BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2010

External Bayesian analysis for computer simulators

Michael Goldstein
Durham University, England
Michael.Goldstein@durham.ac.uk

Summary

Computer simulators offer a powerful approach for studying complex physical systems. We consider their use in current practice and the role of external uncertainty in bridging the gap between the properties of the model and of the system. The interpretation of this uncertainty analysis raises questions about the role and meaning of the Bayesian approach. We summarise some theory which is helpful to clarify and amplify the role of external specifications of uncertainty, and illustrate some of the types of calculation suggested by this approach.

Keywords and Phrases:

Computer simulators; Model discrepancy; Interpretation of Bayesian analysis; Bayes linear analysis; Temporal Sure Preference; Galaxy simulators; History matching.

1. INTRODUCTION

Mathematical models, implemented as computer simulators, offer a powerful methodology for studying complex physical systems. However, the practical use of such simulators to make statements about the behaviour of the systems which the models represent can be problematic, requiring a careful treatment of the uncertainty involved in moving from the model to the system. This paper offers an overview of aspects of this problem, paying particular attention to conceptual issues and discussing some relevant theory which may help to address some of the issues raised.

My first involvement in this area was described in a previous Valencia volume (Craig et al (1996)). Good starting points for getting into the relevant literature relating to the Bayesian treatment of complex models are Kennedy and O'Hagan(2001)

Michael Goldstein is Professor of Statistics at Durham University. Work on this paper was supported by a grant for the Managing Uncertainty for Complex Models consortium, under the Basic Technology initiative. Special thanks to Ian Vernon for all assessments related to the Galform analysis.

and the accompanying discussion, and Santner et al (2003). The methodology has its roots in work on computer experiments, see for example, Sacks et al (1989), which was largely motivated by the aim of choosing decision inputs to optimise system performance. An excellent general resource for Bayesian treatment of uncertainty analysis for complex models is the web-site for the Managing Uncertainty for Complex Models (MUCM)project (MUCM is a consortium of UK Universities funded through the Basic Technology initiative to translate the basic science of the Bayesian approach to these problems into a working methodology with wide applicability for dealing with inferences based on computer simulators.) The project URL is http://mucm.group.shef.ac.uk/index.html

2. COMPUTER SIMULATORS FOR PHYSICAL SYSTEMS

Consider the following three contrasting uses of simulators for complex physical systems, each constructed as the implementation of a scientific model. Firstly, the study of the development of the Universe is carried out by using a Galaxy formation simulator. The aim is purely scientific - to gain information about the physical processes underlying the Universe. Secondly, an oil reservoir simulator is used in order to manage the assets associated with the reservoir. The aim is purely commercial, to develop efficient production schedules, determine whether and where to sink new wells, and so forth. Thirdly, large scale climate simulators are constructed to assess the likely effects of human intervention upon future climate behaviour. Our aims are both scientific - there is much that is unknown about the large scale interactions which determine climate outcomes - and also intensely practical, as such simulators provide suggestions as to the importance of changing human patterns of behaviour before possibly irreversible changes are set into motion.

In all such cases, whether driven by science, commerce or public policy, the simulators help us to understand the underlying processes which determine complex physical phenomena. Using such simulators raises serious challenges in dealing with the uncertainty associated with the analysis. This uncertainty is well suited to Bayesian treatment and various methodologies have been developed for this purpose.

3. INTERNAL AND EXTERNAL UNCERTAINTIES

3.1. *Internal uncertainty analysis*

It is helpful, when dealing with problems around computer simulators, to divide the uncertainty into two basic categories, namely **internal** and **external** uncertainties.

Internal uncertainties are those which arise directly from the problem description. Many analyses in practice are carried out purely on the basis of assessing all of the internal uncertainties, as these are an unavoidable component of any treatment of the problem. External uncertainties are all of the additional uncertainties which arise when we consider whether the treatment of the internal uncertainties indeed provides us with a satisfactory uncertainty description of the physical system itself. Most of the conceptual challenges associated with the Bayesian treatment of computer simulators arise in the appropriate treatment of the external uncertainties.

We introduce some notation to describe the different sources of uncertainty. While the examples of modelling the universe, the climate or the reservoir differ in all physical aspects, the formal structures that we need to analyse are very similar, which is why we may talk of a common underlying methodology. Each simulator can be conceived as a function $f(x)$, where x is the (often high dimensional) input vector, representing unknown properties of the physical system and $f(x)$ is an (often high dimensional) output vector representing various aspects of the behaviour of the system. For example, in a climate model, x might be a specification of a collection of physical parameters which determine the behaviour of the various physical processes (relating to clouds, ice, convection, boundary layer, radiation and so forth) which are needed in order to construct a description of climate behaviour and a typical element of $f(x)$ might be, for example, the global mean temperature in 100 years time. Interest in the model usually centres on the "appropriate" (in some sense) choice, x_0, for x, the extent to which the output $f(x_0)$ can be viewed as informative for actual system behaviour, y, the use that we can make of historical observations z observed with error on a subset y_h of y, and the optimal assignment of any decision inputs, d, in the model. In the climate model, y_h corresponds to historical climate observations recorded over space and time and the decisions might correspond to different carbon emission scenarios.

In the simplest version of this problem, where observations are made, without error, the model is a precise reproduction of the system and the function is simple to invert, we can write

$$z = f_h(x_0) \tag{1}$$

where $f_h(x)$ is the subvector of outputs of $f(x)$ corresponding to the subset $y_h = z$. We invert f_h to find x_0 (either as a unique choice or as a family of precise solutions of (1)) and predict future system behaviour, y_p, exactly from the components $f_p(x_0)$ which correspond to the elements of y_p. If the future output depends on decision inputs, then we may optimise $f_p(x_0, d)$ over choices of d.

In practice, determining x_0, by inverting relation (1), may be extremely complicated if the dimensions of y and x are high and the function $f(x)$ is expensive, in time and computational resources, to evaluate for any choice of x. For example, large climate models may take months to evaluate, for a single input choice, on extremely high specification computers. In such cases, we must recognise that the function f, although deterministic, must be treated as uncertain for all input choices except the relatively small subset for which an actual evaluation has been made. Therefore, an important part of the Bayesian analysis is the construction of a detailed description of the uncertainty about the value of the function at each possible choice of input. Such a representation is sometimes termed an emulator of the function - the emulator both suggests an approximation to the function and also contains an assessment of the likely magnitude of the error of the approximation. A good introduction to emulation is given by O'Hagan (2006).

In order to carry out this programme in practice, we also need to recognise that the observations z are made with error, and separate the uncertainty representation into two relations:

$$z = y_h \oplus e, \tag{2}$$

$$y = f(x_0) \tag{3}$$

where e has some appropriate probabilistic specification, possibly involving parameters which require estimation. (Here and below the notation $U \oplus V$ denotes the sum $U + V$ of two random quantities, U, V which are either independent, if there is a full probabilistic specification, or uncorrelated if there is only a second order specification.)

Specification of an appropriate prior distribution for x_0, likelihood for the observational error e and probabilistic emulator for f, updated by appropriate choices of evaluation of the function and observation of the data z, gives a Bayesian treatment of the statistical inverse problem. We term this an **internal uncertainty analysis**. The analysis is conceptually straightforward, though it may be technically challenging, requiring particular care when constructing the emulator for the function and dealing with the computational difficulties arising from high dimensional and often highly multimodal likelihood functions

3.2. *External uncertainties*

The internal analysis described in section (3.1) is a common way of carrying out an uncertainty analysis based on a computer simulator for a physical system. However, when we consider how relevant such an analysis may be for understanding the behaviour of the actual physical system, then we must take into account the potential mismatch between the simulator and the physical system that it purports to represent. Much of the controversy over the value of climate science centres on the crucial issue as to how much faith we can put in evaluations of climate simulators as meaningful predictors of actual future climate behaviour. This discussion has been very public because of the pressing social concerns, but similar distinctions occur in all areas of model based science. For example, reservoir simulators are used to guide extremely costly investment decisions for which the question of the reliability of the model projections is of enormous concern. Even in pure scientific areas of enquiry, such as Galaxy formation simulators, the relationship between the simulator and the system is fundamental, as the extent to which it would be reasonable to expect the given Galaxy simulator to match actual large scale observable features of the universe determines the amount of mismatch between simulator output and observations that can be tolerated without calling into question the basic science underpinning the modelling.

To complete our analysis, we must address the external uncertainties arising from the potential mismatch between the problem description provided by the computer simulator and the actual behaviour of the physical system. A physical model is a description of the way in which system properties (the inputs to the model) affect system behaviour (the output of the model). This description involves two basic types of simplification.

Firstly, we approximate the properties of the system. Partly, this is because the relevant properties of the system are too complicated and extensive to describe fully and partly this is because, even if we were able to enumerate all of the system properties, then we would not have the requisite knowledge to allow us to specify these values exactly. In our model description, this latter simplification corresponds to uncertainty in features such as the initial and boundary conditions, and forcing functions. Such conditions often require specification of extremely high dimensional spatio-temporal fields which are quite impractical to incorporate into the posterior Bayesian inferences; for example, the Galaxy formation model that we have referenced requires a specification of the precise configuration of all dark matter across space and time. As this is unknown, we must assess the uncertainty that this lack of knowledge introduces into the simulation.

Secondly, we approximate the rules whereby the model assesses system behaviour given system properties. Partly this is because of necessary mathematical simplifications of the extremely complex interactions within the system, partly this results from further necessary simplifications for tractability of the computer implementation and partly this is because we do not fully understand the physical laws which govern the process, so that we are unable to fully replicate actual system behaviour even from the precise description of the system properties.

Neither of these approximations invalidates the modelling process. On the contrary, such simplifications are essential, to give us a practical way of exploring the basic drivers of system behaviour. The problem arises when we ignore the fact that we have made such simplifications and confuse the internal uncertainty analysis of the model with the corresponding uncertainty analysis for the physical system itself. Rather than conflating the model and the system, it is always important to maintain the distinction between the two, so that we can give a proper account as to the information about the system which the model is capable of providing.

We describe the uncertainties about the physical system which result from these approximations as **external** uncertainties. External uncertainties would not be present if the model was a perfect representation of the system, and the extent of the mismatch determines the magnitude and importance of the external contribution to the analysis. In most problems of realistic size and complexity, external components of uncertainty will not be ignorable.

In this view, the collection of evaluations of the simulator provides a partial description of the physical relationships which govern system behaviour which, in combination with historical observations for aspects of the system, reduces our actual uncertainty about the behaviour of the physical system. The internal model analysis is completed by carrying out an external uncertainty analysis which deconflates the model and the system by taking into account all of the additional uncertainties. It may be scientifically challenging and technically difficult to quantify these external uncertainties, but at least this does address the actual question of interest, namely the behaviour of the physical system, rather than the surrogate question of the behaviour of the model.

There are many different ways to take into account all of the external uncertainties in the problem. The simplest, and I would guess most popular, approach is simply to build an extra level of uncertainty into the representation by expanding (3) as

$$y = f(x_0) \oplus \epsilon \tag{4}$$

where ϵ, often termed the model or structural discrepancy, has some appropriate probabilistic specification, possibly involving parameters which require estimation, and is taken to be independent of f, x_0, e. It may appear that e and ϵ are not distinguished in (4), as the two parts of the equation combine to give $z = f(x_0) \oplus e \oplus \epsilon$. However, e and ϵ are usually treated very differently in the formal specification, and, in particular, the correlation structure over the elements of ϵ is an essential part of the uncertainty description, determining, for example, the extent to which underprediction of historical elements in a time series of system observations is likely to be replicated in model predictions of future elements of the series. In principle, the uncertainty structure describing ϵ should be assessed based on a careful analysis of the effect of each of the simplifications and approximations to the system properties and the rules governing the model. However, because such an analysis tends to be extremely complicated, often the uncertainty assessment is made directly as an order of magnitude expert judgement, possibly updated by Bayesian analysis of the mismatch between historical data and the output of the model for a group of carefully chosen model evaluations.

While the range of potential applications of this methodology is enormous, and each application is governed by its own special conditions, in general a Bayesian analysis which proceeds as we have described by careful treatment of the relations in (2) and (4) would be very much the state of the art. Indeed, in most areas of potential application, it would be substantially beyond the state of the art.

The choice of relation (4) is by no means intended to exclude consideration of more careful ways of treating external uncertainty. Goldstein and Rougier (2009) gives a careful treatment of the limitations of this representation. The discussion following that article gives an overview of many of the issues that arise when we attempt to deal seriously with the issues arising from model and system mismatch. However, (4) has the virtue of simplicity and familiarity, and is sufficient for the purposes of this article. Our basic principle is that it is always better to recognise than to ignore uncertainty, even if the modelling and analysis of the uncertainty is difficult and partial. It is hard to imagine a non-trivial problem for which this process of adding uncertainty to distinguish the model from the system is not important and for which even using the simple form (4) would fail to be a considerable improvement on ignoring external uncertainty altogether. Using a simple form for expressing this uncertainty is therefore a good starting point, which reflects current practice even if, for important applications, we may want to go further in our analysis. In any case, the arguments that we shall develop will be essentially the same irrespective of the form that we choose for the external representation.

4. THE MEANING OF AN EXTERNAL UNCERTAINTY ANALYSIS

The problem that we identified with the internal uncertainty analysis of section (3.1) was how to attribute meaning to a treatment which failed to distinguish between the simulator and the system which the model purports to represent. To what extent have we addressed this question by introducing an external treatment of uncertainty?

We must consider the meaning of the analysis arising from our treatment of the uncertainties. While it is understandable to talk informally about, for example, the risk of rapid climate change, when we come to give a careful meaning to words like risk, probability or uncertainty, then we need to be more precise. In the subjectivist Bayes view, the meaning of any probability statement is straightforward in principle as it is the uncertainty judgement of a specified individual, expressed on the scale of probability by consideration of some operational elicitation scheme, for example by consideration of betting preferences. I do not want to rehearse again the relative merits of the subjective Bayes position versus other shades of Bayesianism, but refer instead to the discussion papers Goldstein (2006) and Berger (2006) and the discussion and rejoinder to these papers. Many interesting points were made in this discussion but the one which concerns us is the meaning of the analysis. I restrict my attention here to the subjective Bayes interpretation simply because it does have an agreed and testable meaning which is sufficiently precise that it is capable of forming the basis of a discussion about the meaning of the analysis of a computer simulator. It would be interesting to construct an equivalent form of analysis for any of the other interpretations for uncertainty. In particular, if, in some areas of study, there is a genuinely objective and well defined meaning that can be attributed to certain uncertainty statements, then that does not change any of the following discussion, as these well-defined uncertainty statements simply join the collection of unknown quantities about which we must form our subjective assessments.

The choice of the subjectivist view of uncertainty does not settle the question as to the meaning of the uncertainty analysis, but it does allow us to pose it clearly. So, let us consider again what we mean by a statement such as "the risk of rapid climate change". This quote from the BBC web-site is typical:

'Fortunately, rapid climate change is one area that the UK has taken the lead in researching, by funding the Rapid Climate Change programme (RAPID), the aim of which is to determine the probability of rapid climate change occurring.' See

`www.bbc.co.uk/weather/features/science_nature/the_day_after_tomorrow.shtml`

In the subjectivist interpretation, any probability statement is the judgement of a named individual, so we should speak not of the probability of rapid climate change, but instead of Anne's probability or Bob's probability of rapid climate change and so forth. There is a substantial problem of perception here, as most people expect something more authoritive and objective than a probability which is one person's judgement. However, the disappointing thing is that, in almost all cases, stated probabilities emerging from a complex analysis are not even the judgements of any individual. So, until we produce methods of uncertainty analysis that are so compelling that everyone would have to agree with them, it is not unreasonable to make the more modest requirement that the objective of our analysis should be uncertainties which are asserted by at least one person. If a wider group of people, say a research team, share a consensus view as to the uncertainty, then that is even better, but Bayesian theory only describes how an individual's uncertainties are formed and modified by evidence, so let us start there.

Is the assertion of the uncertainty of an individual scientifically valuable of itself? Usually, not. The fact that an uncertainty statement is the actual judgement of an individual only has value if this individual is sufficiently knowledgeable in the area for his/her judgements to carry weight and if the analysis that has led to this judgement has been both sufficiently careful and thorough to support this judgement and also sufficiently transparent that the reasoning, not simply the conclusions, can be understood and reassessed by similarly knowledgeable experts in the field. So, let us suppose, for the purposes of this discussion, that the objective of the analysis is to produce the "best" current judgements of a specified expert, in a sufficiently transparent form that the reasoning which led to these judgements should be open to critical scrutiny.

The property of "best judgement" is, perhaps necessarily, somewhat imprecise. What we mean is judgements that are sufficiently well founded that the expert is not aware of any further calculations that could feasibly be done which would be judged to lead to substantially improved assessments. Explicitly introducing this notion allows us to formalise the judgement as to when an analysis is good enough for purpose, in the sense that there are no feasible improvements that we can suggest, possibly within some restricted class of options better to define feasibility, that would have substantial practical consequences. The question that we are considering, namely whether the Bayesian analysis of the model does indeed represent the judgements of the expert, can, in a sense, be rendered uninteresting. If experts are too busy, too lazy or too uninterested in the problems, then they are always free to equate their beliefs with the results of the formal analysis, however flawed, faulty or misconceived they perceive the analysis to be. However, best current judgements set a more rigorous standard, and it is a fair and important question for experts to have to assess and reveal just how "second best" they have allowed their declared judgements to be.

We now consider how well the external analysis of the computer model relates to these objectives.

5. INTERPRETING AN EXTERNAL UNCERTAINTY ANALYSIS

Suppose that we specify our uncertainty relations according to (4). We describe all of our uncertainties probabilistically and carry out a Bayesian analysis. Is the output of the analysis our best current judgements about the problem at hand? Is it even our actual current judgements. If not, then what is the meaning of the analysis?

Is the Bayesian approach based on (2) and (4) an analysis of our actual uncertainties or does it, instead, provide a model for such an analysis? We have argued that, in general, an internal model analysis must be completed by taking account of all of the external uncertainties in the problem. Does this mean that the analysis based on (4) is missing a further layer of external uncertainty? If not, what makes uncertainty modelling different from all other forms of modelling? If so, does this lead to an infinite regress of uncertainties?

In section (3.2), we identified two basic reasons why we needed to distinguish between the model analysis and the behaviour of the system by adding a layer of structural discrepancy uncertainty to distinguish the two. Firstly, the model description approximates the properties of the system. That will certainly be the case for most probabilistic specifications. In all but the simplest cases, the sheer volume of quantitative specification which is required in order to carry out a Bayesian analysis inevitably results in a host of pragmatic simplifications to the prior specification.

The second reason attributed for the need to introduce structural discrepancy is that physical models approximate the rules whereby the model assesses system behaviour given system properties. In our case, we must ask whether probabilistic rules provide the correct treatment of the way to construct actual posterior beliefs given prior beliefs and data. This is a fundamental question which has engaged many people and engendered a considerable literature. It is beyond the scope of this article to treat this question properly, so all that I will say is that I have never seen a convincing argument which explains why Bayes theorem (or any other theorem) actually describes the way in which beliefs should change when confronted with evidence. The various coherence arguments which are presented in support of the use of Bayes theorem are based on considerations such as avoiding sure loss concerning the value that you attribute to certain “called off bets” i.e. bets that you place now but which only take effect if certain evidential outcomes occur, otherwise the bets are called off and stakes are returned. No-one has ever offered an even semi-rigorous justification for the argument that, if we place a called off bet now, then at some future time this should actually correspond to our actual betting rate when we learn that the conditioning event did take place. Indeed, this process, of laying our conditional bets now, then completely abstaining from further reflection until such time as we observe the conditioning event, and nothing else, is so far removed from our experience that it is hard to reconcile this description with the actual and manifold considerations that we bring to bear in actually reassessing our beliefs. Further, even were such a situation to arise for which this was an accurate account, then there would still be no logical equivalence between the judgements relating to the called off bet and the judgement that we make when we learn of the outcome and nothing else, as the fact that we learned nothing else is relevant to our revision of judgement but is not incorporated into the conditioning event on which we declare the called off bet. Most careful discussions of the foundations of probabilistic reasoning recognise the element of abstraction within the Bayesian formalism by invoking such considerations as the inferential behaviour of "perfectly rational individuals" or the "small worlds" account of Savage. Such accounts are insightful in helping us to recognise the strengths and limitations of Bayesian reasoning not as a description of inferential reasoning itself but instead as a model for such reasoning.

Just as climate scientists study climate by means of climate models, Bayesian statisticians study belief modification by means of Bayesian models. Just as for climate scientists, the models are a crucial source of information and insight, but to treat the model inference as identical to, rather than as informative for, our actual inferences is to make the same mistake as it would be to conflate the climate model with climate itself. So, let us consider what happens when we treat the Bayesian analysis in the same way as any other model of a complex process. The system properties correspond to the prior specification, the system behaviour is the judgements or alternately the best current judgments of an individual and the probabilistic rules are what the model uses to link the two specifications. As for any other model, we need to deconflate the model output and the best current judgements, by adding a further external layer of uncertainty. There is a certain amount of theory which can help us to do this, which also is revealing as to the strengths of the Bayesian formalism as a natural choice for modelling the inferential process, as we shall now describe.

6. SOME RELEVANT THEORY

6.1. *Adjusted expectation*

In order to develop theory which distinguishes between actual posterior judgements and the results of a Bayesian analysis, we need a formalism that treats the two sets of uncertainty judgements as logically distinct but related. The best way that I know to do this is to start by making expectation, rather than probability, the primitive for the theory. This is in line with de Finetti's treatment of expectation (de Finetti (1974)) where he chooses expectation over probability for the reason that, if expectation is primitive then we can choose to make as many or as few expectation statements as we choose, whereas, if probability is primitive, then we must make all of the probability statements before we can make any of the expectation statements. This distinction is less important within a framework where all of the probability statements are themselves part of a model. However, when we discuss the meaning of the analysis, it is very helpful to be able to identify which subset of statements are to be invested with meaning. As any probability is the expectation of the indicator function for the corresponding event, we can treat a full probabilistic analysis under this formalism if we wish, but we have the option of restricting our attention to whatever subcollection of specifications we are interested in analysing carefully.

We can analyse expectations directly using the Bayes linear approach, in which we make direct prior specifications for that collection of means, variances and covariances which we are both willing and able to assess, and update these prior assessments by linear fitting.

Suppose that we have two collections of random quantities, namely vectors $\boldsymbol{B} = (B_1, ..., B_r)$, $\boldsymbol{D} = (D_0, D_1, ..., D_s)$ where $D_0 = 1$, and we observe $\boldsymbol{D}$

The *adjusted* or *Bayes linear* expectation for B_i given $\boldsymbol{D}$ is the linear combination $\boldsymbol{a}_i^T \boldsymbol{D}$ minimising $\mathrm{E}((B_i - \boldsymbol{a}_i^T \boldsymbol{D})^2)$ over choices of $\boldsymbol{a}_i$ evaluated as

$$\mathrm{E}_{\boldsymbol{D}}(\boldsymbol{B}) = \mathrm{E}(\boldsymbol{B}) + \mathrm{Cov}(\boldsymbol{B}, \boldsymbol{D})(\mathrm{Var}(\boldsymbol{D}))^{-1}(\boldsymbol{D} - \mathrm{E}(\boldsymbol{D}))$$

The *adjusted variance matrix* for $\boldsymbol{B}$ given $\boldsymbol{D}$, is

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{D}}(\boldsymbol{B}) &= \mathrm{Var}(\boldsymbol{B} - \mathrm{E}_{\boldsymbol{D}}(\boldsymbol{B})) \\ &= \mathrm{Var}(\boldsymbol{B}) - \mathrm{Cov}(\boldsymbol{B}, \boldsymbol{D})(\mathrm{Var}(\boldsymbol{D}))^{-1}\mathrm{Cov}(\boldsymbol{D}, \boldsymbol{B})\end{aligned}$$

Adjusted expectation is numerically equivalent to conditional expectation in the particular case where $\boldsymbol{D}$ comprises the indicator functions for the elements of a partition, i.e. where each D_i takes value one or zero and precisely one element D_i will equal one, eg, if B is the indicator for an event, then

$$\mathrm{E}_{\boldsymbol{D}}(B) = \mathrm{P}(B|\boldsymbol{D}) = \sum_i \mathrm{P}(B|D_i)D_i$$

An account of Bayes linear methodology is given in Goldstein and Wooff (2007)

There are a range of differing roles and meanings that we can attribute to a Bayes linear analysis. For our purposes, the relevant considerations arise from the relations between adjusted expectation and posterior judgements, as we now describe.

6.2. *Temporal sure preference*

In general, while our preferences may be rational at each individual time point, there need be no linkage whatsoever between the collections of judgments at different time points. In order to establish links between our judgments at different time points, we need ways of describing 'temporal rationality' which go beyond being internally rational at each time point. Our description is operational, concerning preferences between random penalties, as assessed at different time points, considered as payoffs in probability currency (i.e. tickets in a lottery with a single prize). [With payoffs in probability currency, expectation for the penalty equals the probability of the reward. Therefore, changes in preferences between penalties A and B over time correspond to changes in probability, rather than utility.]

Current preference for random penalty A over penalty B, even when augmented by conditional statements about preferences given possible future evidential outcomes, cannot logically constrain future preferences; for example, you may obtain further, hitherto unsuspected, information or insights into the problem before you come to make your future judgments. It is more compelling to suggest that future preferences may determine prior preferences. Suppose that you must choose between two (probability currency) random penalties, A and B. Suppose that at some future time the values of A and B will be revealed, and you will pay the penalty that you have chosen.

For your future preferences to influence your current preferences, you must know what your future preference will be. Therefore, we introduce the notion of a sure preference. You have a **sure preference** for A over B at (future) time t, if you know now, as a matter of logic, that at time t you will not express a strict preference for penalty B over penalty A. The temporal consistency principle that we impose is that future sure preferences are respected by preferences today. We call this the **temporal sure preference principle**, as follows.

Suppose that you have a sure preference for A over B at (future) time t. Then you should not have a strict preference for B over A now.

Temporal sure preference is not a rationality requirement. It is an extremely weak and operationally testable principle which will often appear reasonable and which has important consequences for statistical reasoning. In Goldstein (1997), the temporal sure preference principle is discussed and it is shown that it implies that your actual posterior expectation, $\mathrm{E}_T(\boldsymbol{B})$, at time T when you have observed $\boldsymbol{D}$, satisfies two relations

$$\boldsymbol{B} = \mathrm{E}_T(\boldsymbol{B}) \oplus \boldsymbol{\epsilon_T} \tag{5}$$

$$\mathrm{E}_T(\boldsymbol{B}) = \mathrm{E}_{\boldsymbol{D}}(\boldsymbol{B}) \oplus \epsilon_D, \tag{6}$$

where $\boldsymbol{\epsilon_T}, \boldsymbol{\epsilon_D}$ each have, a priori, zero expectation and are uncorrelated with each other and with $\boldsymbol{D}$. If $\boldsymbol{D}$ represents a partition, then $\mathrm{E}_T(\boldsymbol{B}) = \mathrm{E}(\boldsymbol{B}|\boldsymbol{D}) \oplus \boldsymbol{\epsilon_D}$ where $\mathrm{E}(\epsilon_D|\boldsymbol{D_i}) = 0, \forall i$.

Equations (5) and (6) establish stochastic relationships between the quantities of interest, the actual posterior judgements for these quantities and the formal Bayes linear or full Bayes evaluations. The conditions required to establish these relations are extremely weak, and therefore very widely applicable. These relations deconflate the Bayesian assessments and the actual posterior judgments allowing us to take account of the difference between the model for the inference and the actual inference.

We can give two interpretations of such relations. Firstly, if we intend, actually, to update our beliefs then we have a direct relation between the formal Bayes analysis and the actual posterior judgements. Secondly, suppose that we carry out the Bayes analysis, but we do not reflect further on our actual posterior judgements. In that case, we may interpret the relations as adding an explicit layer of external uncertainty into our Bayesian analysis representing the amount of uncertainty about what our actual best current judgements would be, were we to make the considerable investment of effort required to determine what these judgements actually were.

7. EXTERNAL BAYESIAN ANALYSIS FOR COMPUTER SIMULATORS

Compare relations (5), (6) with (4). Let us suppose that we have made an appropriate choice for x_0. In assessing our uncertainty about the physical system, given x_0, there are two considerations. Firstly, we do not know the value of $f(x_0)$. Secondly, even if we did know the value of $f(x_0)$, then we still would not know the value of the system behaviour y.

Let us expand the first component, our uncertainty about $f(x_0)$ using (6). We must specify our judgements about the function $f(x)$ given an ensemble of evaluations $F = (f(x_1), \ldots f(x_n))$.

Suppose we employ the formal Bayesian model for updating judgements about the function by assessing the adjusted (Bayes linear or conditional Bayes) expectation $\mathrm{E}_F(f(x))$ at each value of x, by means of functional emulation. According to our discussion above, there are two external uncertainties which are ignored by such an analysis. Firstly, the prior specifications within our model are approximations to specifications which are sufficiently careful that we are justified in applying the temporal sure preference arguments of the preceding section. Therefore our first level of external uncertainty distinguishes $\mathrm{E}_F(f(x))$ from the adjusted expectation, $\mathrm{E}_{F^*}(f(x))$ which would follow from such a careful specification by introducing external uncertainty ϵ_*. The second level of external uncertainty corresponds to the difference between the Bayes or Bayes linear expectation for $f(x)$ and the full posterior judgement $\mathrm{E}_T(f(x))$, so that we introduce external uncertainty ϵ_F. We link $\mathrm{E}_T(f(x))$ with $f(x)$ by adding ϵ_T.

We therefore decompose our view as to the value of $f(x)$, as the composition of three relations as follows.

$$f(x) = \mathrm{E}_T(f(x)) \oplus \epsilon_T(x) \tag{7}$$

$$\mathrm{E}_T(f(x)) = \mathrm{E}_{F^*}(f(x)) \oplus \epsilon_F(x) \tag{8}$$

$$\mathrm{E}_{F^*}(f(x)) = \mathrm{E}_F(f(x)) \oplus \epsilon_*(x) \tag{9}$$

Our specification is completed by linking $f(x_0)$ to y. If we equate $f(x_0)$ with $\mathrm{E}_T(y)$ given x_0, then relation (5) reduces to (4). However, this is a strong requirement and often, we may prefer to view $f(x_0)$ as informative for, but distinct from, the judgement of $\mathrm{E}_T(y)$ given x_0, which we may write as $f^*(x_0)$. In such cases, we may decompose the model discrepancy ϵ into two components, by introducing the functional discrepancy $\epsilon(x)$, as

$$f^*(x) = f(x) \oplus \epsilon(x) \tag{10}$$

and only linking y to the model analysis through the value of $f^*(x_0)$ as

$$y = f^*(x_0) \oplus \epsilon \tag{11}$$

The separation of model discrepancy into the two components ϵ and ϵ_x raises important issues of principle which are discussed in Goldstein and Rougier(2009) .

7.1. *History matching and model adequacy*

So far, we have examined the uncertainty description for the computer simulator and the implications for system behaviour, but we have not paid similar attention to the mismatch between x_0 and the physical properties of the system. This is mainly to simplify the account, as, in principle, there is a missing layer of probabilistic assessment linking inputs to the model with actual properties of the physical system. For simplicity, we are moving this mismatch directly into the discrepancy function for the simulator. This is particularly appropriate when we are making a preliminary assessment of model adequacy.

There are many formal and informal ways of assessing model adequacy. A searching test is whether there is any choice of input x_0 for which the model output $f(x_0)$ is able to reproduce an extensive range of different observable phenomena within a plausible tolerance level. This is a different problem from model calibration, which starts from the view that there is a true but unknown value of x_0 and aims to produce a posterior distribution for this true value given all of the other pieces of the problem. In history matching, the aim is to find the set of all of the choices of

x_0 which give an acceptable match to historical observations, and it is of particular interest if this set is empty as this poses a fundamental challenge to the science underlying the modelling. However, in order to reach a meaningful outcome, we must have a realistic understanding of the potential mismatch between the model and the system which is consistent with the scientific view implemented within the simulator. Hence, we need to think carefully about the probabilistic magnitude of all of the ingredients of the discrepancy. In my view, history matching, for models with extensive history, is almost always of interest provided that model discrepancy has been carefully assessed. If an extensive range of acceptable matches to history can be found, then, depending on the problem, it may be of interest to calibrate the model over the choices within this set.

8. ILLUSTRATION: GALAXY FORMATION MODELS

8.1. *Galform*

As a small illustration of the kinds of external analysis that we might carry out, we refer to part of a study of the Galaxy formation simulator, Galform. The study, carried out in collaboration with the Galform group in the Durham Institute for Computational Cosmology, addresses our basic understanding as to how galaxies form, and, for example, whether the galaxies we observe have been formed in the presence of large amounts of dark matter. The Galform simulation proceeds in two stages.

Firstly, an N-Body dark matter simulation is run to determine the behaviour of fluctuations in the early Universe, and their subsequent growth into millions of galaxy sized lumps of mass in the following 12 billion years. This is a very heavy simulation, done on a supercomputer and cannot be easily repeated.

Secondly, these results are then used by a more detailed Galaxy Formation simulation (Galform) which models the far more complicated interactions of normal matter such as gas cloud formation, star formation and the effects of black holes at the centre of the galaxy.

The second stage simulation cannot be carried out for the whole of the space determined by the first simulation. Instead, the output of the first simulation is divided into 128 different computer models corresponding to different regions of the universe, i.e. different dark matter configurations, determined in the first simulation. For consistency with previous analyses of Galform, the analysis carried out in this study was based on the average of the values of the Galform simulation on each of a specified collection of 40 sub-regions. This simulation, for a given choice of input parameters, took around 20 minutes in total, per evaluation, for the 40 sub-regions. The simulation output comprised many large scale attributes of the universe, for example, the number of galaxies of certain luminosity and colour per unit volume, which could be compared to observational data. The study considered variation in 17 of the input parameters to the simulation controlling features such as the amount of energy in the form of gas blown out of a galaxy due to star formation, the time it takes this energy to return, and the effect the central black hole has in keeping large galaxies ‘hot’. The objective was to history match the Galform output against observed luminosity functions. Crucial to such investigations is the care that must be taken in assessing how much difference may be allowed between functional output and observational data while still considering a match to be potentially acceptable.

8.2. *External Galform analysis*

To illustrate this process, we choose a particular output, corresponding to the log of the observed number of galaxies per unit volume with measured luminosity in a specified range centred on a particular luminosity value, 17, on an appropriate scale.

The variance of the model discrepancy ϵ in (4) was specified as the sum of three components. Firstly, for computational convenience, it was decided to eliminate consideration of those inputs which only appeared to have a small effect on the outputs. The standard deviation of the error introduced by this simplification was assessed to be 0.0412. Secondly, the reliance on the mean of the collection of 40 regions as the choice of function ignored the additional variation as actual observations are made within a particular limited part of the universe. The standard deviation of the error introduced by this simplification was assessed to be 0.0129. Finally, Richard Bower, our lead collaborator in the cosmology group, made a careful assessment of the additional external variation in the Galform model, specifying the standard deviation of this external error to be 0.0753. For comparison, the standard deviation of the observational error on the observed value was assessed as 0.0294.

These assessments determine the closeness of the match that can be required between the computer function and the observation for this output, while still leading to an acceptable match. Because it was impossible actually to evaluate $f(x)$ for each choice of x, this comparison could not be made directly. Therefore, beliefs about each selected component f_i of f, were represented using emulators of the form,

$$f_i(x) = \sum_j \beta_{ij} g_{ij}(x) + u_i(x) \tag{12}$$

where $B = \{\beta_{ij}\}$ are unknown scalars, g_{ij} are known deterministic functions of x, and $u_i(x)$ is a weakly second order stationary stochastic process, with correlation function

$$\mathrm{Corr}(u_i(x), u_i(x')) = \exp(-(\frac{\|x - x'\|}{\theta_i})^2) \tag{13}$$

Emulators were fitted, given a collection of model evaluations, F, using a range of statistical tools, with a substantial component of expert judgement, supported by a careful diagnostic analysis.

Using the emulator, for each x, the Bayes linear expectation $\mathrm{E}_F(f_i(x))$ was constructed, for each component of interest. Instead of comparing $f_i(x)$ with historical observations, z_i was compared to the emulator expectation $\mathrm{E}_F(f_i(x))$. This required the addition of a further element of uncertainty, namely the adjusted standard deviation of $f_i(x)$ given F, $\mathrm{SD}_F(f_i(x))$. $\mathrm{SD}_F(f_i(x))$ varied with the choice of x, but for the component $f_i(x)$ that we have chosen above, a typical order of magnitude value would be between 0.03 and 0.05, given the initial collection of 1,000 evaluations of Galform.

This analysis is described in detail in Bower, Vernon et al (2010) which covers both the internal and external uncertainty analysis for the Galform model and the use of such analysis in the context of History Matching. Use of this methodology did eventually lead to discovery of a large number of acceptable history matches, suggesting that a reasonable representation had been made for the various elements of uncertainty in the problem. However, it does raise the question as to the meaning of the analysis. While the uncertainty analysis was carried out carefully, it would be overstating the case to claim that the results produced were Richard Bower's actual posterior judgements, still less his best judgements. Therefore, had we failed to find any history matches to within the declared tolerance, it is unclear as to what conclusions we would have been justified in drawing. To address these concerns requires us to consider the external form of the Bayesian analysis itself. Carrying out such an analysis properly is as large a task as was carrying out the original external analysis on the Galform model. There is no simple and automatic way to carry this analysis out. However, for illustration, we now carry out two demonstration portions of such an analysis.

8.3. *External Bayesian analysis*

The first example calculation that we shall consider is the external uncertainties in (9). We must consider the extent to which we might have come to a different assessment of the adjusted expectation of the functional output had we made a more careful prior specification. While every feature of the prior specification may be subject to scrutiny, an obvious concern is the form of the correlation function, (13), which forms the basis for the residual variation in the emulator. The judgement of equal smoothness of the function across the whole of the input space is made for reasons of simplicity rather than out of conviction. This is an issue that arises widely in most methods for functional emulation. We aim to minimise the force of this assumption by fitting an informative regression surface, and only imposing the requirement on the residual, but we should still consider the impact of the assumption upon the analysis. There are many ways to do this. Here is a simple assessment that we carried out in order to place an order of magnitude standard deviation on the term ϵ_* in (9) for the output $f_i(x)$ for which we discussed the external analysis in section (8.2).

We chose twenty well separated input values x. For each, we considered the effect on the value of $\mathrm{E}_F(f_i(x))$ both of increasing and of reducing the selected value of θ_i by 20%, in each case re-assessing the value of $\mathrm{E}_F(f_i(x))$ and therefore assessing the difference between the original and revised adjusted expectation for $f(x)$. From these changes, we assessed roughly the order of magnitude variation that we would need to specify for ϵ_* to be consistent with these calculations. (A value of 20% for the change in θ was chosen on the grounds that this would be just about large enough for careful study of the local residual surface to reveal discrepancies of such magnitude as the basis for a more careful calculation, while not being so large as to have already shown up in our diagnostic analysis. However, we should emphasise that this reasoning is purely informal and illustrative, in order to reach quickly a not totally implausible order of magnitude for the additional variance that we would like to specify.)

The results of the analysis depended on the choice of input x. As we noted in section 8.2, $\mathrm{SD}_F(f(x))$ varied up to a value of around 0.05, and for each of the choices that we inspected, the effect of the above calculations suggested a standard deviation for ϵ_* of magnitude around 10% of $\mathrm{SD}_F(f(x))$, which is small but perhaps not completely ignorable.

Secondly, let us consider one of the elements of the external Galform analysis itself. We observed that one component of the external variation was the error introduced by ignoring the variation in dark matter across the universe, and thus the variation attributed to making observations within our limited portion of the universe. The standard deviation assigned for this component, was taken to be

0.0129. In our formulation, this variation, as part of the variation of ϵ in (4), does not change if we make different choices for x_0. However, it is possible that the amount of regional variation should be considered x_0 dependent.

We explore this as follows. The original function evaluations consisted of 1,000 choices of input parameters, each separately evaluated for each of the forty regions. Calculation of the variation in the standard deviation over this collection of evaluations was carried out to give an indication of the variation that might be assigned to ϵ as a function of x_0. The results of this analysis showed that over the great majority of input choices, the value that we had assigned was similar enough to the actual choice that was employed, namely 0.0129. However, there were a few parameter choices where the sample variation across regions would have been better assessed as around 50% larger than our chosen value. This effect might deserve further attention. We can introduce this effect into our external analysis by giving the variance of ϵ a degree of x dependence, or simply increasing the variation attributed to this term to be large enough to be appropriate to all parameter choices. Alternately, if we consider the variation of this uncertainty term to be large enough to call into question the results of our analysis, then we may more seriously address the issue raised by building the two stage representation (10), (11) to account for the external uncertainty which is attributed to the relationship between regional variation and parameter choice. The way to do this is to evaluate the individual functions $f_{[R]}(x)$ assessed over region $R = 1, \ldots 40$. Considering the functions $f_{[R]}(x)$ to be exchangeable over R, we may create an exchangeability representation which allows the precise deduction as to the uncertainty of f in our region of the universe. Full details as to how to do this are given in House et al (2009) and the analogous construction to (5), (6) for assessing external uncertainty for exchangeable structures is given in Goldstein (1994).

9. CONCLUSIONS

This paper considers a very large subject, namely the consequences of recognising that the Bayesian approach is a model for inference, and therefore needs to be treated as does any other model, namely by considering all of the external sources of uncertainty that are necessary to relate the model analysis to real and useful things in the world. The motivation and methodology is related to but logically distinct from assessments of sensitivity and robustness for Bayesian models. For comparison, we may assess sensitivity and robustness of a climate model to certain key choices, but this is not the same as identifying and quantifying the limitations of the model in informing us about climate. Similarly, the external Bayesian analysis aims to help bridge the gap between a formal Bayesian analysis and actual learning about aspects of the world. In each problem, we should seek to clarify the meaning of our analysis, by considering why the resulting uncertainty statements should claim to have value. Are they actually uncertainties asserted by a knowledgeable expert who has done all that could reasonably be done in order to reach these conclusions? If we are not claiming such an interpretation for our analysis, then are we almost making such a claim or, if not, then what alternative meaning can we give?

The external assessment may be difficult, but only because it usually is genuinely difficult to be sure of the worth of our analysis. In any case, it is no more difficult than it was to create the mathematical model and to implement the theory as a computer simulator. It is only that those activities are recognised and resourced, whereas the uncertainty analysis is very often treated as an afterthought.

It might be argued that this type of external analysis is making more difficult something that was already hard anyway. In a way this is true. Just as we have to build and understand a climate model before we can carry out a meaningful external uncertainty analysis for the model, it may be that we have to explore Bayesian uncertainty modelling before addressing its limitations. However, eventually, modellers need to move out of their comfort zones and face the consequences of their modelling choices within the real world, and statisticians are no different from any other modellers in this regard. Unlike statisticians of other persuasions, Bayesian are well placed to meet this challenge as their structures can meaningfully be embedded in a larger probabilistic formalism within which the strengths and weaknesses of their modelling can be fully explored.

REFERENCES

Berger, J. (2006) The case for objective Bayesian analysis *Bayesian Analysis* **1** 385-402

Bower, R.G. Vernon, I., Goldstein, M. Lacey, C.G., Benson, A.J., Baugh, C.M. Cole, S., Frenk, C.S. (2010) The Parameter Space of Galaxy Formation, Monthly Notices of the Royal Astronomical Society Main Journal, to appear

Craig, P. S., Goldstein, M., Seheult, A. H. and Smith, J. A. (1996) Bayes linear strategies for history matching of hydrocarbon reservoirs, 69-98 *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press

de Finetti, B (1974) *Theory of Probability, vol 1* New York: Wiley

Goldstein, M. (1994) Revising exchangeable beliefs: subjectivist foundations for the inductive argument, in *Aspects of Uncertainty: a Tribute to D. V. Lindley* (P. R. Freeman, and A. F. M. Smith, eds.) Chichester: Wiley, 201-222

Goldstein, M (1997) Prior inferences for posterior judgements (1997), in *Structures and norms in Science*, M.C.D. Chiara et. al. eds., Dordrecht: Kluwer, 55-71.

Goldstein, M. (2006) Subjective Bayesian analysis: principles and practice *Bayesian Analysis* **1** 403-420

Goldstein, M. and Rougier, J.C. (2009) Reified Bayesian modelling and inference for physical systems, *J. Statist. Planning and Inference* **139**, 1221-1239

Goldstein, M. and Wooff, D. (2007) *Bayes linear Statistics: Theory and Methods* Chichester: Wiley

House, L., Goldstein, M. and Vernon, I.R. (2009) Exchangeable Computer Models MUCM Technical report 10/02.

Kennedy, M.C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *J. Roy. Statist. Soc. B* **63**, 425-464.

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering and System Safety 91, 1290-1300.

Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, H.P. (1989) Design and analysis of computer experiments. *Statist. Science* **4**, 409-435.

Santner, T., Williams, B. and Notz, W. (2003). The Design and Analysis of Computer Experiments. New York: Springer