Bayes Linear Analysis of Imprecision in Computer Models, with Application to Understanding Galaxy Formation

Ian Vernon
Department of Mathematical Sciences
Durham University
I.R.Vernon@durham.ac.uk

Michael Goldstein
Department of Mathematical Sciences
Durham University
Michael.Goldstein@durham.ac.uk

Abstract

Imprecision arises naturally in the context of computer models and their relation to reality. An imprecise treatment of general computer models is presented, illustrated with an analysis of a complex galaxy formation simulation known as Galform. The analysis involves several different types of uncertainty, one of which (the Model Discrepancy) comes directly from expert elicitation regarding the deficiencies of the model. The Model Discrepancy is therefore treated within an Imprecise framework to reflect more accurately the beliefs of the expert concerning the discrepancy between the model and reality. Due to the conceptual complexity and computationally intensive nature of such a Bayesian imprecise uncertainty analysis, Bayes Linear Methodology is employed which requires consideration of only expectations and variances of all uncertain quantities. Therefore incorporating an Imprecise treatment within a Bayes Linear analysis is shown to be relatively straightforward. The impact of an imprecise assessment on the input space of the model is determined through the use of an Implausibility measure.

Keywords. Bayesian Inference, Computer models, Calibration, Imprecise model discrepancy, Implausibility, Galaxy Formation, Graphical Representation of Model Imprecision.

1 Introduction

Computer models make imprecise statements about physical systems. This arises because of compromises made in the physical theory and in approximations to solutions of very complex systems of equations. Therefore any statement about a physical system, for example climate change, which is derived from the analysis of computer models will be necessarily imperfect, as it will usually be very difficult to put a precise quantification on the discrepancy between the model analysis and the physical system [1]. A full probabilistic representation of the imprecision arising from such model discrepancy will typically be very complex and difficult to analyse. However, there is an alternative way to express such imprecision, based on viewing expectations rather than probability as the natural primitive for expressing uncertainty statements. This formulation allows us to focus directly on 'high level' summary expressions of imprecision. This approach is termed Bayes Linear Analysis; for a detailed treatment see [2].

In this paper we show how the Bayes Linear approach may be used to capture the most important features of the imprecision arising from the use of complex physical models. We illustrate our approach with the galaxy formation model known as Galform. Galform simulates the formation and evolution of approximately 1 million galaxies from the beginning of the Universe until the current day (a period of approximately 13 billion years). It gives outputs representing various physical features of each of the galaxies which can be compared with observational data [3].

This paper is structured as follows: in section 2 we discuss the Galform model in more detail, in section 3 the theory of computer models and the incorporation of the imprecise model discrepancy is described, and in section 4 we develop appropriate graphical displays for such imprecise analyses and demonstrate the application of these methods to the Galform model.

2 Cosmology and Galaxy Formation

2.1 Understanding the Universe

Over the last 100 years, major advances have been made in understanding the large scale structure of the Universe. Current theories of cosmology suggest that the Universe began in a hot, dense state approximately 13 billion years ago, and that it has been expanding rapidly ever since. However, there exists a major problem: observations of galaxies imply that

there must exist far more matter in the Universe than the visible matter that makes up stars, planets and us. This is referred to as 'Dark Matter' and understanding its nature and how it has affected the evolution of galaxies within our Universe is one of the most important problems in modern cosmology.

In order to study many of the effects of Dark Matter, cosmologists try to model Galaxy formation using complex computer models. In this paper, we develop the Bayesian treatment of imprecision for computer models, and illustrate our analysis using one such model, known as Galform (developed by the Galform group at the Institute for Computational Cosmology, Durham University).

2.2 Galform: a Galaxy Formation Simulation

Simulating the formation of large numbers of galaxies from the beginning of the Universe until the current day is a difficult task and so the process is split into two parts. First a Dark Matter simulation is performed to determine the behaviour of fluctuations of mass in the early Universe, and their subsequent growth into millions of galaxy sized lumps in the following 13 billion years. Second, the results of the Dark Matter simulation are used by a more detailed model called Galform which models the far more complicated interactions of normal matter including: gas cloud formation, radiative cooling, star formation and the effects of central black holes.

The first simulation is run on a volume of space of size $(1.63 \text{ billion light-years})^3$. This volume is split into 512 sub-volumes which are independently simulated using the second model Galform, which is the subject of the Imprecise Uncertainty Analysis in this paper (see figure 1). Each run of Galform takes 20-30 minutes per subvolume per processor.

2.3 Galform Inputs and Outputs

The Galform simulation provides many outputs related to approximately 1 million simulated galaxies. We consider the two most important types of output: the bj and K band luminosity functions. The bj band luminosity function gives the number of blue (i.e. young) galaxies of a certain luminosity per unit volume, while the K band luminosity function describes the number of red (i.e. old) galaxies (see Figure 1). The colour of a galaxy comes from the stars it contains, stars which on average burn bluer early in their lifecycle and redder as they age. These outputs can be compared to observational data gathered by the 2dFGRS galaxy survey (see [3] and references therein).

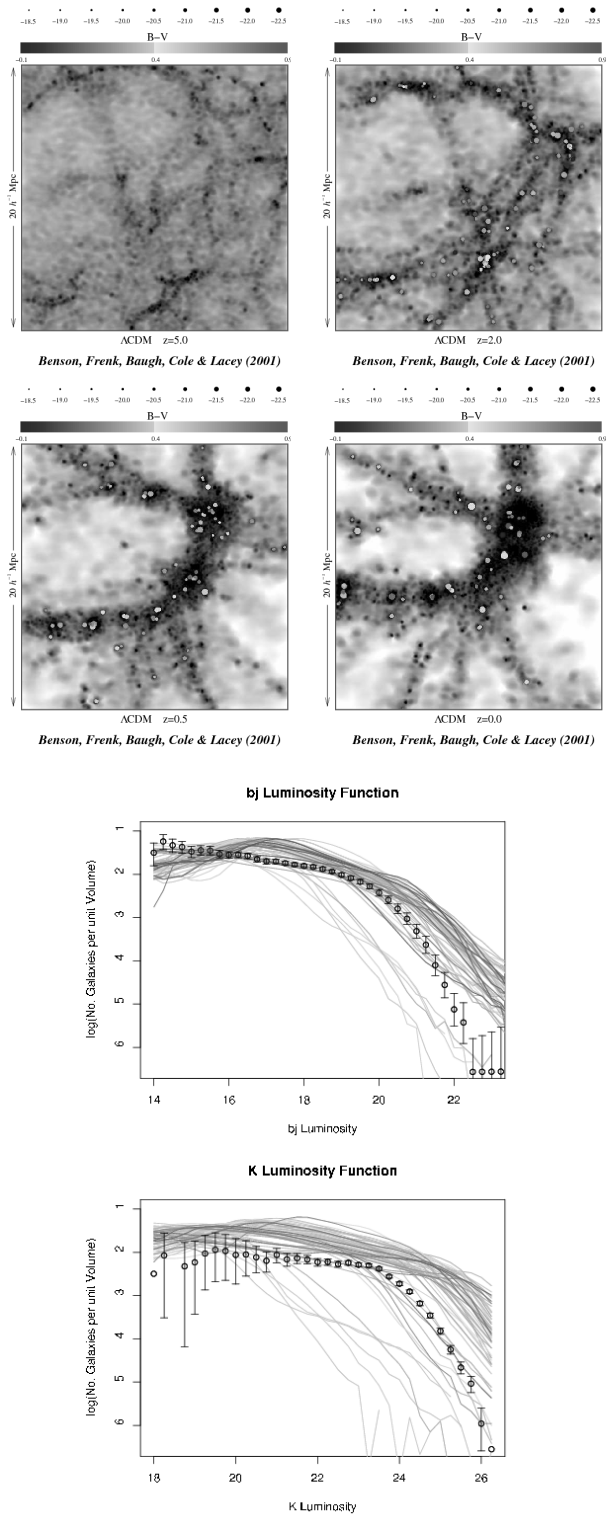

Figure 1: Top 4 panels: the evolution of both the Dark Matter Simulation and Galform over a 13 billion year period. Darker areas show higher concentrations of Dark Matter, leading to the formation of bright galaxies (the white dots). Bottom 2 panels: the bj and K luminosity functions. The grey lines are from 60 runs of the Galform simulation. The black points are observed data from the 2dFGRS survey with associated measurement errors.

Galform has 17 input parameters that the cosmologists were interested in varying. Due to expert judgements regarding the impact of these inputs on the luminosity functions we attempted to calibrate Galform over only 8 of the input parameters (while taking into account the possible effects of the remaining 9). These input parameters and their initial ranges are:

vhotdisk:	100 - 550
aReheat:	0.2 - 1.2
alphacool:	0.2 - 1.2
vhotburst:	100 - 550
epsilonStar:	0.001 - 0.1
stabledisk:	0.65 - 0.95
alphahot:	2 - 3.7
yield:	0.02 - 0.05

The other 9 parameters are: VCUT, ZCUT, alphastar, tau0mrg, fellip, fburst, FSMBH, epsilonSMBHEddington and tdisk.

2.4 Galaxy Formation: Main Issues

The main physical questions that the cosmologists are interested in are: do we understand how galaxies form, and could the galaxies we observe have been formed in the presence of large amounts of dark matter? In order to answer these questions it is vital to correctly analyse all relevant sources of uncertainty within this situation. Many of the sources of uncertainty derive from aspects of the problem for which we have a good physical understanding, for example, the various types of measurement error associated with the observational data (which mainly come from optical deficiencies of telescopes).

However, by far the most important uncertainties arise from the fact that we are uncertain about the discrepancy between the Galform model and the real system, and we are also uncertain about which choice of input should be made when running the model.

3 Bayes Linear Analysis for Computer Simulators

To understand and describe all the sources of uncertainty in the Galform simulator we apply computer model emulation techniques. Although here we will only discuss the Galform simulator, these techniques are very general and can be applied to any complex model of a physical system. Indeed they have been successfully applied to a wide variety of physical models (see [5] for a Bayes Linear approach, [4] for a fully Bayesian approach, and for an overview of computer experiments in general see [6] or the Managing Uncertainty in Complex Models website http://mucm.group.shef.ac.uk/index.html).

3.1 Main Objectives

A common aim of computer experiment analysis is to use observed data to reduce uncertainty about possible choices of the input parameters x (see [5] and [4]). In many problems the major interest lies in whether there is any choice of x that would lead to an acceptable match between model outputs and observed data. The larger the assessed discrepancy between model and system, the weaker the constraints the observations will impose on this choice. In this work we treat this discrepancy as imprecise. Therefore one of the most important aspects of the analysis of the model lies in identifying and quantifying the impact of such imprecision on the choice of possible input values.

3.2 Computer simulators

The simulator (Galform) is represented as a function, which maps the input parameters x to the outputs $f(x)$. We use the "Best Input Approach", where we assume there exists a value x^* independent of the function f such that the value $f^* = f(x^*)$ summarises all the information the simulator conveys about the system. In order to make meaningful statements about the system, denoted y, in relation to the model, we link the simulator to the system using the *model discrepancy* denoted ϵ_{md} via the equation:

$$y = f^* + \epsilon_{md}, \tag{1}$$

and assume that ϵ_{md} is independent of f and x^*, that is, independent in terms of our own beliefs.

The Model Discrepancy term ϵ_{md} links the real system y to the best evaluation of the model represented by f^*. This is distinct from other sources of uncertainty in our analysis and comes directly from expert opinion regarding the 'accuracy' of the model. Understanding the nature of ϵ_{md} is a non-trivial task as there are various other sources of uncertainty that are present that interfere with any assessment of ϵ_{md}. For example, we can never measure the real system y directly. Instead we have measurements z observed with experimental error ϵ_{obs} which are linked to the system by:

$$z = y + \epsilon_{obs}. \tag{2}$$

Another important source of uncertainty is due to lack of knowledge about the form of the function $f(x)$. As the model takes a significant time to run and has a high dimensional input space we only have limited knowledge about its behavior. Further, there is uncertainty regarding the best input value of x^* that features in the definition of ϵ_{md} (equation (1)).

These other types of uncertainty make understanding ϵ_{md} difficult, which is a significant problem as often ϵ_{md} is the most important source of uncertainty due to its size and nature. Due to these difficulties, the expert will often be imprecise over the assessment of the model discrepancy, and even more imprecision could occur when we consider the opinions of a group of experts. It is therefore reasonable to analyse ϵ_{md} within an imprecise framework, while treating other less significant (and more understood) sources of uncertainty as precise.

We need to understand the behavior of the Galform simulation $f(x)$: this is done by representing our beliefs about $f(x)$ as a statistical function known as an Emulator, described in the next section. We address the calibration problem (that of finding inputs x that give rise to good matches between the outputs of $f(x)$ and the observed data z) by use of a technique known as History Matching [5]. This involves discarding regions of the input parameter space that we are reasonably sure will give bad fits to the observed data, and we do this using an Implausibility measure. Analysing the effect on this measure of having an imprecise Model Discrepancy ϵ_{md} (and the corresponding effect on the History Match) is the main goal of this work.

3.3 Representing beliefs about f using emulators

An *emulator* is a stochastic belief specification for a deterministic function. This would be constructed after performing a large, space filling set of runs of the model [6]. Our emulator for component i of f is given by:

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

where $B = \{\beta_{ij}\}$ are unknown scalars, g_{ij} are known deterministic functions of x, and $u(x)$ is a weakly stationary stochastic process. A simple specification is to suppose, for each x, that $u_i(x)$ has zero mean with constant variance and $\mathrm{Corr}(u_i(x), u_i(x'))$ is a function of $\|x - x'\|$. From the emulator, we may extract the mean, variance and covariance for the function, at each input value x.

$$\mu_i(x) = \mathsf{E}[f_i(x)], \qquad \kappa_i(x, x') = \mathrm{Cov}(f_i(x), f_i(x'))$$

Often, because of the mode of construction, the expectation of the emulator interpolates known runs of the model, while the variance represents uncertainty of the function at x inputs that have not been run. A key feature of an emulator is that it is (in most cases) several orders of magnitude faster to evaluate than the model itself. This is important as we will be exploring high dimensional input spaces that necessitate large numbers of evaluations. Emulator techniques are vital in the analysis of any model that has a moderate/long run time and a high dimensional input space.

3.4 Bayes Linear approach

For large scale problems involving computer models, a full Bayes analysis is hard for the following reasons. Firstly, it is very difficult to give a meaningful full prior probability specification over high dimensional input spaces. Secondly, the computations for learning from both observed data and runs of the model, and choosing informative runs, may be technically very challenging. Thirdly, in such computer model problems, often the likelihood surface is extremely complicated, and therefore any full Bayes calculation may be extremely non-robust. However, the idea of the Bayesian approach, namely capturing our expert prior judgements in stochastic form and modifying them by appropriate rules given observations, is conceptually appropriate.

The Bayes Linear approach is (relatively) simple in terms of belief specification and analysis, as it is based only on the mean, variance and covariance specification which, following de Finetti, we take as primitive. It also allows a relatively straightforward description of imprecision which is vital for this work.

We replace Bayes Theorem (which deals with probability distributions) by the Bayes Linear adjustment which is the appropriate updating rule for expectations and variances. The Bayes Linear adjustment of the mean and the variance of y given z is:

$$\begin{aligned} \mathsf{E}_z[y] &= \mathsf{E}[y] + \mathrm{Cov}(y, z)\mathrm{Var}(z)^{-1}(z - \mathsf{E}[z]), \\ \mathsf{Var}_z[y] &= \mathrm{Var}(y) - \mathrm{Cov}(y, z)\mathrm{Var}(z)^{-1}\mathrm{Cov}(z, y) \end{aligned}$$

$\mathsf{E}_z[y]$, $\mathsf{Var}_z[y]$ are the expectation and variance for y adjusted by z.

The Bayes linear adjustment may be viewed as an approximation to a full Bayes analysis, or more fundamentally as the "appropriate" analysis given a partial specification based on expectation (with methodology for modelling, interpretation and diagnostic analysis). For more details see [2].

3.5 History Matching using Implausibility Measures.

We can now use the emulator, the model discrepancy and the measurement errors to calculate a Univariate Implausibility Measure, at any input parameter point x, for each component i of the computer model $f(x)$.

This is given by:

$$I^2_{(i)}(x) = |\mathsf{E}[f_i(x)] - z_i|^2 / \mathrm{Var}(f_i(x) - z_i) \qquad (3)$$

which now becomes:

$$I^2_{(i)}(x) = |\mathsf{E}[f_i(x)] - z_i|^2 / (\mathrm{Var}(f_i(x)) + IMD + OE) \qquad (4)$$

where $\mathsf{E}[f_i(x)]$ and $\mathrm{Var}(f_i(x))$ are the emulator expectation and variance, z_i are the observed data and $IMD = \mathrm{Var}(\epsilon_{md})$ and OE are the (univariate) Imprecise Model Discrepancy variance and Observational Error variance.

When $I_{(i)}(x)$ is large this implies that, even given all the uncertainties present in the problem, we would be unlikely to obtain a good match between model output and observed data were we to run the model at input x. This means that we can cut down the input space by imposing suitable cutoffs on the implausibility function (a process referred to as History Matching). Regarding the size of $I_{(i)}(x)$, if we assume that for fixed x the appropriate distribution of $(f_i(x^*) - z)$ is unimodal, then we can use the 3σ rule which implies that if $x = x^*$, then $I_{(i)}(x) < 3$ with a probability of approximately 0.95 (even if the distribution is asymmetric). Values higher than 3 would suggest that the point x should be discarded.

It should be noted that since the implausibility relies purely on means and variances (and therefore can be evaluated using Bayes Linear methodology), it is both tractable to calculate and simple to specify and hence to use as a basis of imprecise analysis.

One way to combine these univariate implausibilities is by maximizing over outputs:

$$I_M(x) = \max_i I_{(i)}(x) \qquad (5)$$

Using the above unimodal assumptions, values of $I_M(x)$ of around 3.5 might suggest that x can be discarded, as is discussed in section 4.2.

If we construct a multivariate model discrepancy, then we can define a multivariate Implausibility measure:

$$I^2(x) = (\mathsf{E}[f(x)] - z)^T \mathrm{Var}(f(x) - z)^{-1} (\mathsf{E}[f(x)] - z),$$

which becomes:

$$(\mathsf{E}[f(x)] - z)^T (\mathrm{Var}(f(x)) + IMD + OE)^{-1} (\mathsf{E}[f(x)] - z).$$

Again, large values of $I(x)$ imply that we would be unlikely to obtain a good match between model output and observed data were we to run the model at input x. Choosing a cutoff for $I(x)$ is more complicated. As a simple heuristic, we might choose to compare $I(x)$ with the upper critical value of a χ^2 distribution with degrees of freedom equal to the number of outputs.

4 Application to a Galaxy Formation Simulation

One of the long-term goals of the Galform project is to identify the set of input parameters that give rise to acceptable matches between outputs of the Galform model and observed data. We do this using the History Matching ideas outlined above, the full details of which will be reported elsewhere. Before one can embark on such a process, the imprecise model discrepancy must be constructed, and its impact understood, as we now describe.

We proceed to analyse the Galaxy Formation model Galform using the computer model techniques described above. We choose to examine the mean of the first 40 sub-volumes (following the cosmologists' own attempts to calibrate) and select 11 output points from the bj and K luminosity graphs for use in our analysis, as shown in figure 2.

First, 1000 evaluations of the model were made (also shown in figure 2) using a space filling latin hypercube design across the 8-dimensional input space. These runs were used to construct an emulator for Galform as discussed in section 3.3.

We now describe the imprecise model discrepancy used to capture the cosmologist's assessment of the discrepancy between model and reality, and then go on to examine the imprecise implausibility measures this generates, and their impact on the judgement as to which inputs x are deemed acceptable.

4.1 Imprecise Model Discrepancy

At this stage we need to assess the Model Discrepancy ϵ_{md} related to all 11 outputs of interest. This is obtained from an expert opinion regarding the discrepancy between the model and reality, derived from opinions about potential deficiencies of the model. As this is a difficult assessment to make, an imprecise quantification of the model discrepancy will often be the most realistic representation of such uncertainty.

As we are doing a Bayes Linear analysis we only need to consider the assessment of $\mathsf{E}[\epsilon_{md}]$ and $\mathrm{Var}(\epsilon_{md})$. This is a major benefit of the Bayes Linear approach as we can represent any imprecision by letting some of these quantities vary over specified ranges and can then explore the consequences in the rest of our analysis. This is straightforward in comparison to a fully probabilistic analysis where such an imprecise specification would be extremely difficult, and a subsequent examination of the impact of such imprecision would often be intractable.

A leading expert stated that his beliefs regard-

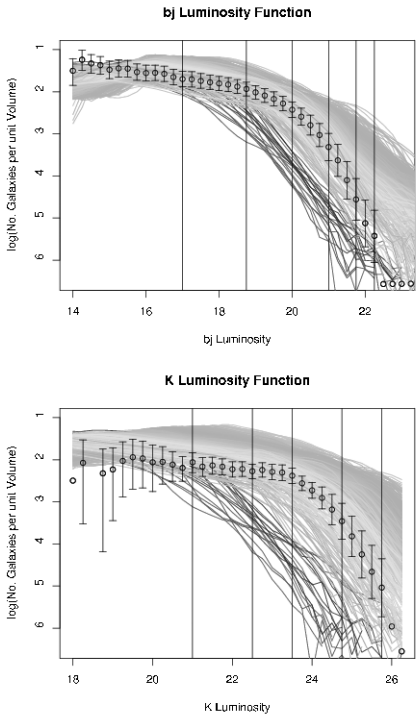

Figure 2: The bj and K luminosity outputs from 1000 runs of the model. The vertical black lines show the 11 outputs chosen for emulation. The error bars now incorporate the (univariate) model discrepancy with $a = \overline{a}$.

ing the model discrepancy were symmetric in that $\mathsf{E}[\epsilon_{md}] = 0$. Define $IMD = \mathrm{Var}(\epsilon_{md})$. Even for the univariate case (i.e. considering only one of the 11 outputs) the individual expert was unwilling to assess the size of *IMD* precisely. However, the expert was willing to make an imprecise assessment by specifying lower and upper bounds $\underline{IMD}$ and $\overline{IMD}$.

For the multivariate case, we needed to assess $IMD = \mathrm{Var}(\epsilon_{md})$ which is now an 11x11 matrix. The structure of this matrix will come from the expert's opinion as to the deficiencies of the model. In the case of Galform there are two major physical defects that can be identified. The first is the possibility that the model has too much (too little) mass in the simulated universe. This would lead to the 11 luminosity outputs all being too high (or too low), and would lead to positive correlation between all outputs in the MD matrix. The second possible defect is that the galaxies might age at the wrong rate leading to more/less blue galaxies and therefore less/more red galaxies. This would be represented as contributing a smaller negative correlation between the bj and K luminosity outputs. To respect the symmetries of these possible defects, the multivariate Imprecise Model Discrepancy (*IMD*) was parameterised in the following form:

$$IMD \;=\; a^2 \begin{pmatrix} 1 & b & .. & c & .. & c \\ b & 1 & .. & c & . & c \\ : & : & : & : & : & : \\ c & .. & c & 1 & b & .. \\ c & .. & c & b & 1 & .. \\ : & : & : & : & : & : \end{pmatrix} \qquad (6)$$

where now a, b and c are imprecise quantities, and we obtain the following expert assessments: $\underline{a} = 3.76 \times 10^{-2}, \overline{a} = 7.52 \times 10^{-2}, \underline{b} = 0.4, \overline{b} = 0.8$, and $\underline{c} = 0.2, \overline{c} = b$.

It is possible to build in far more structure into *IMD* if required. The more detailed the structure, the more difficult eliciting expert information becomes. However, note the relative ease of specifying useful high-level imprecise statements using expectation as primitive, as compared to the corresponding effort for a fully probabilistic analysis. Exploring the effects of these specifications is also an easier task, as we now show by examining the effects of varying choices of a, b and c on the appropriate implausibility measures.

4.2 Implausibility Measures

In section 3.5 we showed how to construct the maximised and multivariate Implausibility measures $I_M(x)$ and $I(x)$. As these are derived using the imprecise model discrepancy we can write the dependence of these two implausibility measures on a, b and c explicitly. We can now explore the effects on $I_M(x, a)$ and $I(x, a, b, c)$ of varying a, b and c within the credal set C defined by:

$$\underline{a} < a < \overline{a}, \;\; \underline{b} < b < \overline{b}, \;\; \underline{c} < c < b,$$

as is described in the next section. As the implausibility measures are now imprecise, in order for regions of the input space x to be discarded as Implausible, they must violate the implausibility cutoff for all values of a, b and c, that is:

$$I(x, a, b, c) \;>\; I_{cut} \quad \forall\, a, b, c \in C, \qquad (7)$$

with a similar relation for $I_M(x, a)$:

$$I_M(x, a) \;>\; I_{Mcut} \quad \forall\, a \in C. \qquad (8)$$

In section 4.3 we set $I_{cut} = 26.75$ corresponding to a critical value of 0.995 from a χ^2 distribution with 11 degrees of freedom (and $I_{Mcut} = 3.5$) which were felt to be appropriate, conservative choices for the cutoffs. Note that if an input x satisfies either constraint (7) or constraint (8) then it is deemed implausible and will be discarded. As can be seen from equations (6),(3)

and (5), $I_M(x, a)$ is a monotonically decreasing function of a and hence constraint (8) will be equivalent to:

$$\min_{a \in C} I_M(x, a) \;=\; I_M(x, \overline{a}) > I_{Mcut} \qquad (9)$$

The constraint for $I(x, a, b, c)$ is more complex and in general no such monotonicity arguments can be used. In a full calibration analysis we would, for fixed x, evaluate $I(x, a, b, c)$ over a large number of points in the credal set C, and only discard the input x if it does not satisfy the implausibility cutoffs for every one of these points. However, here we are more interested in understanding the impact of different choices of a, b and c on the input space, which we do in the next section.

4.3 Effect of the Imprecise Model Discrepancy on the Assessment of the Best Input x^*

The most important effect of an imprecisely specified model discrepancy is its impact upon the choice of acceptable input parameters x^*. Above we showed how to construct the implausibility measures and described their use in deciding which inputs would be deemed acceptable. Here we will explore the impact of the imprecision on the multivariate measure itself, then on the percentage of input space remaining, by analysing the effects of varying a, b and c. Note that while we present all the pictures in greyscale, these displays are designed for presentation in colour.

Figure 3 shows the multivariate implausibility $I(x, a, b, c)$ as a function of a, b and c for two different fixed values of x. In the top (bottom) panel x_7 i.e. alphahot is set to its minimum (maximum) value of 2 (3.7). In both panels x_1 i.e. vhotdisk is at its maximum value of 550, and all other inputs are at their midrange values. In these and subsequent figures we examine slightly larger ranges for a, b and c than are defined by the Credal Set: here they satisfy $0.5\underline{a} < a < 2\overline{a}$, $0 < b < 0.95$ and $0 < c < b$. The top panel shows that $I(x, a, b, c)$ is minimised for large values of a, b and c attaining a minimum of approximately $I(x, a, b, c) = 14.2$. In the bottom panel however, the implausibility is minimised for low values of b and c and only attains a minimum of $I(x, a, b, c) = 38.3$. This shows the dramatically different behaviour of the implausibility measure as a function of a, b and c for two different parts of the input space, and specifically that general monotonicity arguments (such as used in equation (4.2)) cannot be applied to the imprecise parameters b and c. Plots such as those shown in figure 3 are very useful in helping to understand the impact of an imprecise assessment. However, one cannot examine such plots

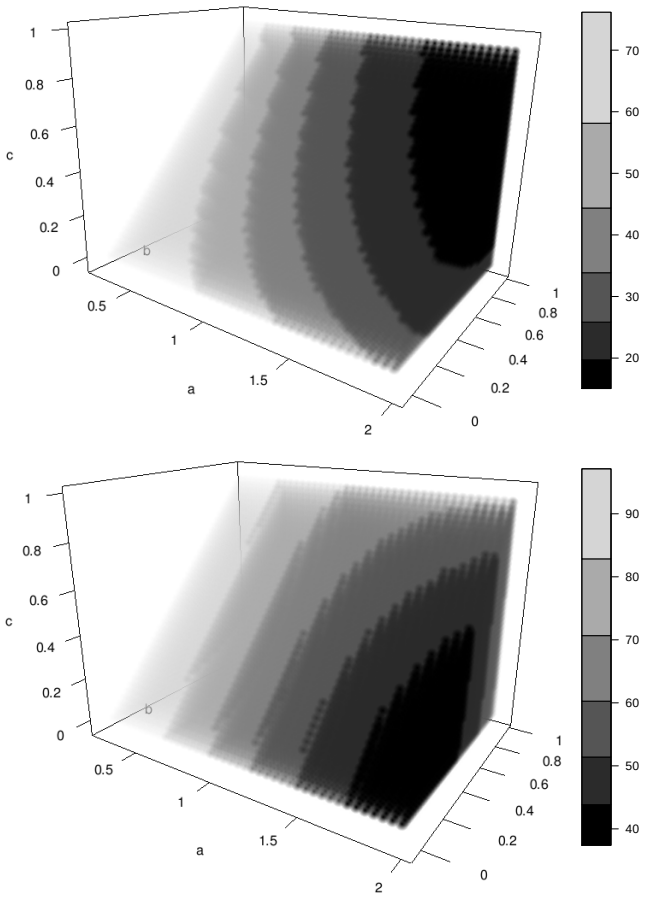

Figure 3: Both panels shows the multivariate implausibility $I(x, a, b, c)$ as a function of a, b and c for two different fixed values of x, with darker colours representing lower implausibility. Here a, b and c vary over the ranges $0.5\underline{a} < a < 2\overline{a}$, $0 < b < 0.95$ and $0 < c < b$. Note that the scale on the a-axis is in terms of multiples of $\overline{a}$. Top panel: vhotdisk = 550, alphahot = 2, Bottom panel: vhotdisk = 550, alphahot = 3.7, all other inputs set to their midrange values.

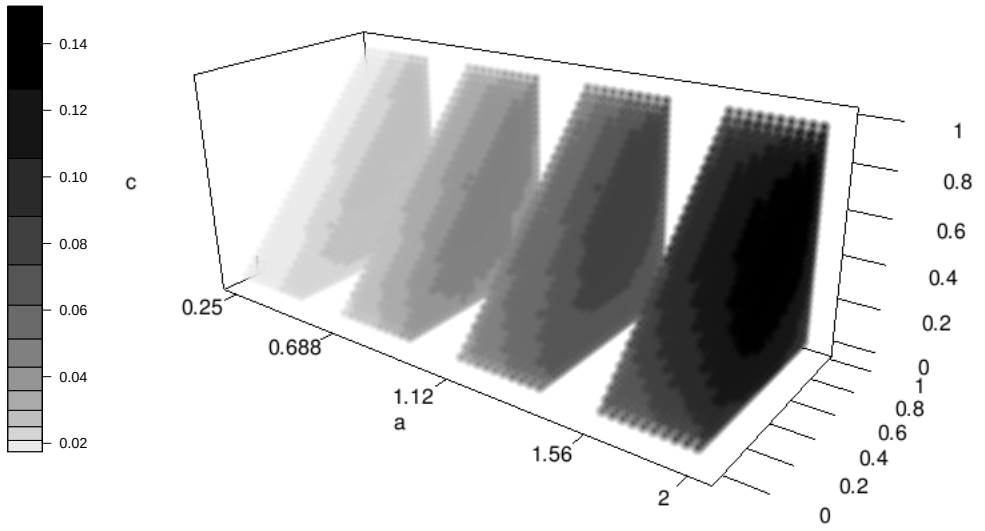

Figure 4: Fraction of input space that survives the multivariate implausibility cutoff given by equation (7) with $I_{cut} = 26.75$ as a function of a, b and c in the ranges $0.5\underline{a} < a < 2\overline{a}$, $0 < b < 0.95$ and $0 < c < b$. Note that the scale on the a-axis is in terms of multiples of $\overline{a}$.

for all points in the 8-dimensional input space. We therefore look at other ways to summarise and visualise the analysis.

We can summarise the effect of the imprecise multivariate implausibility cutoff given by equation (7) on the whole of the input space by looking at the fraction of space remaining once the cutoff has been imposed. Here we display the results corresponding to $I_{cut} = 26.75$, a value which was thought to be a reasonably conservative choice. Figure 4 shows this fraction of space remaining as a function of a, b and c in the ranges $0.5\underline{a} < a < 2\overline{a}$, $0 < b < 0.95$ and $0 < c < b$, with darker colours representing higher fractions. Figure 5 shows the same 3D plot from a different perspective. The 3D object has been cut in 3 places to allow one to see slices of the function at fixed values of a. This shows that for large values of a, the maximum space remaining would occur for intermediate values of b and c (approximately $b = 0.7$ and $c = 0.6$ for $a = 2$), however for smaller values of a the space remaining would be maximised by large b and c (e.g. for $a = 0.5\overline{a} = \underline{a}$, $b = 0.95$ and $c = 0.95$: see figure 5). These plots also suggest that the space remaining is far less sensitive to variation in b and c than in a: it is useful for the expert to know therefore that their assessment for a is more significant than for b and c.

Figure 6 shows the fraction of space remaining as a function of a for fixed choices of b and c. The boundaries of the Credal Set are shown by dotted vertical lines. Again one can see that to maximise the space remaining requires intermediate values of b and c for large a, and large values of b and c for small a. Also note that as a tends to small values, the fraction of space remaining varies only slowly: in fact setting $a = 0$ (which is not shown in this figure) leads to 0.017 of the input space remaining: this is important for the expert to know as it shows that some of the input space would survive the cutoff even for zero model discrepancy.

Examining the space remaining is useful in understand the effects of the imprecise specification of model discrepancy. However, it is also vital to assess the effect on the input space directly i.e. to determine which inputs x would not be discarded due to the imprecise specification. One way to analyze this is to ask what is the minimum value of a that is required to ensure that a particular input point x satisfies the implausibility cutoff. Figure 7 shows 3D plots of the required value of a as a function of the input parameters x_1 and x_7, and of b (with the other inputs at their midrange values , with $c = 0$, and the key in terms of multiples of $\overline{a}$). The darkest areas are those that have a required a of less than $\overline{a}$ and hence would survive the cutoff for the current specification. These plots show that while the value of b has effects in some parts of the input space, the region defined by required $a < \overline{a}$ is relatively independent of the value of b (a similar result is seen for plots with varying c and fixed b). This demonstrates that the required value of a is far more sensitive to the value of x_1 and x_7 as opposed to the specified range of the imprecise quantity b, and gives more evidence to suggest that the experts assessment

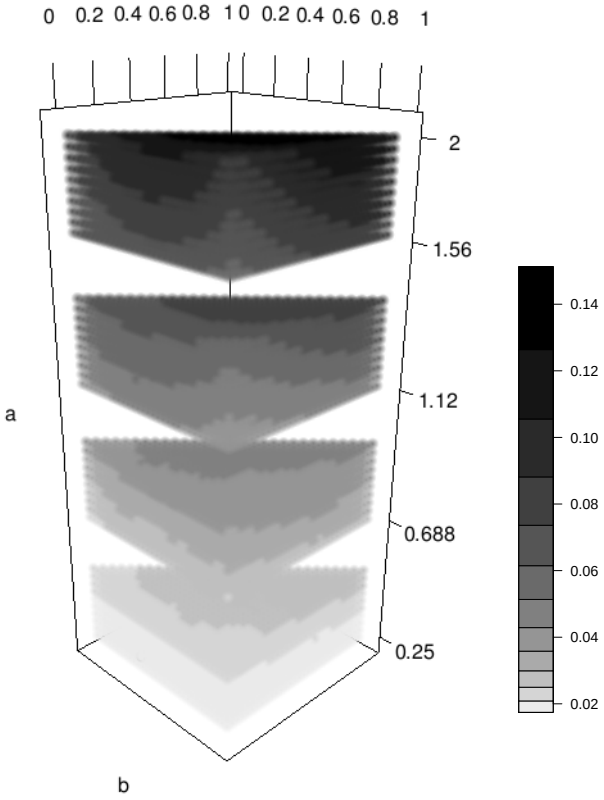

Figure 5: Alternative view of fraction of input space that survives the multivariate implausibility cutoff given by equation (7) with $I_{cut} = 26.75$ as a function of a, b and c in the ranges $0.5\underline{a} < a < 2\overline{a}$, $0 < b < 0.95$ and $0 < c < b$. Note that the scale on the a-axis is in terms of multiples of $\overline{a}$.

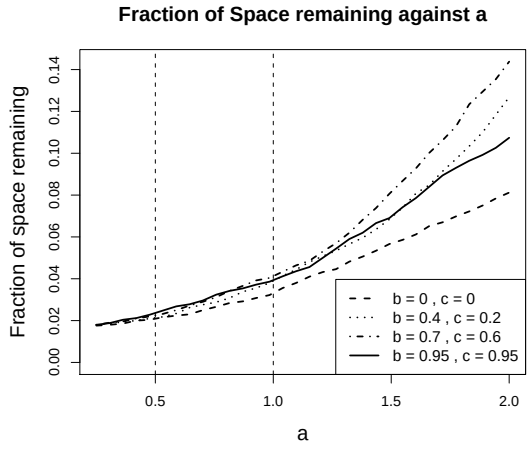

Figure 6: Fraction of input space that survives the multivariate implausibility cutoff given by equation (7) with $I_{cut} = 26.75$ as a function of a for the range $0.25\overline{a} < a < 2\overline{a}$, for various choices of b and c. The scale on the a axis is in terms of multiples of $\overline{a}$. Note that $\underline{a} = 0.5\overline{a}$. It can be seen that more space survives when $b = 0.7$ and $c = 0.6$ for large a, however, for smaller a the more extreme values $b = 0.95$ and $c = 0.95$ are preferred (which are not in the Credal Set).

for a is far more significant than that for b and c.

We have seen the effects of the imprecise assessment on the multivariate implausibility measure $I(x, a, b, c)$, on the fraction of space remaining after the cutoff is imposed, and on the set of allowed values of x_1 and x_7. We showed that these effects are non-trivial as the multivariate implausibility measure is a complicated function of x, a, b and c.

5 Conclusions

We have discussed how computer models make imprecise statements about physical systems. This imprecision arises due to the immense difficulty in giving a precise quantification on the discrepancy between the model analysis and the system. We have shown how use of Bayes Linear methods can provide a relatively straightforward description of this imprecision, allowing a meaningful elicitation of imprecise model discrepancy while leading to a tractable analysis of the issues involved in computer model calibration, which we demonstrated in the context of the galaxy formation simulation Galform.

The mathematical tractability of treating expectation as primitive also allows a detailed study of the effects of such imprecise assessments. In this case this in-

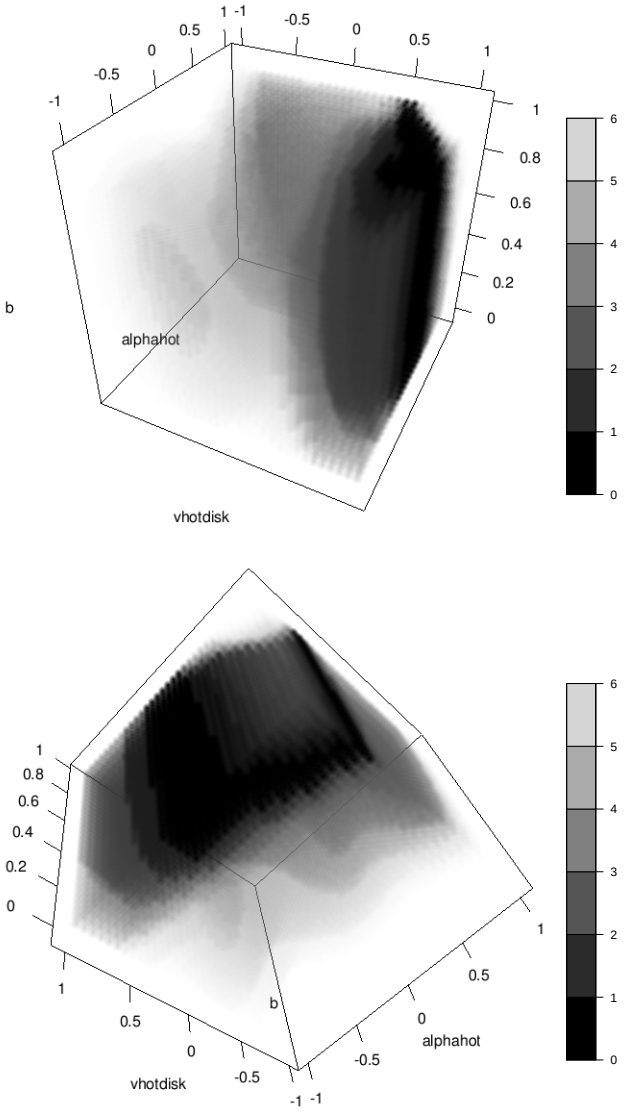

Figure 7: Plots showing the value of a that is required to ensure a point in input space satisfies the multivariate cutoff, as a function of the input parameters vhotdisk and alphahot, and of the imprecise quantity b (with c set to 0). The key is in terms of multiples of $\overline{a}$, and the darker areas represent low required a. All other input parameters have been set to their midrange values.

volved understanding the impact of the imprecision on the implausibility measures; measures that were used to discard regions on input parameter space thought to be very unlikely to give rise to acceptable matches between model output and observed data. In this way we were able to show the direct impact on parts of the input space of the expert's imprecise judgements regarding model deficiency. The effects of the imprecise assessments were found to be non-trivial and a variety of methods were used to summarise the data in order to produce meaningful visual representations of such effects.

Acknowledgements

This paper was produced with the support of the Basic Technology initiative as part of the Managing Uncertainty for Complex Models project, and with EPSRC funding through a mobility fellowship. We would like to thank Prof Richard Bower (Institute for Computational Cosmology, Physics Department, Durham University) for providing the expert assessments that feature in this work. We would also like to thank Prof Richard Bower and the Galfrom group (also based at the Institute for Computational Cosmology, Physics Department, Durham University) for access to the Galform model and to their computer resources.

References

[1] M. Goldstein and J.C.Rougier (2008). Reified Bayesian modelling and inference for physical systems (with discussion), *JSPI*, to appear, .

[2] Goldstein, M.,Wooff, D. (2007). Bayes Linear Statistics: Theory and Methods. *Wiley*

[3] Bower, R.G., Benson, A. J. et.al.(2006). The Broken hierarchy of galaxy formation, *Mon.Not.Roy.Astron.Soc.* 370, 645-655

[4] Kennedy, M.C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society*, B,63, 425-464

[5] P.S. Craig, M. Goldstein, A.H. Seheult, J.A. Smith (1997). Pressure matching for hydocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments (with discussion), in *Case Studies in Bayesian Statistics*, vol. III, eds. C. Gastonis et al. 37-93. Springer-Verlag.

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