nutrients for cell proliferation. The hypoxic cells that are supplied with these nutrients are thus transformed into proliferative cells. This remarkable and complex process is called angiogenesis.

There are many accounts of these biological processes affecting cancer in the literature. The summary here follows Reynolds and Schecker (1995) and Macklin, Cristini and Lowengrub (2010).

10.3. A general class of tumour models

The following system of nonlinear evaluation equations is derived (Lima *et al.* 2016, Lima, Oden and Almeida 2014, Oden *et al.* 2016) using a combination of continuum mixture theory, under the assumption that the balance laws of physics must hold (balance of mass, momentum, energy, and the second law of thermodynamic) and that the basic 'hallmarks of cancer' (Hanahan and Weinberg 2011, Hanahan and Weinberg 2000) serve as guidance for modelling the cancer-cell biology described above. The principal dependent variables in the model classes are scalar-valued functions $\phi_\alpha : \bar{D} \times [0, T_{end}]$ $\phi_\alpha = \phi_\alpha(\boldsymbol{x}, t)$, defined over a spatial domain $D \subset \mathbb{R}^3$, $\alpha = 1, 2, \ldots, 6$, and μ. We have

$$\begin{aligned}
&\phi_1 = \phi_T = \text{VF (volume fraction) of tumour cells,}\\
&\phi_2 = \phi_H = \text{VF of hypoxic tumour cells,}\\
&\phi_3 = \phi_N = \text{VF of necrotic tumour cells,}\\
&\phi_4 = \phi_\sigma = \text{VF of nutrient,}\\
&\phi_5 = \phi_{TAF} = \text{VF of tumour angiogenesis factors,}\\
&\phi_6 = \phi_e = \text{VF of non-activated endothelial cells,}\\
&\mu,\mu_e = \text{chemical potentials.}
\end{aligned}$$

In general,

$$\phi_T = \phi_P + \phi_T + \phi_N, \tag{10.1}$$

where ϕ_P is the volume fraction of proliferative cells. The general model is

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

characterized by the system of partial differential equations

$$\begin{aligned}
\frac{\partial \phi_T}{\partial t} &= \boldsymbol{\nabla} \cdot \mathcal{M}(\phi_T, \phi_H, \phi_N) \boldsymbol{\nabla} \mu + S_T, \\
\mu &= \Psi'(\phi_T) - \epsilon_T^2 \Delta \phi_T - \chi_0 \phi_\sigma, \\
\frac{\partial \phi_H}{\partial t} &= \boldsymbol{\nabla} \cdot \bar{M}_H \phi_H^2 \boldsymbol{\nabla} \mu + S_H, \\
\frac{\partial \phi_N}{\partial t} &= \boldsymbol{\nabla} \cdot \bar{M}_N \phi_N^2 \boldsymbol{\nabla} \mu + S_N, \\
\frac{\partial \phi_\sigma}{\partial t} &= \boldsymbol{\nabla} \cdot M_\sigma \delta_\sigma^{-1} \boldsymbol{\nabla} \phi_\sigma - \boldsymbol{\nabla} \cdot M_\sigma \chi_0 \boldsymbol{\nabla} \phi_T + S_\sigma, \\
\frac{\partial \phi_{TAF}}{\partial t} &= \boldsymbol{\nabla} \cdot M_{TAF} \delta_{TAF}^{-1} \boldsymbol{\nabla} \phi_{TAF} + S_{TAF}, \\
\frac{\partial \phi_e}{\partial t} &= \boldsymbol{\nabla} \cdot \bar{M}_e \phi_e^2 \boldsymbol{\nabla} \mu_e + S_e, \\
\mu_e &= \Psi'(\phi_e) - \epsilon_e^2 \Delta \phi_e,
\end{aligned} \tag{10.2}$$

in $\Omega \times (0, T)$, where $S_T = \lambda_P \phi_\sigma (\phi_T - \phi_H - \phi_N) - \lambda_A (\phi_T - \phi_N)$, and

$$\begin{aligned}
S_P &= \lambda_P \phi_\sigma \phi_P - \lambda_A \phi_P - \lambda_{PH} \mathcal{H}(\sigma_{PH} - \phi_\sigma) \phi_P \\
&\quad + \lambda_{HP} \mathcal{H}(\phi_\sigma - \sigma_{HP}) \phi_H, \\
S_H &= -\lambda_A \phi_H + \lambda_{PH} \mathcal{H}(\sigma_{PH} - \phi_\sigma) \phi_P - \lambda_{HP} \mathcal{H}(\phi_\sigma - \sigma_{HP}) \phi_H \\
&\quad - \lambda_{HN} \mathcal{H}(\sigma_{HN} - \phi_\sigma) \phi_H, \\
S_N &= \lambda_{HN} \mathcal{H}(\sigma_{HN} - \phi_\sigma) \phi_H, \\
S_\sigma &= -\lambda_P \phi_\sigma \phi_P - \lambda_{P_h} \phi_\sigma \phi_H + \lambda_A (\phi_P + \phi_H) + \alpha_R (1 - \phi_\sigma) \phi_e, \\
S_{TAF} &= \lambda_{TAF} (1 - \phi_{TAF}) \phi_H - \alpha_{TAF} \phi_{TAF} \phi_e \mathcal{H}(\phi_{TAF} - \phi_{TAF_c}), \\
S_e &= \alpha_P (\phi_{TAF}) \phi_e \mathcal{H}(\phi_{TAF} - \phi_{TAF_c}),
\end{aligned} \tag{10.3}$$

Here $\mathcal{H}$ denotes the Heaviside step function, defining switches from proliferative to hypoxic, back to proliferative cells, apoptosis, *etc.* depending on the nutrient content ϕ_σ and cell concentration ϕ_T. The λ and α are material parameters and M's cell mobility functions. Here we take

$$\begin{gathered}
\mathcal{M}(\phi_T, \phi_H, \phi_N) = \bar{M}_P (\phi_T - \phi_H - \phi_N)^2 + \bar{M}_H \phi_H^2 + \bar{M}_N \phi_N^2, \\
\text{for } M_H = \bar{M}_T \phi_H^2,\ M_N = \bar{M}_N \phi_N^2,\ M_e = \bar{M}_e \phi_e^2,
\end{gathered} \tag{10.4}$$

where $\bar{M}_P$, $\bar{M}_H$, $\bar{M}_N$ and $\bar{M}_e$ are positive constants (or time-independent densities). Among typical boundary conditions on the boundary $\partial\Omega$ over a time interval $(0, \tau)$, we list the following:

$$\begin{aligned}
\boldsymbol{\nabla} \phi_T \cdot \boldsymbol{n} = \boldsymbol{\nabla} \mu_T \cdot \boldsymbol{n} = \boldsymbol{\nabla} \phi_H \cdot \boldsymbol{n} = \boldsymbol{\nabla} \phi_N \cdot \boldsymbol{n} = \boldsymbol{\nabla} \mu_\sigma \cdot \boldsymbol{n} = 0, \\
\boldsymbol{\nabla} \phi_{TAF} \cdot \boldsymbol{n} = \boldsymbol{\nabla} \phi_e \cdot \boldsymbol{n} = \boldsymbol{\nabla} \mu_e \cdot \boldsymbol{n} = \boldsymbol{\nabla} \phi_\sigma \cdot \boldsymbol{n} = 0,
\end{aligned} \tag{10.5}$$

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

on $\partial\Omega \times (0,\tau)$, $\boldsymbol{n}$ being a unit exterior normal to $\partial\Omega$.

10.4. OPAL: predicting glioma growth in murine subjects

As emphasized in earlier sections of the article, the predictability of computational models brings on the ability and use of relevant observational data to calibrate and validate the models, a fact frequently overlooked by modellers. Here we describe an application of OPAL of Section 8.5 in which sparse data on the progression of brain cancer, glioma, in murine subjects (laboratory rats) are provided over a 20-day period through MRI images.

These subclasses of the models in (10.2) and (10.3) are considered.

(i) Reaction–diffusion models (μ=constant$\times\phi_T$):

$$\frac{\partial \phi_T}{\partial t} = \boldsymbol{\nabla} \cdot M_T^* + \lambda_T^{grow} \phi_T (1 - \phi_T). \tag{10.6}$$

(ii) Reaction–diffusion models with mechanical coupling:

$$\begin{aligned} \frac{\partial \phi_T}{\partial t} &= \boldsymbol{\nabla} \cdot M_T^* + \lambda_T^{grow} \phi_T (1 - \phi_T), \\ \mu &= c\phi_T - \lambda \boldsymbol{\nabla} \cdot \boldsymbol{u}. \end{aligned} \tag{10.7}$$

(iii) Phase-field models:

$$\begin{aligned} \frac{\partial \phi_T}{\partial t} &= \boldsymbol{\nabla} \cdot M_{PF}(\phi_T) \boldsymbol{\nabla} \mu + \lambda_T^{grow} \phi_T (1 - \phi_T), \\ \mu &= 2\bar{E}\phi_T(1 - 3\phi_T + 2\phi_T^2) - \epsilon_T^2 \delta \phi_T - \lambda \boldsymbol{\nabla} \cdot \boldsymbol{u}. \end{aligned} \tag{10.8}$$

Here $M_T, \lambda_T^{grow}, c, \lambda, \bar{E}, \varepsilon_T$ are possibly random material parameters, $\boldsymbol{u}$ is a global displacement field over Ω determined by a model of isotropic linear elasticity with mechanical parameter E and ν. Mechanical coupling with cell mobility is modelled by setting $M_T = \bar{M}_T \exp(\gamma\Sigma)$ and $M_T^* = \bar{M_T}^* \exp(\gamma\Sigma)$ and $\lambda_T^{grow} = \bar{\lambda}_T^{grow} \exp(\gamma\Sigma)$, Σ being the von Mises stress in the tumour mass.

A set of 13 possible models can be derived from (10.6)–(10.8) for different combination of parameters, leading to four reaction–diffusion models (RD01–RD04), four reaction–diffusion models with mechanical effect (MD01–MD04), and four phase field models (PH01–PH05). The models, parameters and Occam categories are indicated in Table 10.1.

Sensitivity analysis of all parameters was conducted using the elementary effect method (Morris method) described in Section 7.4.1 and uniform priors were employed for all parameters for the calibration phase, but the calibration posteriors used as priors in the validation scenario proved to be multi-modal and non-Gaussian. Monte Carlo methods were used to solve the forward problems and MCMC-parallel tempering algorithms were employed for the Bayesian inversion. Full details are given in Lima *et al.* (2016). Results of the OPAL calculations are collected in Table 10.2.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

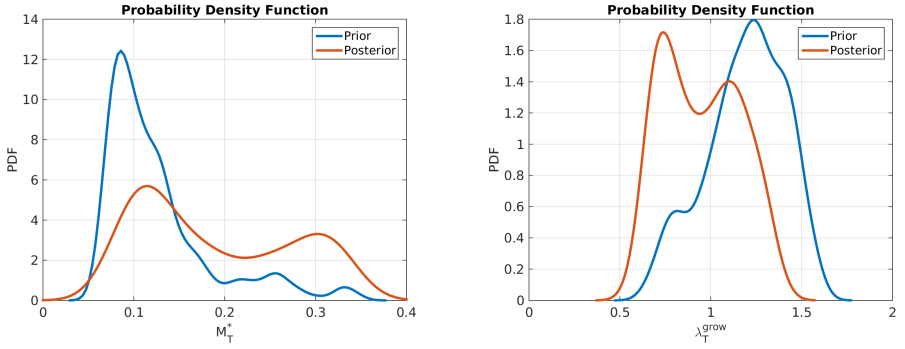
Table 10.1. Initial set of possible models (from Lima *et al.* 2016).

Model	Variables			Parameters											No. of	Occam
	ϕ_T	μ	$\boldsymbol{u}$	M_T	M_T^*	c	λ_T^{grow}	$\bar{E}_T$	ϵ_T	E	ν	λ	γ	γ^g	params	category
RD01	✓				✓		✓								2	1
PF01	✓	✓		✓			✓	✓	✓						4	2
RD02	✓		✓		✓		✓			✓	✓	✓	✓		6	3
RD03	✓		✓		✓		✓			✓	✓	✓		✓	6	3
MD01	✓	✓	✓	✓		✓	✓			✓	✓	✓			6	3
RD04	✓		✓		✓		✓			✓	✓	✓	✓	✓	7	4
MD02	✓	✓	✓	✓		✓	✓			✓	✓	✓	✓		7	4
MD03	✓	✓	✓	✓		✓	✓			✓	✓	✓		✓	7	4
PF02	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓			7	4
MD04	✓	✓	✓	✓		✓	✓			✓	✓	✓	✓	✓	8	5
PF03	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓		8	5
PF04	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓		✓	8	5
PF05	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	9	6

Table 10.2. Results and plausibilities (from Lima *et al.* 2016).

Model	Occam category	Plausibility	D_{KL} (σ_{16}^2)	D_{KL} (σ_{18}^2)
RD01	1	N/A	0.92	0.70
PF01	2	N/A	0.65	0.74
RD02	3	0.22		
RD03	3	0.14		
MD01	3	0.64	0.93	0.72
RD04	4	0.02		
MD02	4	0.06		
MD03	4	0.02		
PF02	4	0.90	0.68	0.76
MD04	5	0.02		
PF03	5	0.34		
PF04	5	0.64	0.70	0.66
PF05	6	N/A		

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press



(a) M_T^* (b) λ_T^{grow}

Figure 10.2. Prior ($\pi^V_{prior}(\boldsymbol{\theta}_1|\mathcal{P}_1, \mathcal{M})$) and posterior ($\pi^V_{post}(\boldsymbol{\theta}_1|\boldsymbol{y}_V, \mathcal{P}_1, \mathcal{M})$) for model RD01 at the validation step. (Reproduced from Lima *et al.* 2016 with permission from World Scientific.)

The first model, RD01, a reaction–diffusion model with only two parameters, M_T^* and λ_T^{grow}, is in Occam category 1. The Kullback–Leibler divergence divided by the variance in data (σ_t^z) at $t = 16$ and $t = 18$ days is used as a validation metric with validation tolerance $\gamma_{tol} = 0.7$ corresponding to a threshold of around 10% of the tumour mass (or the tumour area). The computed validation prior (calibration posterior) and validation posterior are illustrated in Figure 10.2.

Using the stated criteria, RD01 is declared *invalid*. Through similar calculations in the OPAL framework, PF01 (category 2) was also invalid. Model PF04 from category 5 satisfies the prescribed tolerance ($D_{KL}(\sigma_{16}^2) = 0.070$, $D_{KL}(\sigma_{18}^2) = 0.66$), and is declared *not* invalid (or 'valid'). Thus, it is not necessary to consider more complex models such as PF05.

Extensions of such model selection and validation have been represented for experiments involving radiation therapy for glioma (Lima *et al.* 2017), and calculation of highly heterogeneous multispecies models such as (10.7) is based on *in vitro* laboratory data on cell cultures, is a topic currently under study. A comparison of predicted and measured tumour areas is given in Figure 10.3.

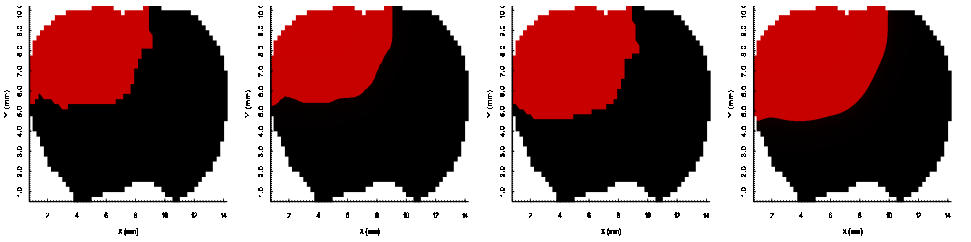
10.5. Multiscale models of tumour growth

We close this section with a discussion of a truly multiscale model of tumour growth that transcends spatial and temporal scales.

(i) *Tissue scale*: phenomenological models involving stochastic partial differential equation (centimetres, days, years).

(ii) *Cell models*: agent-based or rule-based models of cell proliferation, mobility and apoptosis (microns, minutes).

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press



(a) $32.73\,\mathrm{mm}^2$, data (b) $32.45\,\mathrm{mm}^2$, PF04 (c) $37.92\,\mathrm{mm}^2$, data (d) $41.96\,\mathrm{mm}^2$, PF04

Figure 10.3. Comparison of measured and predicted tumour area evolution for a slice at $t = 16$ (a, b) and $t = 18$ (c, d). (Reproduced from Lima *et al.* 2016 with permission from World Scientific.)

(iii) *Transduction*: subcell network models of signalling pathways involving large systems of nonlinear ordinary differential equations (nanometres, microseconds).

One could also add to this hierarchy genome models that use genetic sequencing to regulate and inform events (protein structures) in the cell nucleus.

As a representative of the tissue scale class, we recall the model of mass balance of tumour cells in $(10.7)_1$:

$$\frac{\partial \phi_T}{\partial t} = \boldsymbol{\nabla} \cdot M(\phi_T)\boldsymbol{\nabla}\phi_T + \lambda^{grow}\phi_\sigma(\phi_P) - \lambda^{apop}(\phi_T - \phi_N) \tag{10.9}$$

plus initial and boundary conditions, ϕ_T being the tumour cell volume fraction, ϕ_σ the nutrient VF, $\phi_P = \phi_T - \phi_H - \phi_N$, the VF of proliferative cells, and ϕ_H and ϕ_N the VFs of hypoxic and necrotic cells. As ϕ_P and ϕ_σ increase, ϕ_T grows at a rate governed by the proliferative rate parameter λ^{grow}. As the difference in viable tumour cells and necrotic cells increases, cells die through apoptosis at a rate influenced by parameter λ^{apop}.

At the cellular scale, a point $\boldsymbol{x}$ in D at time t with a tumour cell VF ϕ_T becomes a domain $D_T^{(t)}(\boldsymbol{x})$ occupied by many thousands of tumour cells bathing in a uniform nutrient bath of intensity $\phi_\sigma, \Omega_T^{(t)}(\boldsymbol{x})$. There, depending on the level of a protein *epidermal growth factor* (EGF), the individual cells may or may not traverse the cell cycle, going from quiescent states (G_0) to synthesis (S) to proliferative states (G_1) at which mitosis can occur, usually in cell splitting and, therefore, proliferation. Or the cells may die, either by apoptosis, natural death (homeostasis), or becoming necrotic due to lack of nutrients. Clusters of cells may separate and move in $\Omega_T^{(t)}(\boldsymbol{x})$, coming into contact with other cells, and adhesion of cells may affect proliferation rates. These cell-level events are orchestrated by rules depending on the nutrient concentration and the concentration of a protein EGF, manipulated in the cell nucleus and by cell densities.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

The cells themselves are suspended in an extra-cellular medium in which molecules react with receptor proteins on the surface membrane of the cell. This initiates a biochemical chain of events, called a *signalling cascade* or *signal transduction*, that leads to cell response, such as signalling cell differentiation, apoptosis, proliferation through mitosis, gene expression, *etc.* The rates of reactions of proteins along these subcell pathways are generally believed to be governed by the laws of chemical kinetics, formally *stoichiometry*, and for this reason, chemical kinetics provides a natural framework for developing discrete network models of signal transduction – generally characterized by systems of ordinary differential equations connecting protein concentrations through enzymatic reactions.

One of the principal threads connecting the models of different scales (there are several) is the production of a protein EGF, in the cell nucleus as a product of cell transduction. This is the subcellular QoI: the target output that determines, in part, the cell sequencing at the cell-level model. Depending on the EGF concentrations (and nutrient levels), apoptosis rates rise and fall. A cell-level QoI is the average cell apoptosis and proliferation rates which, on average, determine the rate parameters λ^{grow} and λ^{apop} in the tissue scale model (10.9). A suitable QoI at the tissue scale is the ensemble average of the tumour mass of volume (*e.g.* $\mathbb{E}[\int_\Omega \phi_T(\cdot, T_{end})\, \mathrm{d}\boldsymbol{x}]$. A systematic analysis, calibration and validation of such multiscale models remains an open problem. Aside from formidable conceptual, mathematical and computational issues, the acquisition of relevant observational data at all scales stands as a major challenge.

Comments and acknowledgements

Portions of this article are based on lecture notes (Oden 2017) that can be accessed from https://www.ices.utexas.edu and on portions of a chapter in Oden *et al.* (2018). The calculations given in Section 9.5 are derived from the work of Laura Scarabosio, Barbara Wohlmuth and Danial Faghihi. The author thanks Danial Faghihi, Ernesto Lima and Amir Shahmoradi for their help in typing and proofing an earlier draft of the work. Extensions of the model selection paradigm and implementation of OPAL discussed in Section 10 have been made in the recent article by Lima *et al.* (2017). The support of the US Department of Energy, Office of Science, Office of Advanced Scientific Computing Research and Applied Mathematics Program during the writing of much of this work is gratefully acknowledged.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

REFERENCES[1]

J. Aczel (1966), *Lectures on Functional Equations and Their Applications*, Academic Press.

AIAA (1998), Guide for the verification and validation of computational fluid dynamics simulations. AIAA Standards G-077-1998(2002).

M. Ainsworth and J. T. Oden (2000), *A Posteriori Error Estimation in Finite Element Analysis*, Wiley.

H. Akaike (1974), 'A new look at the statistical model identification', *IEEE Trans. Automat. Control* **19**, 716–723.

H. Akaike (1976), Canonical correlation analysis of time series and the use of an information criterion. In *System Identification: Advances and Case Studies* (R. K. Mehra and D. G. Lainiotis, eds), Academic Press, pp. 27–96.

H. Akaike (1977), On entropy maximization principle. In *Applications of Statistics* (P. R. Krishnaiah, ed.), North-Holland.

H. Akaike (1998), Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike* (E. Parzen, K. Tanabe and G. Kitagawa, eds), Springer Series in Statistics, Springer, pp. 199–213.

S. Arnborg and G. Sjodin (2001), On the foundations of Bayesianism. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Vol. 568 of AIP Conference Proceedings, AIP, pp. 61–71.

ASME (2006), Guide for verification and validation in computational solid mechanics. ASME Committee PTC-60 V&V 10.

I. Babuška and J. T. Oden (2004), 'Verification and validation in computational engineering and science, I: Basic concepts', *Comput. Methods Appl. Mech. Engrg* **193**, 4047–4068.

I. Babuška, R. Tempone and F. Nobile (2008), 'A systematic approach to model validation based on Bayesian updates and prediction-related rejection criteria', *Comput. Methods Appl. Mech. Engrg* **197**, 2517–2539.

M. J. Bayarri, J. O. Berger, R. Paulo, J. Sacks, J. M. Cafeo, C. Cavendish, C.-H. Liu and J. Tu (2007), 'A framework for validation of computer models', *Technometrics* **49**, 138–154.

J. L. Beck (2010), 'Bayesian system identification based on probability logic', *Struct. Control Health Monit.* **17**, 825–847.

J. L. Beck and A. A. Taflanidis (2013), 'Prior and posterior robust stochastic predictions for dynamical systems using probability logic', *Int. J. Uncertainty Quantif.* **3**, 271–288.

J. L. Beck and K.-V. Yuen (2004), 'Model selection using response measurements: Bayesian probabilistic approach', *J. Engrg Mech.* **130**, 192–203.

R. Becker and R. Rannacher (2001), An optimal control approach to *a posteriori* error estimation in finite element methods. In *Acta Numerica*, Vol. 10, Cambridge University Press, pp. 1–102.

G. E. P. Box (1979), 'Robustness in the strategy of scientific model building', *Robustness in Statistics* **1**, 201–236.

[1] The URLs cited in this work were correct at the time of going to press, but the publisher and the authors make no undertaking that the citations remain live or are accurate or appropriate.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

G. E. P. Box and N. R. Draper (1987), *Empirical Model-Building and Response Surfaces*, Wiley Series in Probability and Statistics, Wiley.
M. Braack and A. Ern (2003), '*A posteriori* control of modeling errors and discretization errors', *Multiscale Model. Simul.* **1**, 221–238.
K. P. Burnham and D. R. Anderson (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer.
T. Carter (2007), An introduction to information theory and entropy. In *Complex Systems Summer School, Santa Fe*. http://astarte.csustan.edu/~tom/SFI-CSSS/info-theory/info-lec.pdf
G. C. Chow (1981), 'A comparison of the information and posterior probability criteria for model selection', *J. Econometrics* **16**, 21–33.
P. V. Coveney, E. R. Dougherty and R. R. Highfield (2016), 'Big data need big theory too', *Phil. Trans. R. Soc. A* **374**(2080), 20160153.
T. M. Cover and J. A. Thomas (2006), *Elements of Information Theory*, second edition, Wiley-Interscience.
R. T. Cox (1946), 'Probability, frequency and reasonable expectation', *Amer. J. Phys.* **14**, 1–13.
R. T. Cox (1961), *Algebra of Probable Inference*, Johns Hopkins University Press.
R. T. Cox (1979), 'Of inference and inquiry: An essay in inductive logic', In *The Maximum Entropy Formalism* (R. D. Levine and M. Tribus, eds), MIT Press, pp. 119–167.
R. I. Cukier, C. M. Fortuin, K. E. Shuler, A. G. Petschek and J. H. Schaibly (1973), 'Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients, I: Theory', *J. Chem. Phys.* **59**, 3873–3878.
W. A. Curtin and R. E. Miller (2003), 'Atomistic/continuum coupling in computational materials science', *Model. Simul. Mater. Sci. Engrg* **11**, R33.
M. J. Dupre and F. J. Tipler (2009), 'New axioms for rigorous Bayesian probabilities', *Bayesian Analysis* **4**, 599–606.
W. E (2011), *Principles of Multiscale Modeling*, Cambridge University Press.
W. E and B. Engquist (2003), 'The heterogeneous multiscale methods', *Commun. Math. Sci.* **1**, 87–132.
K. Farrell, J. T. Oden and D. Faghihi (2015), 'A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems', *J. Comput. Phys.* **295**, 189–208.
K. Farrell-Maupin and J. T. Oden (2017), 'Adaptive selection and validation of models of complex systems in the presence of uncertainty', *Research Math. Sci.* **4** 14.
R. A. Fisher (1922), 'On the mathematical foundations of theoretical statistics', *Phil. Trans. R. Soc. Lond.* A **222**, 309–368.
C. J. Geyer (2003), 5601 notes: The sandwich estimator. University of Minnesota School of Statistics.
M. B. Giles (2015), Multilevel Monte Carlo methods. In *Acta Numerica*, Vol. 24, Cambridge University Press, pp. 259–328.
P. Glasserman (2003), *Monte Carlo Methods in Financial Engineering*, Vol. 53 of Stochastic Modelling and Applied Probability, Springer.
P. Gregory (2005), *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica® Support*, Cambridge University Press.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

J. Y. Halpern (1999*a*), 'A counter example to theorems of Cox and Fine', *J. Artif. Intell. Research* **10**, 67–85.
J. Y. Halpern (1999*b*), 'Cox's theorem revisited', *J. Artif. Intell. Research* **11**, 429–435.
D. Hanahan and R. A. Weinberg (2000), 'The hallmarks of cancer', *Cell* **100**, 57–70.
D. Hanahan and R. A. Weinberg (2011), 'Hallmarks of cancer: The next generation', *Cell* **144**, 646–674.
Z. Hashin and S. Shtrikman (1963), 'A variational approach to the elastic behavior of multiphase materials', *J. Mech. Phys. Solids* **11**, 127–140.
W. K. Hastings (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**, 97–109.
S. Heinrich (2001), Multilevel Monte Carlo methods. In *LSSC 2001: International Conference on Large-Scale Scientific Computing*, Springer, pp. 58–67.
H. Hoel, E. Von Schwerin, A. Szepessy and R. Tempone (2014), 'Implementation and analysis of an adaptive multilevel Monte Carlo algorithm', *Monte Carlo Methods Appl.* **20**, 1–41.
T. Homma and A. Saltelli (1996), 'Importance measures in global sensitivity analysis of nonlinear models', *Reliab. Engrg Syst. Safety* **52**, 1–17.
C. Howson and P. Urbach (2006), *Scientific Reasoning: The Bayesian Approach*, Open Court Publishing.
C. M. Hurvich and C. Tsai (1989), 'Regression and time series model selection in small samples', *Biometrika* **76**, 297–307.
E. T. Jaynes (1968), 'Prior probabilities', *IEEE Trans. Systems Sci. Cybernet.* **4**, 227–241.
E. T. Jaynes (2003), *Probability Theory: The Logic of Science*, Cambridge University Press.
H. Jeffreys (1939), 'The times of P, S and SKS, and the velocities of P and S', *Geophys. Suppl. Monthly Not. R. Astron. Soc.* **4**, 498–533.
H. Jeffreys (1961), 'Small corrections in the theory of surface waves', *Geophys. J. Internat.* **6**, 115–117.
H. Jeffreys (1998), *The Theory of Probability*, third edition, Oxford University Press.
A. M. Johansen and L. Evers (2007), Monte Carlo methods. Lecture notes, University of Bristol.
M. C. Kennedy and A. O'Hagan (2001), 'Bayesian calibration of computer models', *J. Royal Statist. Soc.* B **63**, 425–464.
B. J. K. Kleijn and A. van der Vaart (2002), The asymptotics of misspecified Bayesian statistics. In *Proceedings of the 24th European Meeting of Statisticians* (T. Mikosch and M. Janžura, eds).
B. J. K. Kleijn and A. van der Vaart (2012), 'The Bernstein–Von-Mises theorem under misspecification', *Electron. J. Statist.* **6**, 354–381.
G. J. Klir (2005), *Uncertainty and Information: Foundations of Generalized Information Theory*, Wiley.
K. H. Knuth and J. Skilling (2012), 'Foundations of inference', *Axioms* **1**, 38–73.
S. Konishi and G. Kitagawa (2008), *Information Criteria and Statistical Modeling*, Springer.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

J.-D. Lebreton, K. P. Burnham, J. Clobert and D. R. Anderson (1992), 'Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies', *Ecological Monographs* **62**, 67–118.
E. Lima, J. Oden and R. Almeida (2014), 'A hybrid ten-species phase-field model of tumor growth', *Math. Models Methods Appl. Sciences* **24**, 2569–2599.
E. Lima, J. Oden, D. Hormuth, T. Yankeelov and R. Almeida (2016), 'Selection, calibration, and validation of models of tumor growth', *Math. Models Methods Appl. Sci.* **26**, 2341–2368.
E. Lima, J. Oden, B. Wohlmuth, A. Shahmoradi, D. Hormuth, T. Yankeelov, L. Scarabosio and T. Horger (2017), 'Selection and validation of predictive models of radiation effects on tumor growth based on noninvasive imaging data', *Comput. Methods Appl. Mech. Engrg* **327**, 277–305.
W. K. Liu, E. Karpov, S. Zhang and H. Park (2004), 'An introduction to computational nanomechanics and materials', *Comput. Methods Appl. Mech. Engrg* **193**, 1529–1578.
P. Macklin, V. Cristini and J. Lowengrub (2010), Biological background. In *Multiscale Modeling of Cancer: An Integrated Experimental and Mathematical Modeling Approach* (V. Cristini and J. Lowengrub), Cambridge University Press, pp. 8–23.
M. Maier and R. Rannacher (2016), 'Duality-based adaptivity in finite element discretization of heterogeneous multiscale problems', *J. Numer. Math.* **24**, 167–187.
M. Maier and R. Rannacher (2018), 'A duality-based optimization approach for model adaptivity in heterogeneous multiscale problems', *Multiscale Model. Simul.* **16**, 412–428.
N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller (1953), 'Equation of state calculations by fast computing machines', *J. Chem. Phys.* **21**, 1087–1092.
P. Ming and P. Zhang (2007), 'Analysis of the heterogeneous multiscale method for parabolic homogenization problems', *Math. Comp.* **76**(257), 153–177.
M. D. Morris (1991), 'Factorial sampling plans for preliminary computational experiments', *Technometrics* **33**, 161–174.
W. L. Oberkampf and C. J. Roy (2010), *Verification and Validation in Scientific Computing*, Cambridge University Press.
J. T. Oden (2014), 'Predictive computational science', *IACM Expressions* **35**, 2–4.
J. T. Oden (2017), Foundations of predictive computational sciences. ICES Reports.
J. T. Oden and S. Prudhomme (2001), 'Goal-oriented error estimation and adaptivity for the finite element method', *Comput. Math. Appl.* **41**, 735–756.
J. T. Oden and S. Prudhomme (2002), 'Estimation of modeling error in computational mechanics', *J. Comput. Phys.* **182**, 496–515.
J. T. Oden and K. S. Vemaganti (2000), 'Estimation of local modeling error and goal-oriented adaptive modeling of heterogeneous materials, I: Error estimates and adaptive algorithms', *J. Comput. Phys.* **164**, 22–47.
J. T. Oden, I. Babuška and D. Faghihi (2018), Predictive computational science: Computer predictions in the presence of uncertainty. In *Encyclopedia of Computational Mechanics* (E. Stein, R. de Borst and T. J. R. Hughes, eds), Wiley, pp. 1–26.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

J. T. Oden, E. A. Lima, R. C. Almeida, Y. Feng, M. N. Rylander, D. Fuentes, D. Faghihi, M. M. Rahman, M. DeWitt, M. Gadde *et al.* (2016), 'Toward predictive multiscale modeling of vascular tumor growth', *Arch. Comput. Methods Engrg* **23**, 735–779.

T. Oden, R. Moser and O. Ghattas (2010), 'Computer predictions with quantified uncertainty, I', *SIAM News* **43**, 1–3.

J. T. Oden, S. Prudhomme and P. T. Bauman (2005), 'On the extension of goal-oriented error estimation and hierarchical modeling to discrete lattice models', *Comput. Methods Appl. Mech. Engrg* **194**(34), 3668–3688.

J. T. Oden, S. Prudhomme, A. Romkes and P. T. Bauman (2006), 'Multiscale modeling of physical phenomena: Adaptive control of models', *SIAM J. Sci. Comput.* **28**, 2359–2389.

J. T. Oden, K. Vemaganti and N. Moës (1999), 'Hierarchical modeling of heterogeneous solids', *Comput. Methods Appl. Mech. Engrg* **172**, 3–25.

J. B. Paris (1994), *The Uncertain Reasoner's Companion: A Mathematical Perspective*, Cambridge University Press.

H. S. Park, E. G. Karpov, P. A. Klein and W. K. Liu (2005), 'Three-dimensional bridging scale analysis of dynamic fracture', *J. Comput. Phys.* **207**, 588–609.

S. Plimpton (1995), 'Fast parallel algorithms for short-range molecular dynamics', *J. Comput. Phys.* **117**, 1–19.

E. E. Prudencio and K. W. Schulz (2012), The parallel C++ statistical library 'QUESO': Quantification of uncertainty for estimation, simulation and optimization. In *European Conference on Parallel Processing* (M. Alexander *et al.*, eds), Vol. 7155 of Lecture Notes in Computer Science, Springer, pp. 398–407.

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

R. J. Reynolds and J. A. Schecker (1995), 'Radiation, cell cycle, and cancer', *Los Alamos Science* **23**, 51–89.

P. J. Roache (1998), *Verification and Validation in Computational Science and Engineering*, Hermosa Press.

A. Saltelli (2002), 'Making best use of model evaluations to compute sensitivity indices', *Comput. Phys. Commun.* **145**, 280–297.

A. Saltelli and I. M. Sobol' (1995), 'About the use of rank transformation in sensitivity analysis of model output', *Reliab. Engrg Syst. Safety* **50**, 225–239.

A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto and S. Tarantola (2010), 'Variance based sensitivity analysis of model output: Design and estimator for the total sensitivity index', *Comput. Phys. Commun.* **181**, 259–270.

A. Saltelli, K. Chan and E. M. Scott (2000), *Sensitivity Analysis*, first edition, Wiley Series in Probability and Statistics, Wiley.

A. Saltelli, K. Chan and E. M. Scott (2009), *Sensitivity Analysis*, Wiley Paperback Series, Wiley.

A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana and S. Tarantola (2008), *Global Sensitivity Analysis: The Primer*, Wiley.

G. Schwarz (1978), 'Estimating the dimension of a model', *Ann. Statist.* **6**, 461–464.

G. Shafer (1976), *A Mathematical Theory of Evidence*, Princeton University Press.

C. E. Shannon (1948), 'A mathematical theory of communication', *Bell System Technical J.* **27**, 379–423, 623–656.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press

L. E. Shilkrot, R. E. Miller and W. A. Curtin (2004), 'Multiscale plasticity modeling: Coupled atomistics and discrete dislocation mechanics', *J. Mech. Phys. Solids* **52**, 755–787.

I. M. Sobol' (1990), 'Sensitivity estimates for nonlinear mathematical models', *Matematicheskoe Modelirovanie* **2**, 112–118.

I. M. Sobol' (1993), 'Sensitivity analysis for non-linear mathematical models', *Math. Model. Comput. Experiment* **1**, 407–414.

I. M. Sobol' (2007), 'Global sensitivity analysis indices for the investigation of nonlinear mathematical models', *Matematicheskoe Modelirovanie* **19**, 23–24.

M. H. Stone (1936), 'The theory of representation for Boolean algebras', *Trans. Amer. Math. Soc.* **40**, 37–111.

B. Szabó and I. Babuška (2011), *Introduction to Finite Element Analysis: Formulation, Verification and Validation*, Wiley.

E. B. Tadmor, M. Ortiz and R. Phillips (1996), 'Quasicontinuum analysis of defects in solids', *Philos. Mag.* A **73**, 1529–1563.

A. Terenin and D. Draper (2015), Rigorizing and extending the Cox–Jaynes derivation of probability: Implications for statistical practice. arXiv:1507.06597

M. Tribus (2016), *Rational Descriptions, Decisions and Designs: Pergamon Unified Engineering Series*, Elsevier.

K. S. Van Horn (2003), 'Constructing a logic of plausible inference: A guide to Cox's theorem', *Int. J. Approx. Reasoning* **34**, 3–24.

K. S. Vemaganti and J. T. Oden (2001), 'Estimation of local modeling error and goal-oriented adaptive modeling of heterogeneous materials, II: A computational environment for adaptive modeling of heterogeneous elastic solids', *Comput. Methods Appl. Mech. Engrg* **190**(46), 6089–6124.

G. J. Wagner and W. K. Liu (2003), 'Coupling of atomistic and continuum simulations using a bridging scale decomposition', *J. Comput. Phys.* **190**, 249–274.

J. Zimmermann and A. B. Cremers (2011), The quest for uncertainty. In *Rainbow of Computer Science: Essays Dedicated to Hermann Maurer on the Occasion of His 70th Birthday* (C. S. Calude *et al.*, eds), Vol. 6570 of Lecture Notes in Computer Science, Springer, pp. 270–283.

https://doi.org/10.1017/S096249291800003X Published online by Cambridge University Press