Interrogation Theory

SUMMARY

The goal of an investigation, scientific or otherwise, is usually to find answers to some specific set of questions about the state of nature: what is the seismic velocity structure? How likely is this volcano to erupt within a certain period? Does a subsurface reservoir contain resources of interest? Background research may reveal the existence of pertinent knowledge and information discovered previously, new data are normally acquired, and an inference problem is solved in order to answer the questions taking both all of the a priori information and the new data into account. Inverse theory, decision theory and the theory of experimental design provide methods to optimise the design of the investigation and to estimate results. However, those theories are normally set in the context of a particular model of the universe, with its particular parameterisation. This requires the investigator to specify a priori a coherent utility (a function that describes the risks and rewards) of all possible outcomes under that parameterisation. Quite commonly, the investigator may not be able to do this.

Ideally an investigator would be able merely to pose a set of questions, define a set of constraints on the data types, acquisition costs and logistics, and provide a functional to relate the questions to any particular parameter space. Theory and methodology would then semi-autonomously drive the interrogation of the state of nature by optimally selecting one or more relevant models and parameter spaces, and designing, acquiring and analysing data, in order to best answer the questions. If necessary this could be done in a sequential or iterative manner, which potentially then involves changing the questions posed in each iteration based both on previous results and on inspiration from the investigator. We present such a theory of interrogation in this paper.

We review the relevant aspects of decision and design theory, and cast them in a framework where the investigator specifies a utility only at the level required by the general questions to be posed. Each model under consideration is then mapped into this utility space of possible answers. We then extend this framework to sequential investigations, where the outcome of each step may affect all aspects of the problem: the models entertained, the utilities, and even the questions themselves.

A variety of examples illustrates the generality of this method: an asset team investigating how best to exploit a subsurface reservoir, Monte Carlo sampling to estimate the Bayesian evidence for geophysical models, discriminating between different rock physics models of strain in laboratory deformation experiments, an organisation sequentially assessing the effectiveness of its methods to evaluate subsurface assets, assessing whether subsurface CO_2 storage should be promoted for climate change mitigation, and examples running through the text of seismic tomography, earthquake characterisation, and autonomous interplanetary robotic exploration.

Key words: Decision Theory – Optimal Design – Risk – Utility – Inference – Bayesian

submitted to *Geophys. J. Int.*

Interrogation Theory

Richard Arnold[1,2⋆] and Andrew Curtis[2,3]

[1] *School of Mathematics and Statistics, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand. E-mail: rich*

[2] *School of Geosciences, Edinburgh University, Grant Institute, West Mains Road, Edinburgh EH9 3JW, United Kingdom. E-mail: a*

[3] *Institute of Geophysics, ETH Zurich, Zurich, Switzerland*

1 INTRODUCTION

In an investigation an individual researcher, or group of researchers usually wishes to find the answer to a set of questions. These questions are often formulated in plain, non-mathematical language: 'How much methane is in this gas field?', 'When will this volcano next erupt?', 'How likely is it that atmospheric CO_2 will pass a given threshold?', 'How should one best assess the properties of particular structures in the Earth's subsurface?' The answers may be found by experimentation, consultation of records, consultations with experts, or some other data collection exercise. Subsequent interpretation of this information is then an exercise in inference in its broadest sense, and usually involves solving inverse problems. Figure 1(a) shows the classical schema for forward and inverse problems, which incorporates the generation of synthetic data from a model of the universe (the forward problem), and inference about parameters in the model from recorded data (the inverse problem) (Snieder 1998).

The answering of a question requires an experimental design, again in its broadest sense: the selection of a method of inquiry which will result in new information, germane to the question at hand. In such investigations the well established theory of statistical experimental design (e.g., Chaloner & Verdinelli 1995; Myung & Pitt 2009; Curtis 2004a,b; Maurer et al. 2010) and theory of Bayesian or statistical inference (e.g., Tarantola 2005; Vehtari & Ojanen 2012) provide machinery which in theory deliver the optimal experimental design, and procedures for efficient estimation and inference. However in order for this machinery to function, the questioner must – explicitly or implicitly – specify models of the system under consideration, quantify prior knowledge, identify the space of possible experimental designs, and quantify in the form of a utility function the risks or rewards associated with drawing all possible conclusions in all possible worlds.

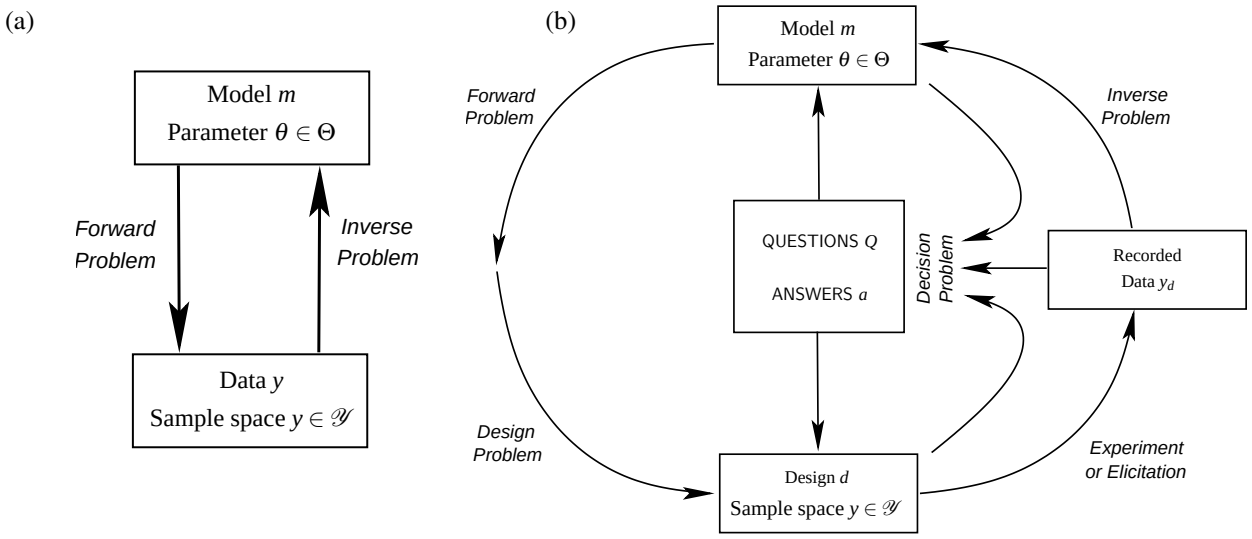

Figure 1. (a) Classical schema for solving Inverse Problems: The data y are assumed to come from a model m, parameterised by θ. The inverse problem is the estimation of the set of parameter values θ that will generate data in the forward problem that best match the observed data. (b) Schema for *Interrogation Problems:* given a set of scientific questions Q and a model m or range of models (Sambridge et al. 2006), an experimental design is selected to generate data y_d that best constrain the model parameters θ and thereby answer the questions. Feedback occurs through the modification of the questions in the light of the answers, and the process can repeat. This schema is embodied algorithmically in Figure 2.

These requirements may prove very demanding for investigators. Specification of models and prior information may require the elicitation of expert opinion, and likewise expert knowledge may be required to identify the set of feasible experimental designs, among which the optimal design can be sought. Moreover, the specification of utilities, which quantify the consequences of different conclusions and which are needed to decide between outcomes using statistical decision theory, often requires the construction of high dimensional functions in Geophysical problems. Some of the properties of these complex utilities may be significant, unintended, but go unnoticed because of their complexity or dimensionality (see e.g. Curtis & Lomax 2001). Thus while much is known about optimal procedures in the context of a fully specified decision or estimation problem, little attention has been given to the optimality of that specification.

Furthermore, it may be that the outcome of an inquiry is the realisation that there are new questions that are relevant, questions which could not have been anticipated at the outset. In such circumstances the inquiry continues, possibly for many iterations, with new questions added and old questions discarded at each step. Each step of such a multi-stage investigation adds to the investigator's knowledge and may stimulate new directions of inquiry.

Our goal in this paper is to synthesise an overarching methodology for what we call an **inter-**

rogation. An interrogation is a potentially open ended investigation, initiated by a set of high level questions. We seek a framework for such an inquiry, which allows for the iterative updating of knowledge and the generation of new questions. A specific requirement we have of such a procedure is that answers be provided in the same high level terms that they are asked, and that investigators only be required to state their preferences and utilities at that same level. An updated schema for such problems is shown in Figure 1(b).

An example motivation for seeking such a framework comes from the Earth resources industry, and in particular from the complex investigations and decision making that take place routinely concerning subsurface reservoirs of fluids contained within the pore space of rock. To make this concrete, consider a depleted gas field managed by a commercial organisation. This organisation may be interested in deciding whether or not to use the field for carbon storage, i.e., storing the carbon dioxide (CO_2) produced by burning oil, gas or coal for electricity production, in a subsurface reservoir at the field in order to mitigate against anthropogenic atmospheric CO_2-related climate change.

Decisions regarding the gas field are made by an asset team, comprising experts of various kinds (geologists, geophysicists, engineers, business managers and field/logistics managers). A sequence of questions are relevant to the decision regarding whether to develop a carbon storage reservoir. At a high level there is the question of whether a potentially secure storage reservoir exists. This requires the team to establish a range of likely scenarios beneath the ground that involve answering lower level questions: What are the subsurface rock types? What are their porosities (the amount of space available for fluids)? A relevant sub-question is usually, what is the depositional origin of the rock (was it deposited in an old tidal delta or in a lake? Around a reef front or in a lagoon?) as this partly controls the reservoir rock properties that would be expected. Is there a cap rock above the reservoir that would form a barrier to contain the carbon dioxide, thus preventing escape upwards into the atmosphere?

At their meetings the team has access to subsurface data, seismic surveys, geological models, and various process models which can generate scenarios given likely sets of inputs. The data can be consulted, displayed, discussed, and new scenarios generated. Ultimately the asset team needs to choose answers to the above questions, and thus make a decision about suitable actions to be taken. Examples of the decision process and flows of information that take place in such asset team discussions are analysed in the sequence of papers by Polson & Curtis (2010); Polson et al. (2012b,a) and Polson & Curtis (2015).

The process of determining the answer to the highest level question is by nature iterative, and requires input from data, modelling, and the use of expert opinion. If questions cannot be resolved at the meeting, then the team may decide to consult other experts, or to collect further data by exploratory drilling, or by conducting a seismic survey (a subsurface imaging experiment). These data collec-

tion exercises may take place simultaneously without reference to each other, and may use different parameterisations of models or of the information obtained. Thereafter, information from all parameterisations must be reconciled and integrated within the subsequent decision-making process. From the point of view of the company, the utility of such a decision made by the asset team is ultimately determined by its cost and potential benefits, so the accurate assigning of costs and estimation of those benefits to different courses of action is an important part of the process.

This example demonstrates the key components of an interrogation: an inquiry with an invariant goal, embodied in high level questions which require one or more data collection exercises to provide answers, connected to a decision-making process (which may simply be to decide which questions to ask next – an experimental design exercise), with different costs and benefits associated with different decisions made. Other examples include geophysical investigations of new areas of the Earth which typically begin with a sparse survey to answer the question of whether suitable targets of interest exist, followed by later iterations that deploy surveys designed to focus information on previously located targets, or deploy more sophisticated data processing to extract more information from existing data. In each iteration, since the result is initially unknown, it is difficult a priori to fix a suitable parameterisation, utility function, and hence to design or decide what the next best course of action would be. Still other examples include the design of semi-autonomous robots for interplanetary exploration (e.g., the ”Mars Rover”) that must iteratively identify targets and find ways to approach them, iterative elicitation of human expert opinion, interrogation of witnesses or suspects, and many other examples of standard, iterative, hypothesis forming and testing in scientific investigations. Some of these examples are revisited below.

In this paper we review all of the components of an interrogation, several of which are displayed in Figure 1(b). Our approach draws on results from Bayesian inverse theory, optimal experimental design theory, and from the decision theoretic approach to inference. We also rely on the existence of established methods of model appraisal and selection (e.g., Snieder 1998; Burnham & Anderson 2002).

The novelty of our approach is the explicit consideration of goals, questions and answers as the motivators of investigations, and in particular the dynamic nature of investigators' questions. The scope of what we describe here is admittedly large: we are proposing an overarching framework for decision making in any setting. However in geoscience and other industries the resources committed to decisions are in many cases so large that the current absence of a formal end-to-end theory for decision making poses an apparent risk of wasteage of the substantial resources deployed.

We are aware of one paper which explicitly models the question–answer sequence in scientific enquiry to the same level of dynamism as in our methodology, namely the paper by Brockmann &

Dawkins (1979) on strategies of the digging wasp *Sphex ichneumoneus*. The layout of their paper, explained in detail in Dawkins (2015) (pp51-82), mirrors the sequence of inquiries they carried out. Each question is motivated by the answer to the question preceding it, and the latter questions could not have been anticipated at the outset. Our work builds on these descriptive papers by providing the first accompanying mathematical formalism with which the overall interrogation process can be quantitatively modelled, analysed and designed.

Our formulation of interrogations is sufficiently broad that it captures the highly practical problems of designed experiments as well as the more high level conceptual approaches to human decision making. When making decisions in the real world we must at some level contemplate all possible contingencies, and make optimal choices among all possible courses of action. Such a comprehensive approach is of course impractical, but as we show below it is theoretically valuable to envision it, even as we make appropriately severe restrictions to smaller sets of possible models for the world, and possible courses of action. The examples we choose to demonstrate our approach below show that an investigator can start with a high level conceptual view and still find practical implementation.

In Section 2 we define more precisely the components of an interrogation problem. This includes some existing results from decision theory, recast in our notation and setting. We present a worked example in Section 3. Sequential interrogations are described in Section 4, and a discussion follows in Section 5. Our notation is summarised in Table 1.

2 INTERROGATION PROBLEMS

A schematic diagram of an interrogation problem is shown in Figure 2, and in this section we introduce and explain each of the components of the problem, and establish our notation. The 'investigator' referred to in this section may typically be an individual in Geophysical investigations, but might otherwise be a team of experts, a company, a government, or any entity capable of posing questions.

2.1 Components of an Interrogation Problem

An investigator has prior knowledge B, and questions about nature Q, the characteristics of which are formulated further below. There is a set of possible answers $\mathbb{A}$ among which a choice will be made. In standard descision theory these answers would be referred to as 'actions' or 'decisions', as the conclusions of an inference procedure which follows the observation of new data (see e.g. Young & Smith 2005, Ch. 2). The precise nature of an answer of course depends on the question being asked. An answer may be the estimate of a particular parameter of interest (e.g., the seismic velocity structure of the Earth, or the capacity of a putative subsurface reservoir), in which case the answer a is a single

Table 1. Notation used for interrogation theory. Additional notation introduced in worked examples has been omitted so as to clearly distinguish the fundamental entities.

Symbol	Description
$\mathbb{M}$	Space of models
m	Model; element of the space of models $\mathbb{M}$
$p(m)$	Prior probability of models in $\mathbb{M}$
$\mathbb{A}$	Answer space
a	Answer; element of answer space $\mathbb{A}$
$a^*(y_d, d)$	Optimal answer given data $y_d \in \mathcal{Y}_d$ collected under design $d \in \mathbb{D}$
$\mathbb{D}$	Space of Experimental Designs
d	Experimental Design; element of design space $\mathbb{D}$
d^*	A priori optimal design
Θ_m	Parameter space of model $m \in \mathbb{M}$
θ_m	Element of Θ_m, the parameter space of model m
$p(\theta_m\|m)$	Prior over parameter space Θ_m for parameters of model m
$\mathcal{Y}_d$	Sample space of observations under design $d \in \mathbb{D}$
y_d	Observation; element of sample space $\mathcal{Y}_d$ observed under design $d \in \mathbb{D}$
$f(y_d\|d, \theta_m, m)$	Statistical likelihood: probability distribution of data y_d under design d if the state of nature is model m with parameter value θ_m
B	State of knowledge
Q	Questions about the state of nature
$T(\theta_m\|m, Q)$	Target function, summarising the state of nature described by model m with parameter values θ_m, relevant to questions Q. Takes values in the Target space $\mathbb{T}$
$\mathbb{T}$	Target space $\mathbb{T}$ of values of $T(\theta_m\|m, Q)$
$U(a\|\theta_m, m, y_d, d)$	Utility of answer $a \in \mathbb{A}$ if the true model is $m \in \mathbb{M}$ with parameter values $\theta_m \in \Theta_m$, and data $y_d \in \mathcal{Y}_d$ are observed under experimental design $d \in \mathbb{D}$
$U(a\|t, d)$	Utility of answer $a \in \mathbb{A}$ if the summarised state of nature is $t \in \mathbb{T}$ and the design is $d \in \mathbb{D}$
$U_p(a\|y_d, d)$	Posterior expected utility of answer $a \in \mathbb{A}$ given data $y_d \in \mathcal{Y}_d$ observed under experimental design $d \in \mathbb{D}$
$U^*(y_d, d)$	Posterior expected utility for the optimal answer $a^*(y_d, d)$ given data $y_d \in \mathcal{Y}_d$ observed under experimental design $d \in \mathbb{D}$
$U^*(d)$	A priori expected optimised utility for experimental design $d \in \mathbb{D}$

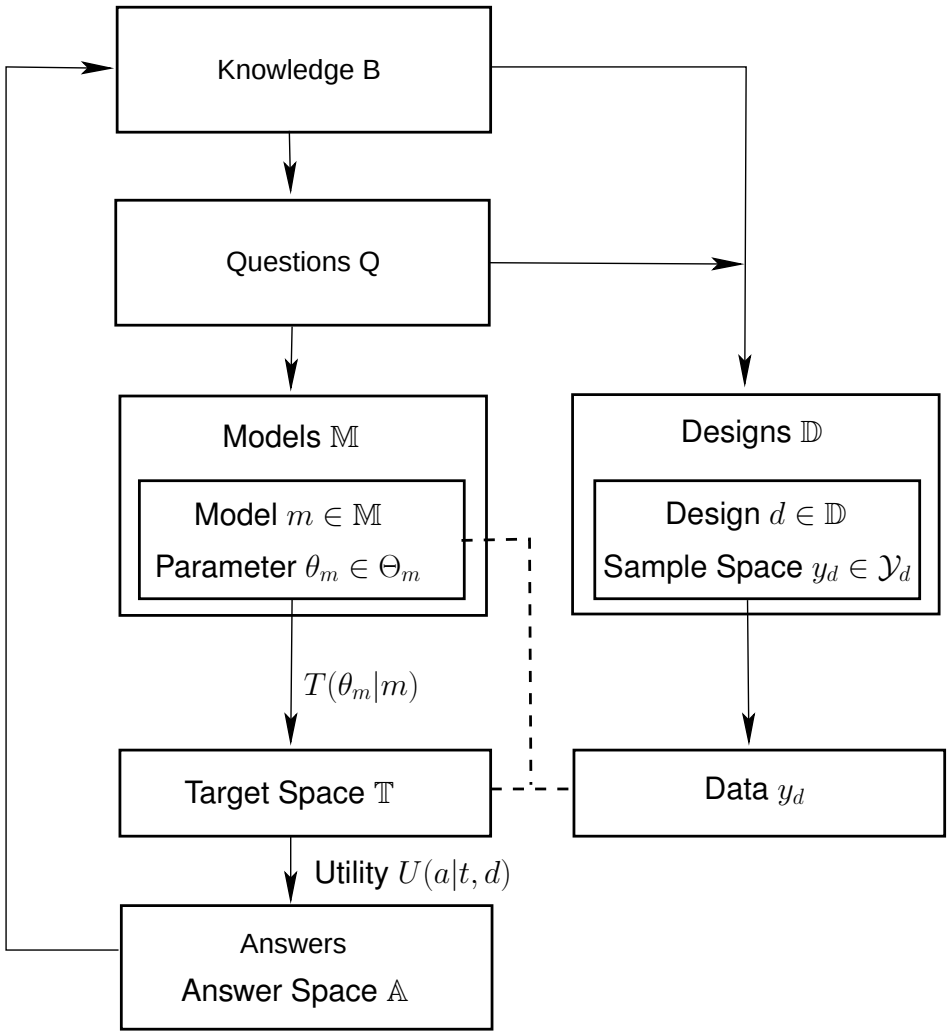

Figure 2. Algorithmic schema of an interrogation problem. See text for details.

value or vector of values. Alternatively the answer may be the estimate of the entire distribution of possible values of a parameter, and a is then a probability mass or density function. In the case of an exercise which results in a choice being made amongst a set of options (e.g. to establish a carbon storage reservoir, or not) then the answer a is the label of the option that is chosen. Equivalently when choosing among a set of options the answer may be a vector with binary (0/1) entries, one for each option, with exactly one entry being 1 and all others zero. Thus a can take many different forms, and it is this variety of definition that provides the flexibility to represent a spectrum of interrogation problems within a single schema and methodology.

The **space of models** $\mathbb{M}$ is the (countable) set of all models of nature that are deemed relevant to the investigation by the investigator. Here the term 'model' is used in a mathematical sense, to represent a relationship between observed data and the parameter of the model. For example in earthquake source characterisation we might use a model $m_1 \in \mathbb{M}$ based on a simple one-dimensional representation of

seismic wave velocities in the Earth – with velocity dependent only on depth – and also an alternative model $m_2 \in \mathbb{M}$ that embodies a full three dimensional velocity structure. We note that in practical settings such mathematical models can be, and usually are, implemented in some kind of software.

The content of $\mathbb{M}$ therefore depends on the investigator's prior knowledge B, and on the questions Q, and is assumed to be rich enough that some element of $\mathbb{M}$, or some average over its elements, provides a sufficiently accurate description of nature.

Each element $m \in \mathbb{M}$ has an associated parameter value θ_m in a **parameter space** Θ_m. The parameter θ_m may be multidimensional (e.g., to describe the subsurface seismic velocity structure, or earthquake source characteristics), and may have discrete and continuous components. In this paper we treat θ_m as continuous, but all of our results hold if it is discrete, or has discrete components, in which case integrals $\int_{\Theta_m} \ldots \mathrm{d}\theta_m$ can be replaced by sums $\sum_{\theta_m \in \Theta_m}$. The investigator has a prior distribution on the space of models $p(m)$ – the prior probability that model m is the true model, and $\sum_{m \in \mathbb{M}} p(m) = 1$. Within each model m the investigator has a prior $p(\theta_m|m)$ on the parameter $\theta_m \in \Theta_m$, and $\int_{\Theta_m} p(\theta_m|m)\mathrm{d}\theta_m = 1$.

The investigator seeks to collect new information in order to answer the questions Q. This involves collecting data, and this in turn requires some protocol or **experimental design.** For simplicity we use the term 'experiment' to cover all possible data collection exercises, whether they would be classed as observational or experimental, or even the acquisition of existing data already collected by other investigators. In geophysical problems this could include deployment of seismometers at chosen locations, samples of rock at a variety of depths, or acquisition of seismic records from an archive.

The **design space** $\mathbb{D}$ includes all possible experimental designs. These designs are independent of all of the models. Each element $d \in \mathbb{D}$ defines a **sample space** $\mathcal{Y}_d$ of possible observations $y_d \in \mathcal{Y}_d$ which are observable under that design. Note that y_d may be a (vector of) continuous and/or discrete random variables. For every design $d \in \mathbb{D}$ and every model $m \in \mathbb{M}$ there is a statistical likelihood for the observable data $f(y_d|d, \theta_m, m)$, which describes the probability (density) of observing y_d if model m holds with parameter values θ_m.

When considering plausible models for nature, the investigator may entertain models with considerable variation in complexity and structure. Some models may be very simple, with only a small number of parameters – others extremely complex. However no matter which model is actually true, the investigator's questions Q must be answerable from the parameter θ_m of model m. Thus for each model m there exists a **target function** $T(\theta_m|m, Q)$ mapping the values of the parameters θ_m of model m into a **target space** $\mathbb{T}$. This space $\mathbb{T}$ is common to all models $m \in \mathbb{M}$, and the functions summarise the state of nature in exactly and only the terms specified in the questions Q.

$T(\theta_m|m, Q)$ may for example be a restriction of the dimension of the parameter space, eliminat-

ing nuisance parameters and retaining only those parameters with a common, pertinent interpretation which exists in all models $m \in \mathbb{M}$. For example, we might be interested in the predicted earthquake recurrence times on a large fault at a tectonic plate boundary – and might consider a simple model m_1 which only includes the large scale structure of the boundary summarised in a small parameter set θ_1. We might also entertain a more complex model m_2 with a larger parameter set θ_2 which includes additional parameters quantifying the slip rates on an array of minor faults. In such a case $T_1(\theta_1|m_1, Q)$ and $T_2(\theta_2|m_2, Q)$ simply extract the recurrence time on the major fault from each model, ignoring all other parameters.

In another setting an investigator might wish to answer a question such as, does the seismic velocity structure indicate the presence of a specified set of subsurface properties of interest? Then $T(\theta_m|m, Q)$ might be an indicator function, signaling whether or not the model confirms the presence of each property.

The existence of a common target space $\mathbb{T}$ places constraints on the set of models $\mathbb{M}$: all models $m \in \mathbb{M}$ must be able to be mapped to the same target space through some $T(\theta_m|m, Q)$. We refer to a value $t \in \mathbb{T}$ of the target function as the **summarised** state of nature: t embodies only those aspects of the state of nature relevant to the questions Q.

The **answer space** $\mathbb{A}$ contains the set of possible answers to the investigator's question Q. For each answer $a \in \mathbb{A}$ there is, given the investigator's knowledge and preferences B, a **utility** function $U(a|t, d)$ associated with accepting answer a if the summarised true state of nature is in fact $t \in \mathbb{T}$ and design $d \in \mathbb{D}$ is executed. The utility depends on the experimental design d so that the costs of carrying out the design d are included, conditional on the true summarised state of nature t.

Our expressions for the utility $U(a|t, d)$ and target function $T(\theta_m|m, Q)$ are all conditioned on the state of knowledge B. We make this conditioning explicit in Section 4 when we discuss sequential interrogations which involve updating B, but until that section we suppress dependence on B for notational simplicity.

2.2 Identifying optimal answers

The investigator wishes to select the best answer a to the question Q: i.e. that which is optimal given the above construction. Optimality is determined by maximising the investigator's utility U. Generally speaking we have problems of two kinds, conditioned on the investigator's state of knowledge and utility:

1. **Decision Problem.** Given a particular design d, and an already observed dataset y_d, what is the best answer $a^*(y_d, d)$? Conditional on a particular design d, $a^*(y_d, d)$ maps data onto answers. If $a^*(y_d, d)$ takes discrete values then it induces a partition of the sample space $\mathcal{Y}_d$.

2. **Design Problem.** What is the best experimental design d^* which will lead to the choice of answer with the highest utility? The solution to this problem identifies d^* as the best design, and consequently $a^*(y_d, d^*)$ as the best decision rule.

In principle we can solve both problems using the framework of optimal Bayesian experimental design (Chaloner & Verdinelli 1995). This approach requires the specification of a highly structured utility, $U(a|\theta_m, m, y_d, d)$ for answer $a \in \mathbb{A}$, parameter $\theta_m \in \Theta_m$, embedded in model $m \in \mathbb{M}$, and data $y_d \in \mathcal{Y}_d$ collected under the experimental design $d \in \mathbb{D}$.

However the investigator may in general have no means of constructing a utility function of such structure and complexity. Moreover, when agreeing to a function of this dimensionality for the utility, an investigator cannot generally be expected to appreciate all of the consequences of the choice of a specific functional form (Curtis & Lomax 2001). There is also no easy way to guarantee that the utilities specified are coherent across the various models under consideration.

Instead the investigator may only be able to specify with confidence a utility with respect to the answers to the questions being posed. This is the motivation for our construction of the Target space $\mathbb{T}$ in Section 2.1 above. It is more reasonable to expect that an investigator will be able to specify most readily the utility $U(a|t, d)$, i.e. the utility of accepting answer a if the true summarised state of nature in the Target space is t, optionally incorporating the costs of the design d. Recall that in our construction the summarised state of nature always exists for every model under consideration (if it does not exist, the model is irrelevant to question Q). Requiring a utility only at the level of the Target space relieves the investigator of the need to specify utilities for every parameter value for every model m in the space of models.

As a simple example, say Q is the question, "What is the depth of the Moho beneath a particular geographical location?". The Moho is generally expected to mark the transition from crustal to mantle seismic velocities. So one might seek an answer by estimating the velocity structure with depth beneath that location (parameters θ_m) by inverting measured surface wave dispersion curves (data y_d acquired in an experiment with design d) using a surface wave modal approximation for the forward problem (model m). It is not a trival task to specify the form of a utility function $U(a|\theta_m, m, y_d, d)$ of the answer a to question Q directly from any potentially encountered multi-dimensional velocity structure θ_m, which may use a variety of different parameterisations in different models m (Sambridge et al. 2006). Instead, say a target function $t = T(\theta_m|m, Q)$ is defined such that it transforms any velocity structure into the corresponding inferred depth of the Moho. Then defining a utility $U(a|t, d)$ is relatively easy: for example setting $U(a|t, d) = -(a - t)^2$ means that the utility is maximised when our estimate (the answer a) is as close as possible to the true depth (t).

We can therefore construct the expected posterior utilities of answers given data by first integrating

over the model and parameter spaces:

$$U_p(a|y_d, d) = \sum_{m\in\mathbb{M}} \int_{\Theta_m} U(a|T(\theta_m|m), d)p(\theta_m, m|y_d, d)\,\mathrm{d}\theta_m \tag{1}$$

Here $p(\theta_m, m|y_d, d)$ is the Bayesian posterior distribution over the space of models and over the parameter spaces for each model.

Given a set of data y_d observed under experimental design d the optimal answer a^* that solves the Decision Problem is the answer that maximises the utility above:

$$a^*(y_d, d) = \underset{a\in\mathbb{A}}{\text{argmax}}\, U_p(a|y_d, d) \tag{2}$$

The maximised utility corresponding to $a^*(y_d, d)$ is then denoted $U^*(y_d, d) = U_p(a^*|y_d, d)$.

On the other hand, before any data are observed the expected utility that will result from a design $d \in \mathbb{D}$ is the value of $U^*(y_d, d)$ after it has been averaged over all possible datasets observable under the experimental design:

$$\begin{aligned} U^*(d) &= \int_{\mathcal{Y}_d} U^*(y_d, d)p(y_d|d)\,\mathrm{d}y_d \\ &= \int_{\mathcal{Y}_d} \max_{a\in\mathbb{A}} \sum_{m\in\mathbb{M}} \int_{\Theta_m} U(a|T(\theta_m|m), d)p(\theta_m, m|y_d, d)p(y_d|d)\,\mathrm{d}\theta_m\,\mathrm{d}y_d \end{aligned} \tag{3}$$

using equation (1). The optimal design a priori that solves the Design Problem is thus design $d^* \in \mathbb{D}$ that optimises $U^*(d)$:

$$d^* = \underset{d\in\mathbb{D}}{\text{argmax}}\, U^*(d) \tag{4}$$

with maximised utility $U^{**} = U^*(d^*)$.

2.3 Estimating the Summarised State of Nature

We now make this procedure more concrete by considering the case where the goal is simply to estimate the summarised state of nature T. The answer, and the answer to the investigator's question, is then $a = \widehat{T}$ where $\widehat{T}$ is an estimate of T. We set the utility of estimate $\widehat{T}$ when the true value of T is t, to be the negative of the squared error function:

$$U(a|t, d) = U(a|t) = -(t - a)^T W(t - a) \tag{5}$$

Here t and a are vectors, and the target and answer spaces $\mathbb{T}$ and $\mathbb{A}$ are identical. W is a known, symmetric, positive definite weight matrix, and the first equality in equation (5) acknowledges that in this utility we have neglected any costs or benefits of acquiring data d other than those pertaining to estimate t. The answer $a^*(y_d, d)$ that maximises the utility $U(a|t)$ is the one that maximises the

posterior expected utility

$$\begin{aligned} U_p(a|y_d,d) &= E[U(a|T)|y_d,d] \\ &= -\text{tr}(W\text{Var}[T|y_d,d]) - (E[T|y_d,d]-a)^T W(E[T|y_d,d]-a) \end{aligned} \tag{6}$$

where tr$(\cdot)$ is the trace of a square matrix and Var$[\cdot]$ is the variance operator. The second equality follows from the definition of the variance-covariance of a vector, and is the vector generalisation of the scalar result: $E[(t-a)^2] = \text{Var}[t] + (E(t)-a)^2$, where we have dropped conditional dependencies and weights for simplicity. Since the first term on the right of equation (6) is invariant with respect to a, the optimal answer is that which minimises the magnitude of the second term. This is achieved by setting the estimate of T equal to the posterior mean of T averaged over all models m and parameters θ_m since for this choice of T the second term on the right of equation (6) is zero:

$$\begin{aligned} a^*(y_d,d) &= E[T|y_d,d] \\ &= \sum_{m\in\mathbb{M}} \int_{\Theta_m} T(\theta_m|m)p(\theta_m,m|y_d,d)\,\mathrm{d}\theta_m \end{aligned} \tag{7}$$

The maximised utility corresponding to $a^*(y_d,d)$ is then

$$U^*(y_d,d) = -\text{tr}(W\text{Var}[T|y_d,d]) \tag{8}$$

A priori the optimal experimental design is the design $d^* \in \mathbb{D}$ that optimises $U^*(y_d,d)$ after it has been averaged over all possible datasets observable under the design d:

$$U^*(d) = -\int_{\mathcal{Y}_d} \text{tr}(W\text{Var}[T|y_d,d])p(y_d|d)\,\mathrm{d}y_d \tag{9}$$

and so the optimal design is that which minimises the expected posterior variance. This design d^* is known as the A-optimal Bayesian design (see also §2.2 of Chaloner & Verdinelli 1995), however its expression here differs from standard treatments because of our use of the common target space.

An important special case arises when the investigator's interest is only in model selection. The question Q is, "What is the best representation of nature out of the set of M models in the space of models $\mathbb{M}$?" The answer a is identity of this best model. This amounts to estimating the value of an $M \times 1$ indicator vector T, which is zero except for its m^*th entry which is 1, indicating that the true model is m^*. We represent such an indicator vector as δ_{m^*}.

For each model m, the function $T(\theta_m|m)$ is conditioned on model m being true. The indicator vector for $T(\theta_m|m)$ is therefore δ_m. This is independent of parameter values θ_m, thus all of the parameters θ_m are in fact nuisance parameters.

From equation (7) the optimal answer under the squared error loss function in equation (5) is

$$a^*(y_d,d) = E[T|y_d,d] = \sum_{m\in\mathbb{M}} \delta_m p(m|y_d,d) \tag{10}$$

which implies that the optimal estimate $a^*(y_d, d)$ is simply the $M \times 1$ vector of posterior probabilities of each of the models m in the space of models $\mathbb{M}$ being true.

From equation (8) the maximised utility corresponding to this estimate is

$$U^*(y_d, d) = -\sum_{m\in\mathbb{M}} W_{mm} p(m|y_d, d) + \sum_{m\in\mathbb{M}} \sum_{m'\in\mathbb{M}} W_{mm'} p(m|y_d, d) p(m'|y_d, d) \tag{11}$$

If the weight matrix W is the identity, then this expression simplifies further to

$$U^*(y_d, d) = -\sum_{m\in\mathbb{M}} p(m|y_d, d)(1 - p(m|y_d, d)) \,. \tag{12}$$

This sum achieves its maximum possible value if the posterior probabilities $p(m|y_d, d)$ only take values 0 or 1 in which case $U^*(y_d, d) = 0$ (in words, our utility is maximised because we make zero loss). Moreover since those probabilities must add up to 1, this largest possible value is achieved when all of the posterior probability concentrates on a single model and all other models have zero posterior probability.

Integrating equation (12) over all possible datasets y_d from the design d weighted by their probabilities $p(y_d|d)$ we obtain

$$\begin{aligned} U^*(d) &= -1 + \int_{\mathcal{Y}_d} \sum_{m\in\mathbb{M}} p(m|y_d, d) p(y_d|m, d) p(m) \,\mathrm{d}y_d \\ &= -1 + E[p(m|y_d, d)] \end{aligned} \tag{13}$$

so that the a priori optimal design d^* for model selection is thus the one that maximises the expected posterior probability $E[p(m|y_d, d)]$ where the expectation in this expression is over all models m and all data y_d observable under the design d.

3 A WORKED EXAMPLE

We now consider a simple example of a generic inverse problem with which we can demonstrate the methodology above.

3.1 Problem Motivation

Over recent years, the method that has become most widely used for finding the solution to an inverse problem is to deploy Bayes rule:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int_{\theta\in\Theta} p(y|\theta)p(\theta)\mathrm{d}y} \tag{14}$$

where $p(\theta)$ is the prior distribution on parameters θ, and $p(y)$ is the marginal distribution of the observations y and is called the evidence. The solution on the left is expressed as a conditional distribution

of θ given the observed values of y which usually involves solving an inverse problem. The attraction of equation (14) is that it converts that inverse problem on the left into a forward problem (finding the probability of observing y given the values of parameters θ) as shown on the right. However, as can be seen in the second equality, a remaining difficulty is that to calculate the evidence requires an integration to be performed over parameter space $\mathcal{Y}_d$. This is a serious issue as in many Geophysical problems either performing or avoiding direct computation of this integral requires the use of Monte Carlo methods, incurring substantial computational expense (e.g., Mosegaard & Tarantola 1995; Tarantola 2005; Bodin & Sambridge 2009).

Motivated by this problem, we now show how interrogation theory can be applied to obtain optimal results for the integration of probabilistic variables. We demonstrate the method using one dimensional variables, but the multi-parameter case works similarly.

3.2 Problem Specification

Assume that there is some observable scalar property $\mu(x)$ of the universe that varies with position x along the real line. An investigator is not interested in the details of the form of $\mu(x)$, but only wants to determine the scalar value of the integral

$$T = \int_{\mathbb{X}} \mu(x)\ w(x)\ \mathrm{d}x \tag{15}$$

for some specified weight function $w(x)$. Here the domain $\mathbb{X}$ of x is a portion or all of the real line $\mathbb{R}$. For example, Figure 3 shows T as the integral of $\mu(x)$ with $w(x)$ simply being an indicator function $w(x) = I(x \in [a, b])$ that restricts the range of x to an interval of interest $[a, b]$. This clearly relates to the problem of finding the evidence $p(y)$ in equation (14) with $\mu(x) = p(y|x)$ and $x = \theta$. Alternatively, a practical example of such an integral occurs where $\mu(x)$ is energy loss due to attenuation during transmission of waves over the spatial interval $[a, b]$; we may neither know nor be interested in the exact form of $\mu(x)$, but nevertheless need to estimate total energy loss T. Each of these scenarios might constitute types of questions Q that have answers given by target T of the form in equation (15).

Other forms of functions that we could have chosen instead of equation (15) include calculating an average of the gradient $d\mu(x)/dx$ over some interval, the maximum value of $\mu(x)$ in an interval, or the value of $\mu(x)$ at some inaccessible (or future) value of x. We consider equation (15) principally for its simplicity, which means we are able to derive some explicit analytic results below. More complex (possibly multidimensional) target functions T may require numerical and/or simulation methods in what follows, but the interrogation theory remains similar.

We assume that the value of $\mu(x)$ is observed at a selected set of points x, and from these values an estimate of T will ultimately be obtained. It is common that physical models of this sort need to

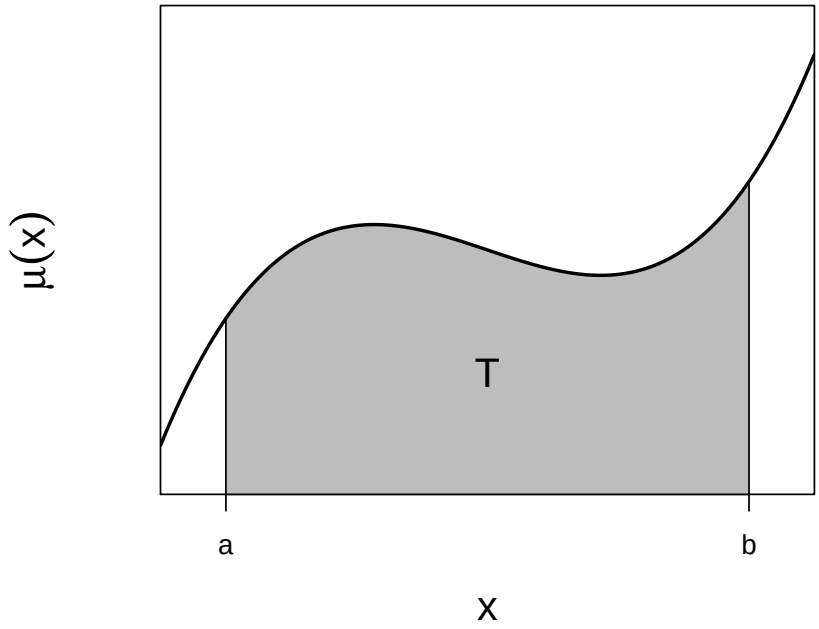

Figure 3. Example integral $T = \int_a^b \mu(x)\mathrm{d}x$ giving the shaded area under the graph. The investigator wants to estimate the value of T, with the precise form of $\mu(x)$ being only of secondary interest.

incorporate some natural variability around an overall trend function, for example. It is also always the case that there is some measure of observational uncertainty associated with every measurement of $\mu(x)$. These two components of uncertainty can often be separated through repeated measurements of $\mu(x)$ at the same locations. For the purposes of the current example we leave them combined as additive, independent and identically distributed zero mean Normal errors. Furthermore, for simplicity in this first example, we assume that the variance of these errors σ^2 is known. This means that an observation y of $\mu(x)$ at x takes the form

$$y(x, \varepsilon) = \mu(x) + \varepsilon \tag{16}$$

where ε is the $N(0, \sigma^2)$ additive error.

Equation (15) may be defined for any model for $\mu(x)$ deemed to be plausible. Again to keep things simple we assume that the investigator entertains only constructions for $\mu(x)$ that are linear in parameters θ_m, and that there are M possible models of interest, with

$$\mu(x|\theta_m, m) = g_m(x)^T\theta_m \qquad \text{for } m = 1, \ldots, M, \text{ and } x \in \mathbb{X} \tag{17}$$

Here the vector $g_m(x)$ is a vector of functions of x, effectively basis functions, appropriate to the model under consideration. For example model m_1 might be a quadratic function in which case $\mu(x|\theta_1, m_1) = \theta_{11} + \theta_{12}x + \theta_{13}x^2$ and $g_1(x) = (1, x, x^2)^T$. Model m_2 might only have two

parameters, but depend on x through the log function, e.g., $\mu(x|\theta_2, m_2) = \theta_{21} + \theta_{22} \log x$ with $g_2 = (1, \log x)^T$. We leave the actual content of the functions $g_m(x)$ unspecified for the moment.

It follows that under the assumption that model m is true, an observation y of $\mu(x|\theta_m, m)$ at location x has distribution

$$y(x|\theta_m, m) \sim N(\mu(x|\theta_m, m), \sigma_m^2) \qquad \text{for } m = 1, \ldots, M, \text{ and } x \in \mathbb{X} \tag{18}$$

with σ_m^2 assumed known. The models in equation (17) form the entire space of models $\mathbb{M} = \{m_1, \ldots, m_M\}$, and the investigator gives them prior probabilities $(p(m_1), \ldots, p(m_M))$.

The parameter spaces of each model are $\Theta_m = \mathbb{R}^{p_m}$, where p_m is the dimension of θ_m. The investigator adopts Normal priors for these

$$\theta_m|m \sim N(\theta_{0m}, \sigma_m^2 R_{0m}) \tag{19}$$

where θ_{0m} is the p_m dimensional prior mean for θ_m, R_{0m} is a specified $p_m \times p_m$ positive definite matrix, and σ_m^2 is again the known error variance.

The scientific questions of interest Q are entirely summarised by the question "What is the value of the integral in equation (15)?" So, for the target functions T we have the expression

$$T(\theta_m|m) = \int_{\mathbb{X}} \mu(x|\theta_m, m)\ w(x)\ \mathrm{d}x = \bar{g}_m^T \theta_m \tag{20}$$

where we have defined

$$\bar{g}_m = \int_{\mathbb{X}} g_m(x)\ w(x)\ \mathrm{d}x \tag{21}$$

For all models $m \in \mathbb{M}$ the values of T lie in the space of real numbers $\mathbb{T} = \mathbb{R}$. The answer to the question is the estimated value of the integral, so that the space of answers $\mathbb{A}$ is the same as $\mathbb{T} = \mathbb{R}$. The investigator chooses the squared error utility function $U(a|t, d) = -(t - a)^2$, the form taken by equation (5) when a scalar is to be estimated.

The investigator specifies a design space $\mathbb{D}$: an element d of this space is the choice of a set of n_d sampling locations $\{x_{di} : i = 1, \ldots, n_d\}$ along the x-axis.

Given a design d and a model m, the functions of x contained in each $g_m(x)$ in (17) give the rows of the $n_d \times p_m$ model matrix X_{md}, with $X_{md;ij} = g_{mj}(x_{di})$ for $i = 1, \ldots n_d$ and $j = 1, \ldots, p_m$. This fully specifies the statistical model $y_d|d, \theta_m, m$ that gives rise to the observed data y_d:

$$y_d|d, \theta_m, m \sim N(X_{md}\theta_m, \sigma_m^2 I) \tag{22}$$

Given the above, and assuming we have chosen the functions $g_m(x)$ in (17) that define each model and hence design matrices X_{md}, the Decision Problem is now fully specified, and the Design Problem is to select an optimal set of n_d sampling locations at which to observe the values of y.

3.3 Solution

We now seek the form of solutions to the Decision and Design Problems outlined in Section 2.2. Before proceeding to find these solutions we note by applying standard linear inverse theory that the priors and likelihood from (19) and (22) imply that within model m the posterior for θ_m given data y_d from design d is

$$\theta_m|y_d, d, m \sim N(M_{md}^{-1}u_m(y_d), \sigma_m^2 M_{md}^{-1}) \tag{23}$$

where we have defined

$$M_{md} = R_{0m}^{-1} + X_{md}^T X_{md} \quad \text{and} \quad u_{md}(y_d) = R_{0m}^{-1}\theta_{0m} + X_{md}^T y_d \tag{24}$$

The posterior mean of θ_m in model m, conditional on data y_d and design d can thus be written

$$\widehat{\theta}_{md}(y_d) \equiv E[\theta_m|y_d, d, m] = M_{md}^{-1}u_{md}(y_d) = M_{md}^{-1}(R_{0m}^{-1}\theta_{0m} + X_{md}^T y_d) \tag{25}$$

The marginal distribution of $y_d|d, m$, integrating over θ_m, is

$$y_d|d, m \sim N(X_{md}\theta_{0m}, \sigma_m^2 S_{md}^{-1}) \tag{26}$$

where we have defined

$$S_{md} = I_{n_d} - X_{md}M_{md}^{-1}X_{md}^T \tag{27}$$

and I_{n_d} is the $n_d \times n_d$ identity matrix. Before seeing any data the expected distribution of the posterior mean $\widehat{\theta}_{md}(y_d)$ in (25) is

$$\widehat{\theta}_{md}|d, m \sim N(\theta_{0m}, \sigma_m^2(R_{0m} - M_{md}^{-1})) \tag{28}$$

Since our chosen utility is the squared error function, we can use the results of Section 2.3. In particular, given a particular design d, and an observed dataset y_d under that design, the optimal answer $a^*(y_d, d)$ is given by (7):

$$a^*(y_d, d) = \sum_{m\in\mathbb{M}} \bar{g}_m^T M_{md}^{-1}(R_{m0}^{-1}\theta_{m0} + X_{md}^T y_d)\ p(m|y_d, d) \tag{29}$$

The posterior $p(m|y_d, d)$ for the models m in (29) is given by Bayes rule using the marginal distribution $p(y_d|d, m)$ from (26) and the prior probabilities on the space of models $p(m)$. This solves the Decision Problem conditional on observations y_d from design d.

The maximised utility corresponding to $a^*(y_d, d)$ is given by (8):

$$\begin{aligned} U^*(y_d, d) &= -E[T^2|y_d, d] + (a^*(y_d, d))^2 \\ &= -\sum_{m\in\mathbb{M}} \bar{g}_m^T \left[\sigma_m^2 M_{md}^{-1} + M_{md}^{-1}u_{md}(y_d)u_{md}^T(y_d)M_{md}^{-1}\right]\ \bar{g}_m\ p(m|y_d, d) \\ &\qquad + (a^*(y_d, d))^2 \end{aligned} \tag{30}$$

To solve the Design Problem we find the a priori expected value of $U^*(y_d, d)$ by averaging over all possible observations y_d:

$$U^*(d) = -\sum_{m\in\mathbb{M}} \bar{g}_m^T \left[\sigma_m^2 R_{0m} + \theta_{0m}\theta_{0m}^T\right] \bar{g}_m \; p(m) + E\left[\left(a^*(y_d,d)\right)^2 \middle| d\right] \tag{31}$$

The optimal design maximises $U^*(d)$. Note that the first term of (31) is independent of the design, so that the optimal design is the one that maximises

$$\begin{aligned} \check{U}^*(d) &= E\left[\left(a^*(y_d,d)\right)^2 \middle| d\right] \\ &= \int_{\mathcal{Y}_d} \left[\sum_{m\in\mathbb{M}} \bar{g}_m^T \widehat{\theta}_{md}(y_d)\; p(m|y_d,d)\right]^2 p(y_d|d)\; \mathrm{d}y_d \end{aligned} \tag{32}$$

Thus we see that an optimal design is obtained by maximising the expectation over the data space of the squared target value's expectation over the space of models. An ideal sampling strategy is therefore one that chooses samples (perhaps sequentially - see below) that maximise this expected utility.

3.4 Demonstration

As a demonstration of estimating the integral T, consider a laboratory experiment to estimate the creep deformation characteristics of a rock sample by applying a constant stress over time x and measuring the resulting strain $\mu(x)$ across the sample for some function μ. It is well known that the dominant initial (primary) mode of deformation is approximately logarithmic in time ($\mu(x)$ takes a form similar to $\log(1+x)$); as strain builds, the response transitions to a secondary mode which is dominated by strain that is linear in time ($\mu(x)$ is approximately proportional to x) – see Boukharov et al. (1995). For any new material, there can be significant uncertainty about which measured data correspond to primary and which to secondary creep mechanisms, hence any inference we wish to make based on the creep properties must take this uncertainty into account.

Thus motivated we consider the case where there are two single parameter models with different basis functions:

$$\begin{aligned} m_1: &\quad \mu_1(x) = \theta_1 x && \text{for } \theta_1 \in \mathbb{R} \\ m_2: &\quad \mu_2(x) = \theta_2 \log(1+x) && \text{for } \theta_2 \in \mathbb{R} \end{aligned} \qquad \text{and } x \in [0, b] \tag{33}$$

The two basis functions which are being compared are the linear function $g_1(x) = x$ and the curved function $g_2(x) = \log(1+x)$. They both pass through the origin, and for sufficiently large b (i.e. beyond the linear regime close to $x = 0$), they are distinguishable by the curvature of the $\log$ function. We expect that detection of that curvature would require an adequate coverage of x values across the interval $[0, b]$.

We assume the error variance $\sigma_m^2 = \sigma^2$ is known and is the same for both models, we assume equal

prior probabilities for the models $p(m_1) = p(m_2) = 1/2$, and the same diffuse prior $\theta_m \sim N(0, \sigma_0^2)$ for θ_m in both models.

Initially, say our goal is to estimate the integral

$$T = \int_0^b \mu(x)\, \mathrm{d}x \tag{34}$$

When evaluated for each of the models, this takes the values

$$T(\theta_m|m) = \bar{g}_m \theta_m \quad \text{with} \quad \bar{g}_m = \begin{cases} \frac{1}{2}b^2 & \text{if } m = 1 \\ (1+b)\log(1+b) - b & \text{if } m = 2 \end{cases} \tag{35}$$

The design space consists of choices of n observation locations $\mathbf{x}_d$ at which observations of y will be made. Given a choice of locations the design matrices X_{md} are simply $n \times 1$ column vectors $\mathbf{x}_{md}$:

$$X_{1d} = \mathbf{x}_{1d} = \begin{bmatrix} x_{d1} \\ x_{d2} \\ \vdots \\ x_{dn} \end{bmatrix} \quad \text{and} \quad X_{2d} = \mathbf{x}_{2d} = \begin{bmatrix} \log(1+x_{d1}) \\ \log(1+x_{d2}) \\ \vdots \\ \log(1+x_{dn}) \end{bmatrix} \tag{36}$$

An observation y_d under design d is a vector of n real values $\mathbf{y}_d$. Under model m with parameters θ_m, and design d these are drawn from the distribution

$$\mathbf{y}_d|d, \theta_m, m \sim N(\mathbf{x}_{md}\theta_m, \sigma^2 I_n) \tag{37}$$

Integrating out θ_m over its Normal prior distribution, the distribution of $\mathbf{y}_d$ is

$$\mathbf{y}_d|d, m \sim N(0, \sigma^2 S_{md}^{-1}) \tag{38}$$

where

$$S_{md} = I_n - \frac{\sigma_0^2 \mathbf{x}_{md}\mathbf{x}_{md}^T}{\sigma^2 + \sigma_0^2 \mathbf{x}_{md}^T \mathbf{x}_{md}} \tag{39}$$

$$S_{md}^{-1} = I_n + \frac{\sigma_0^2}{\sigma^2}\mathbf{x}_{md}\mathbf{x}_{md}^T \tag{40}$$

with $|S_{md}| = \sigma^2/(\sigma^2 + \sigma_0^2 \mathbf{x}_{md}^T \mathbf{x}_{md})$. The marginal distribution of $\mathbf{y}_d$ averaging over the prior on models is a mixture (sum) of Normal distributions:

$$p(\mathbf{y}_d|d) = (2\pi\sigma^2)^{-n/2} \sum_{m \in \mathbb{M}} p(m) |S_{md}|^{1/2} \exp(-\tfrac{1}{2\sigma^2}\mathbf{y}_d^T S_{md} \mathbf{y}_d) \tag{41}$$

The posterior distribution of the parameter θ_m within model m is then found to be

$$\theta_m|y_d, d, m \sim N\left(\frac{\sigma_0^2 \mathbf{x}_{md}^T \mathbf{y}_d}{\sigma^2 + \sigma_0^2 \mathbf{x}_{md}^T \mathbf{x}_{md}}, \frac{\sigma^2 \sigma_0^2}{\sigma^2 + \sigma_0^2 \mathbf{x}_{md}^T \mathbf{x}_{md}}\right) \tag{42}$$

The posterior distribution of models m given data $\mathbf{y}_d$ from design d is

$$p(m|\mathbf{y}_d, d) = \frac{p(m)|S_{md}|^{1/2} \exp(-\frac{1}{2\sigma^2}\mathbf{y}_d^T S_{md}\mathbf{y}_d)}{\sum_{m' \in \mathbb{M}} p(m')|S_{m'd}|^{1/2} \exp(-\frac{1}{2\sigma^2}\mathbf{y}_d^T S_{m'd}\mathbf{y}_d)} \tag{43}$$

Thus the optimal answer $a^*(\mathbf{y}_d, d)$ given data $\mathbf{y}_d$ from design d is given by (29):

$$a^*(\mathbf{y}_d, d) = \sum_{m \in \mathbb{M}} \frac{\bar{g}_m \sigma_0^2 \mathbf{x}_{md}^T \mathbf{y}_d}{\sigma^2 + \sigma_0^2 \mathbf{x}_{md}^T \mathbf{x}_{md}} p(m|\mathbf{y}_d, d) \tag{44}$$

This answer to the question is our best estimate of the integral T, and has variance equal to the negative optimised utility by equation (30):

$$\begin{aligned} \text{Var}[T|\mathbf{y}_d, d] &= -U^*(\mathbf{y}_d, d) \\ &= \sum_{m \in \mathbb{M}} \frac{\bar{g}_m^2 \sigma_0^2}{\sigma^2 + \sigma_0^2 \mathbf{x}_{md}^T \mathbf{x}_{md}} \left[\sigma^2 + \frac{\sigma_0^2 (\mathbf{x}_{md}^T \mathbf{y}_d)^2}{\sigma^2 + \sigma_0^2 \mathbf{x}_{md}^T \mathbf{x}_{md}}\right] p(m|\mathbf{y}_d, d) \\ &\quad - \left(a^*(\mathbf{y}_d, d)\right)^2 \end{aligned} \tag{45}$$

To identify the optimal design a priori, by (32) we need to find the design d (the set of sampling locations) that maximises

$$\check{U}^*(d) = \int_{\mathcal{Y}_d} \left[a^*(\mathbf{y}_d, d)\right]^2 p(\mathbf{y}_d|d) \, \mathrm{d}\mathbf{y}_d \tag{46}$$

where $p(\mathbf{y}_d|d)$ is the mixture of Normal distributions given in equation (41) and $a^*(\mathbf{y}_d, d)$ is given in equation (44). Note that the integral over $\mathbf{y}_d$ is an $n-$dimensional integral, so that in designs with large samples (i.e., with many values of x at which $\mu(x)$ is observed) the optimisation to find the best design d^* may be computationally costly.

As an explicit example, we have implemented the model above with the fixed parameter settings $b = 10$, $\sigma_0 = 3$ and $\sigma = 0.2$. We considered five designs, each with three sampling locations, which are listed in Table 2. We evaluated $\check{U}^*(d)$ in equation (46) for each design, and the results are also shown in Table 2. They identify design d_4 as the optimal design, i.e. the design with maximal utility a priori, with the data points spread widely across the interval. Similarly spread out designs d_2 and d_5 are almost as good. The worst design d_1 puts all of the sampling points close to zero, meaning that there is no hope of discriminating between the linear and log funcitons.

We simulated a dataset for each of the five designs using Model 1 and a true parameter value of $\theta_m = 0.4$, corresponding to a true value of the integral of $T = c_1\theta_1 = 20$. We then approached these datasets as an analyst ignorant of the true model would. We computed the posterior probability $p(m|\mathbf{y}_d, d)$ of each model $m \in \{1, 2\}$ given the data and design using equation (43), and the best estimate of the integral $\widehat{T} = a^*(\mathbf{y}_d, d)$ using equation (44) with its standard error $\text{SE}(\widehat{T}) = \sqrt{-U^*(\mathbf{y}_d, d)}$ (square root of the variance from equation (45)). These values are listed in Table 3. The design d_4, identified a priori as the best, is also the best a posteriori (smallest standard error), again with d_2 and

Table 2. Experimental designs and their corresponding a priori utilities for integral estimation. The values $\check{U}^*(d)$ are calculated using (46) with $a^*(\mathbf{y}_d, d)$ from (44) and $p(\mathbf{y}_d|d)$ from (43).

Design, d	x_{d1}	x_{d2}	x_{d3}	Description	$\check{U}^*(d)$	
d_1	1	1.5	2	Clustered close to 0	12356	(worst)
d_2	1	5.0	10	Spaced widely	12456	
d_3	9	9.5	10	Clustered far from 0	12411	
d_4	1	9.5	10	One near 0, two clustered far	12505	(best)
d_5	1	1.5	10	Two near 0, one far	12456	

d_5 performing almost as well. All three designs strongly discriminate between the two models by assigning effectively zero probability to model 2, and yield very precise estimates $\widehat{T} = a^*(\mathbf{y}_d, d)$ of T. Notice that (by chance) design d_5 leads to a posterior estimate of the target integral that is closest to the true value of 20. However, since the true value is unknown, the investigator is unaware of this fortuitous outcome, and therefore can not take advantage. The worst design a priori d_1 has poor discrimination and the least precise estimate. The basis functions and the best fitting models for the best design d_4 and the worst design d_1 are displayed in Figure 4.

It might be that instead of estimating the integral value in equation (34) our interest was defined by the question: "which is the best model from the two candidates in (33)?" Then, from Section 2.3, the optimal answer is the estimate of the 2×1 indicator vector T which takes the value $(1, 0)^T$ if Model 1 holds, and $(0, 1)^T$ if Model 2 holds.

Given observations y_d from design d then the optimal estimate of T that solves this Decision

Table 3. Simulated Data. For each of five designs we list a set of simulated observations $\mathbf{y}_d$, posterior probabilities of each model $p(m|\mathbf{y}_d, d)$, and estimates of the integral T from (35) (the true value is 20), with associated standard errors from (45). The true model is $m = 1$.

	Observations			$p(m\|\mathbf{y}_d, d)$				
Design, d	y_{d1}	y_{d2}	y_{d3}	$m = 1$	$m = 2$	$\widehat{T}$	SE($\widehat{T}$)	
d_1	0.393	1.094	0.660	0.290	0.710	15.89	5.32	(worst)
d_2	0.327	1.945	4.164	1.000	0.000	20.51	0.89	
d_3	3.476	3.690	3.695	0.196	0.804	24.01	2.57	
d_4	0.301	3.748	3.939	0.999	0.001	19.69	0.75	(best)
d_5	0.264	0.520	4.035	1.000	0.000	20.05	0.98	

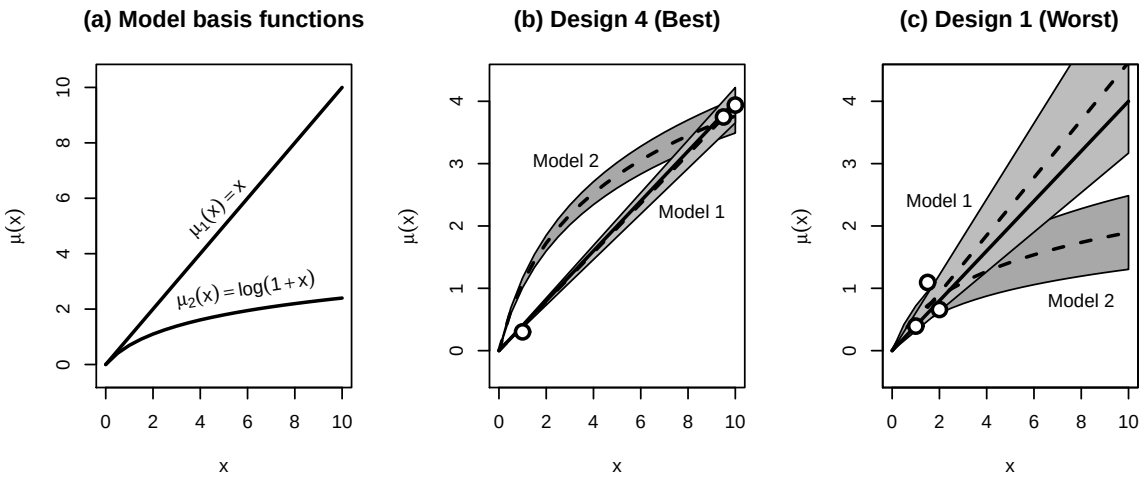

Figure 4. (a) Basis functions for the two models; (b) and (c) show the best fits for Models 1 and 2 under the best (d_4) and worst (d_1) of the five designs, given the data simulated under Model 1 with $\theta = 0.4$ (Table 3). The true model is shown as a solid line, the simulated data are shown as open circles. The shaded areas around the fitted curves are the 95% posterior credible intervals for the fitted models.

Problem is given by equation (10). Thus for our simulated data the estimates of the vector T are the pairs of $p(m|y_d, d)$ values from Table 3, reproduced in Table 4. The data from designs d_2, d_4 and d_5 strongly favour Model 1 (the true model), with posterior probabilities of Model 1 close to 1. The data from designs d_1 and d_3 favour Model 2, but less strongly than the other three designs favour Model 1 – the posterior probabilities for Model 2 being only 0.71 and 0.80 under d_1 and d_3 respectively.

The optimal design d^* for model selection is the one that maximises the model selection utility $\check{U}^*_{MS}(d) = E[p(m|y_d, d)]$ in equation (13). Values of $\check{U}^*_{MS}(d)$ for the five candidate designs are given in Table 4. On the basis of this prior utility the ordering of the five designs is similar to that for the estimation of the integral value (Table 2) in that the top three optimal designs a priori are d_2, d_4 and d_5 with the candidate points spaced widely, and the worst is again d_1 in which they cluster close to zero. There is a slight preference for design d_5 in this case, showing that the optimal experiment to perform varies with the question posed.

4 SEQUENTIAL ANALYSES

We now consider the situation where a sequence of data collection exercises are to be carried out. At each stage the optimal experimental design is selected, and data are collected. The new data result in an updated state of knowledge, then the process repeats. We refer to this as a **sequential interrogation**.

Table 4. Experimental designs, a priori utilities for model selection, and estimates of the summarised state of nature (the latter repeated from the simulated data in Table 3). The a priori utilities $\check{U}^*_{MS}(d)$ are given by (13) and estimates $\widehat{T}_m = p(m|\mathbf{y}_d, d)$ come from (43).

Design, d	Description	$\check{U}^*_{MS}(d)$	$\widehat{T} = (\widehat{T}_1, \widehat{T}_2)$
d_1	Clustered close to 0	0.75	(0.290, 0.710)
d_2	Spaced widely	0.96	(1.000, 0.000)
d_3	Clustered far from 0	0.83	(0.196, 0.804)
d_4	One near 0, two clustered far	0.95	(0.999, 0.001)
d_5	Two near 0, one far	0.97	(1.000, 0.000)

Such sequential exercises arise from the recognition that the first questions posed by the investigator may ultimately be found to be irrelevant. In the example of interrogating a human criminal suspect the interrogator is motivated to Solve The Crime; they may commence the interrogation already convinced of the broad features of how a crime was committed, and be seeking information about the finer detail. The utility for the initial questions may relate to accurate estimates of the sequencing and timings of certain key events. It may however be found during questioning that the entire logical reasoning of the interrogator was based on a flawed assumption. That being the case, questioning may proceed in a new and entirely unanticipated direction. A geophysical example might be that a surprising revelation about the nature of subsurface structures may induce a change in the focus of enquiry: questions about the exact location and size of a particular known fault may shift to the mapping of hitherto undetected faults that are revealed by an imaging experiment conducted as part of the interrogation.

To formalise the sequential process we first need a definition of the state of knowledge. We then define exactly what we mean by a sequential interrogation, and show how the state of knowledge is updated after each step in the sequence. We illustrate this procedure with a simple example.

4.1 The state of knowledge

At the start of the k^{th} step the state of knowledge of the investigator is B_k. This state of knowledge includes, or can be translated into, the following components:

1. A set of questions Q_k
2. The space of models $\mathbb{M}_k$, containing all of the models that are considered by the investigator at step k;
3. For each model $m \in \mathbb{M}_k$:

- a prior probability $p_k(m) = p(m|B_k)$ of the model;
- a parameter space Θ_m;
- a prior distribution $p_k(\theta_m|m) = p(\theta_m|m, B_k)$ of the parameter θ_m in Θ_m.

4. Target functions $T_k(\theta_m|m) = T(\theta_m|m, B_k)$ which encode the questions Q_k for each model m. For all models $m \in \mathbb{M}_k$ the function $T_k(\theta_m|m)$ takes values in the single target space $\mathbb{T}_k$.

5. An answer space $\mathbb{A}_k$.

6. A design space $\mathbb{D}_k$ of experimental designs amongst which a design is to be chosen at step k.

7. A utility $U_k(a|t, d) = U(a|t, d, B_k)$ for answer $a \in \mathbb{A}_k$, defined as the utility of accepting answer a if the true value of the function $T(\cdot|\cdot)$ is t and the design chosen is $d \in \mathbb{D}_k$.

8. For each design $d \in \mathbb{D}_k$ and model $m \in \mathbb{M}_k$ there is a likelihood $f(y_d|d, \theta_m, m)$ for data $y_d \in \mathcal{Y}_d$. Since the likelihood is conditioned only on the chosen design d and model m, it is otherwise independent of the state of knowledge B_k.

9. The stock of any other knowledge Ω_k that the investigator has, which is not captured by 1–8 above.

While knowledge Ω_k appears to be superfluous at iteration k, it may prove to contain significant information at later iterations. Hence, Ω_k represents our stock of intellectual capital that we hold in reserve.

In summary we can represent the state of knowledge in symbolic terms as

$$\begin{aligned} B_k \;=\; & (\Omega_k, Q_k, \mathbb{M}_k, \mathbb{T}_k, \mathbb{A}_k, \mathbb{D}_k, \\ & \{p_k(m), \{p_k(\theta_m|m), T_k(\theta_m|m)|\theta_m \in \Theta_m\}|m \in \mathbb{M}_k\}, \\ & \{U_k(a|t, d)|a \in \mathbb{A}_k, t \in \mathbb{T}_k\}) \end{aligned} \tag{47}$$

4.2 Interrogation procedure

The procedure is as follows. At each step k the methods described in earlier sections can be applied to solve the Decision and Design Problems:

1. Given the state of knowledge B_k at step k the optimal design d_k^* can be identified (as the solution of the appropriate Design problem):

$$\begin{aligned} d_k^* \;&=\; \operatorname*{argmax}_{d \in \mathbb{D}_k} U_k^*(d) \\ &=\; \operatorname*{argmax}_{d \in \mathbb{D}_k} \int_{\mathcal{Y}_d} \max_{a \in \mathbb{A}_k} \sum_{m \in \mathbb{M}_k} \int_{\Theta_m} U_k(a|T(\theta_m|m), d_k) \\ &\qquad \times p_k(\theta_m, m|y_d, d) p_k(y_d|d)\, \mathrm{d}\theta_m\, \mathrm{d}y_d \end{aligned} \tag{48}$$

2. Design d_k^* is then implemented, and data $y_{d_k^*}$ are collected.

3. The optimal answer is

$$a_k^* = \underset{a \in \mathbb{A}_k}{\operatorname{argmax}} \sum_{m \in \mathbb{M}_k} \int_{\Theta_m} U_k(a|T_k(\theta_m|m), d_k^*) p_k(\theta_m, m|y_{d_k^*}, d_k^*) \, \mathrm{d}\theta_m \tag{49}$$

and this achieves an estimated utility of

$$U_k(a_k^*|y_{d_k^*}, d_k^*) = \sum_{m \in \mathbb{M}_k} \int_{\Theta_m} U_k(a_k^*|T_k(\theta_m|m, d_k^*)) p_k(\theta_m, m|y_{d_k^*}, d_k^*) \, \mathrm{d}\theta_m \tag{50}$$

4. The state of knowledge is updated combining the previous state B_k and new data $y_{d_k^*}|d_k^*$ to form a new state of knowledge B_{k+1}. Details of the updating process are given below.

5. The procedure repeats for step $k+1$ unless a condition for terminating the sequence has been met. At termination the final answer a_k^* has been accepted and the final state of knowledge is B_{k+1}.

Updating the state of knowledge to B_{k+1} involves updating some or all components in equation 47, potentially including the questions Q_{k+1}. The application of the termination condition and the updating of the questions in step 5 (note that questions are included in B_{k+1}) are generally controlled by some **supra-utility** $\mathcal{U}(U_k^\dagger, B_{k+1}, C_k)$. This is a function of the (estimated) utility $U_k^\dagger = U_k(a_k^*|y_{d_k^*}, d_k^*)$ achieved at the end of step k, the updated state of knowledge B_{k+1}, and C_k which is some measure of accumulated time and/or cost at the end of step k. $\mathcal{U}$ embodies any overall *raison d'etre* of the interrogation procedure, and all logistical or cost constraints.

For example, if questions Q are sufficient to describe the *raison d'etre* of the interrogation problem, then if the sequence is scheduled to stop after a prechosen number of steps K, then the 'cost' is the number of steps $C_k = k$ and the supra-utility is simply $\mathcal{U}(C_k) = C_k$. Termination occurs once $\mathcal{U}$ reaches the value K.

Alternatively termination may occur if the achieved step k utility $U_k^\dagger = U_k(a_k^*|y_{d_k^*}, d_k^*)$ is sufficient (large enough according to some criterion). For the negative squared error utility such a stopping criterion is equivalent to estimates having a variance below some specified tolerance. In this situation the supra-utility is $\mathcal{U} = U_k^\dagger$.

To continue the example of the interrogation of a human suspect the iterations represent successive lines of questioning. In that case $U_k^\dagger$ would likely represent a measure of the strength of evidence of guilt of a crime; however, the nature of the crime being investigated may change as questions Q_k evolve at successive iterations, as investigators become aware of new information. The invariant goal of the investigator is to Solve the Crime, and this is encoded in the supra-utility $\mathcal{U}$, which is some measure of whether $U_k^\dagger$ has exceeded a threshold where the evidence is expected to pass muster in a court of law, and would also depend on any time limits during which suspects can be held for questioning without a specific charge being brought. Hence, if cost C_k measures the cumulative time for rounds of questioning 1 to k, then $\mathcal{U} = f(U_k^\dagger, C_k)$ for some chosen function f.

4.3 Updating the state of knowledge

The updating process from $B_k, y_{d_k^*}|d_k^*$ to B_{k+1} in principle requires the respecification of every component of the state of knowledge from Section 4.1.

Changing the questions will result in an alteration to almost every part of the state of knowledge. In particular the target and answer spaces may differ completely from the previous step, as will the design space and the utility.

The space of models $\mathbb{M}_{k+1}$ may be the same as $\mathbb{M}_k$. However we may eliminate some models now thought to be impossible, and might include new models which are more complex or more detailed versions of models in $\mathbb{M}_k$.

For each model $m \in \mathbb{M}_{k+1}$ we need to define priors for the models and priors on the parameters of those models. This is straightforward in the case that the spaces of models $\mathbb{M}_k$ and $\mathbb{M}_{k+1}$ are identical, in which case we use Bayesian updating in which the posteriors from step k are the priors for step $k+1$:

- the model prior $p_{k+1}(m) = p_k(m|y_{d_k^*}, d_k^*)$;
- the parameter space Θ_m is unchanged;
- the parameter prior $p_{k+1}(\theta_m|m) = p_k(\theta_m|y_{d_k^*}, d_k^*, m)$.

If we eliminate some models now thought to be impossible, only the first bullet above changes to $p_{k+1}(m) = 0$ for those m that are dropped, and otherwise $p_{k+1}(m) = Rp_k(m|y_{d_k^*}, d_k^*)$ for normalisation constant R that ensures that $p_{k+1}(m)$ still sums to unity over all m. While the parameter prior need not change, some of the values of $p_{k+1}(\theta_m|m)$ become redundant if the corresponding $p_{k+1}(m) = 0$. If nevertheless calculated, such redundant information therefore becomes part of the background stock of knowledge Ω_{k+1}.

If the questions and spaces of models remain the same in iteration $k+1$ then the target functions $T_{k+1}(\theta_m|m)$ will be the same as $T_k(\theta_m|m)$, and $\mathbb{T}_{k+1} = \mathbb{T}_k$. Otherwise new target functions must be specified for new models added to the space of models, or these functions may change completely if the questions have changed. If the questions Q_{k+1} are a subset of Q_k, some questions having been answered satisfactorily, the new target space may simply reduce in dimension.

For any pair of design $d \in \mathbb{D}_{k+1}$ and model $m \in \mathbb{M}_{k+1}$ that already existed at step k, then the likelihood function $f(y_d|d, \theta_m, m)$ must be the same as at step k. Any new pair (d, m) not seen before must have a new likelihood constructed. However, conditional on (d, m) this construction does not otherwise depend on the state of knowledge.

The stock of knowledge Ω_{k+1} is updated to include any additional information from the observed data y_d that is informative but which lies outside the scope of the space of models $\mathbb{M}_{k+1}$.

Even if it can neither be enumerated nor even described other than conceptually, the role of Ω_k is important. It represents the background knowledge and human experience that does not seem relevant at step k, but which nevertheless provides the information required to pose new questions, introduce new models, contemplate new possible answers and introduce new designs. In other words it describes the knowledge and experience base that may underpin human *inspiration*. Note also that there are situations where Ω_k can be enumerated and represented explicitly. For example, an artificially intelligent robot might store a parameterised form of information collected during its experience to-date. This information might be partitioned into that deemed relevant and irrelevant for a particular problem at hand; Ω_k represents the latter partition element. This is important in a variety of situations, for example if interrogation theory is used to describe robotic decision-making over very long time periods where accumulated experience is significant, and where inspiration is important and must be simulated, such as will likely be required in future unmanned space missions to investigate other planets (e.g., for autonomous geological mapping see Candela et al. 2017). This is discussed further below.

The simplest case of all is that our models, questions and possible answers do not change from step to step – and therefore that there is no inspiration after the initial setup of the problem. That implies that the priors are updated using Bayes rule, the utility function remains the same throughout, and there is no change to the state of knowledge Ω_k which remains fixed at its initial value Ω_1. In that case:

$$\begin{aligned} B_{k+1}(B_k, y_{d_k^*}|d_k^*) \;\; = \;\; & (\Omega_1, Q, \mathbb{M}, \mathbb{T}, \mathbb{A}, \mathbb{D}, \\ & \{p_{k+1}(m) = p_k(m|y_{d_k^*}, d_k^*), \\ & \{p_{k+1}(\theta_m|m) = p_k(\theta_m|y_{d_k^*}, d_k^*, m), T(\theta_m|m)|\theta_m \in \Theta_m\}|m \in \mathbb{M}\}, \\ & \{U(a|t,d)|a \in \mathbb{A}_k, t \in \mathbb{T}_k\}) \end{aligned} \tag{51}$$

A sequential interrogation set up in this way is also the simplest candidate for automation e.g., when inspiration from an autonomous robotic system is neither necessary nor desired.

4.4 Example Sequential Interrogation

As an example of a sequential investigation we consider the case of a subsurface Earth resources company reviewing its portfolios of assets in the Earth's subsurface (e.g., water, hydrocarbon or CO_2 reservoirs; ore bodies or mineral assets). It is interested in assessing the proportion of such assets that were improperly assessed by asset teams at the exploration phase – before the eventual economic success or failure of each asset was known. It wishes to make this assessment by inspecting the original paperwork for each asset for any significant irregularities or deficiencies in procedures followed. This task is costly so the company decides to take a random sample of assets to inspect, and to stratify the

sample across those assets that were eventually successful and those that were unsuccessful in terms of providing a resource that proved economically feasible to produce or use. The company decides to take independent samples from each of these two portfolios of assets.

There are N_j assets in each portfolio j, ($j = 1, 2$) and the (unknown) proportion of assets with irregularities in portfolio j is θ_j. Say the overall proportion of assets with irregularities is the quantity of interest

$$T = \frac{N_1\theta_1 + N_2\theta_2}{N_1 + N_2} = c_1\theta_1 + c_2\theta_2 \,, \tag{52}$$

with $c_j = N_j/N$, $N = N_1 + N_2$ and $c_1 + c_2 = 1$. The overall aim is to estimate this proportion as accurately as possible, given a fixed overall budget that will allow a total of n individual asset inspections. To achieve the maximum efficiency the allocation of the n assets between the two portfolios must be chosen to minimise the variance of the ultimate estimate of T. However that variance depends on the unknown values of θ_1 and θ_2. Therefore the company plans a sequential analysis. At each stage a sample is allocated between the two portfolios, the sample is collected, and new estimates of θ_1 and θ_2 are made. These estimates improve (become more precise) at each step, enabling an improved allocation of the sample at the next step. Thus we have a sequence of design problems: each one being the choice of the optimal design based on the accumulated evidence.

The space of models $\mathbb{M}_k$ at every step k contains the same single model m: there are two possibly distinct proportions θ_1 and θ_2 of assets with irregularities. The parameter space is $\Theta_k = [0, 1] \times [0, 1]$.

At Step 1 the prior for θ_j is assumed to be a Beta$(\alpha_{1j}, \beta_{1j})$ distribution, for $j = 1, 2$ (see e.g. Gelman et al. (2013) for a definition of the Beta distribution). The function $T(\theta_1, \theta_2)$ is defined in equation (52) above, and the answer to the question is an estimate of T with a squared error loss function chosen to define the utility. We therefore use the method and results laid out in Section 2.3.

At step k the design space $\mathbb{D}_k$ consists of possible allocations of n_k units between the two portfolios, with n_{kj} samples taken from portfolio j, for $j = 1, 2$, and $n_{k1} + n_{k2} = n_k$. We assume that K steps will be taken, and that the size of the sample n_k to be allocated at step k has been fixed in advance, with $\sum_{k=1}^{K} n_k = n$. We discuss the choice of K and n_k further below.

Within portfolio j there are n_{kj} assets inspected, chosen by a simple random sample of assets not inspected so far, of which y_{kj} are found to have irregularities or deficiencies. Assuming independence among the assets within portfolios these data can be assumed to arise from Binomial probability distributions (Gelman et al. 2013):

$$y_{kj}|\theta, d_k \sim \text{Binomial}(n_{kj}, \theta_j) \quad \text{for } j = 1, 2 \text{ and } k = 1, \ldots, K. \tag{53}$$

We assume that the proportion of assets to be sampled is small compared to the number of assets in either portfolio, so no finite-population corrections are necessary. At the first step this standard Beta-

Binomial set up means that the posterior for θ_j is a Beta$(\alpha_{1j} + y_{1j}, \beta_{1j} + n_{1j} - y_{1j})$ (Gelman et al. 2013). This posterior serves as the prior Beta$(\alpha_{2j}, \beta_{2j})$ at the second step, with $\alpha_{2j} = \alpha_{1j} + y_{1j}$ and $\beta_{2j} = \beta_{1j} + n_{1j} - y_{1j}$. The posterior at step k is used as the prior at step $k+1$, so it follows that the priors at every step are

$$\theta_j | B_k \sim \text{Beta}(\alpha_{kj}, \beta_{kj}) \tag{54}$$

with

$$\alpha_{kj} = \alpha_{1j} + \sum_{\ell=1}^{k-1} y_{\ell j} \quad \text{and} \quad \beta_{kj} = \beta_{1j} + \sum_{\ell=1}^{k-1} (n_{\ell j} - y_{\ell j}) \tag{55}$$

The optimal answer at step k (i.e. the best estimate of T), is given by equation (7), which in this case is

$$a_k^*(y_{d_k}, d_k) = \sum_{j=1}^{2} c_j \frac{\alpha_{(k+1)j}}{\alpha_{(k+1)j} + \beta_{(k+1)j}} \tag{56}$$

with optimised utility from equation (8):

$$U_k^*(y_{d_k}, d_k) = -\sum_{j=1}^{2} c_j^2 \frac{\alpha_{(k+1)j}\beta_{(k+1)j}}{(\alpha_{(k+1)j} + \beta_{(k+1)j})^2(\alpha_{(k+1)j} + \beta_{(k+1)j} + 1)} \tag{57}$$

The optimal design at step k maximises the a priori expected utility

$$U_k^*(d_k) = -\sum_{j=1}^{2} \frac{\lambda_{kj}}{(\alpha_{kj} + \beta_{kj} + n_{kj})} \tag{58}$$

with respect to design $d_k = n_{k1}$ (since $n_{k2} = n_k - n_{k1}$), where

$$\lambda_{kj} = c_j^2 \frac{\alpha_{kj}\beta_{kj}}{(\alpha_{kj} + \beta_{kj})(\alpha_{kj} + \beta_{kj} + 1)} \, . \tag{59}$$

By setting derivatives of $U_k^*(d_k)$ to zero, this optimisation problem reduces to the solution n_{k1} of the quadratic equation

$$\lambda_{k1}(\alpha_{k1} + \beta_{k1} + n_k - n_{k1})^2 - \lambda_{k2}(\alpha_{k2} + \beta_{k2} + n_{k1})^2 = 0 \, . \tag{60}$$

In the case when $c_1 = c_2$, $\alpha_{k1} = \alpha_{k2}$ and $\beta_{k1} = \beta_{k2}$, this gives the simple solution $n_{k1} = n_k/2$, in other words equal allocation. This would be the likely situation at step 1 if the priors for the two portfolios are the same and if the two portfolios were the same size ($N_1 = N_2$). The solution to equation (60) is not guaranteed to be an integer; however the solution n_{k1} can just be rounded to the nearest whole number to produce a physically implementable design.

The sample sizes n_k at each step were fixed in advance in this example, as was the number of steps K. At one extreme the whole exercise could be done in one step, with $n_k = n$ and $K = 1$. At the other extreme we could have $n_k = 1$ and take $K = n$ steps, at each step allocating the next sample member to the porfolio which will increase the utility the most. This latter approach will yield the best

possible result in terms of the utility, but is likely to be ruled out as impractical. Instead the number of steps K and workloads n_k may be chosen in advance for convenience and ease of administration, or can be designed by maximising the expected utility as in equation 58.

Note that an alternative stopping rule is to fix the n_k in advance, but have the number of steps K undetermined. The supra-utility $\mathcal{U}$ is then equal to the optimised achieved utility $U_k^*(y_{d_k}, d_k) = -\text{Var}[T|y_{d_k}, d_k]$ from equation (57), and the process terminates once this exceeds some chosen value $-V_0$, where $\sqrt{V_0}$ is the desired maximum standard error of the estimate of T.

After initial assessments of the proportion of incorrectly assessed assets in equation (52), at some step the company might decide that it would also like to know the answer to a related but different question Q: "Is improper asset evaluation associated with the success rate of the asset portfolio?" This question can be answered by finding the values of both θ_1 and θ_2 in order to assess any differences in their values. Our goal is then to estimate the vector $T = (\theta_1, \theta_2)^T$ with minimum variance, and we must specify the weight matrix W in equation 5 in the construction of the utility. If we set W to be diagonal then the relative sizes of the (non-negative) diagonal entries W_{11} and W_{22} characterise the importance we place on knowing θ_1 and θ_2. It follows directly that the solutions to the design and decision problems are the same as the above (equations (56)–(60)), but with c_1 and c_2 replaced with $\sqrt{W_{11}}$ and $\sqrt{W_{22}}$ respectively.

5 DISCUSSION

Many different types of problems can be addressed using the schema in Figures 1 and 2. Herein the worked examples were all linear in the parameters Θ to be estimated, but the general approach and expressions presented in Section 2 also apply to nonlinear problems. Hence, for example, nonlinear inversion, model selection, and experimental design problems fit equally well within the same interrogation methodology.

The theory presented herein appears highly structured and formalised. Nevertheless, it can be used to represent processes that are apparently less structured such as human discursive decision-making, interrogation or elicitation. For example, the study of Polson & Curtis (2010, 2015) recorded the dynamics of the uncertainty perceived by individual geoscientific experts in a sequential manner – before, during and after a group elicitation session. That session was carried out to assess the suitability of a particular analogue geological reservoir for subsurface storage of CO_2 for climate change mitigation, and the result showed the dynamic state of subjective opinions, even of experts with a well defined common data set (Bond et al. 2007, 2012; Curtis 2012). In a separate expert elicitation experiment in 2011, the same authors applied their methodology with a group of six other experts. The investigator in the exercise was interested in the answer to the question Q_1 : "Should we *promote* carbon capture

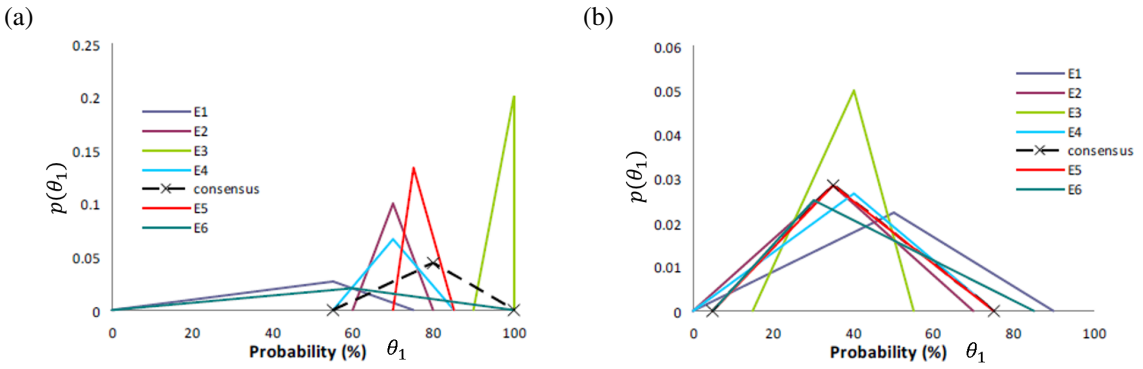

Figure 5. Individual and consensus probability distributions for $p(\theta_1)$ elicited from six experts (E1-E6) under (a) design d_1, and (b) design d_2 discussed in the text. From a study similar to that of Polson & Curtis (2010, 2015).

and subsurface storage as a technology for climate change mitigation *now*?" The relevant state of nature θ_1 is the truth of this statement: either $\theta_1 = 0$ if it is false, and $\theta_1 = 1$ if it is true. The investigator had no expert opinion about this, and at the outset assigned a prior $p(\theta_1 = j) = \frac{1}{2}$ to both options $j = 0, 1$.

The investigator then called together a panel of experts, and following a discussion, collected both their individual views and their group consensus view about θ_1. This discussion was an experiment, with design d_1 (which at this stage was simply to elicit their views about θ_1), and their responses constituted the investigator's data y_{d_1}. These data were translated by the investigator, using averaging techniques appropriate to the synthesis of expert opinion (see Polson & Curtis (2010)), into a posterior probability distribution for θ_1: $p(\theta_1|y_{d_1})$.

However the result of this exercise was judged unsatisfactory, because the views of the individual experts were highly divergent (Figure 5a). The reason for this was apparently that during the discussion d_1 the experts had found it difficult to answer the question about θ_1 directly, because it depended on the related question of whether alternative low carbon technologies or strategies (such as the development of new nuclear power options, or energy use reduction) would also be promoted fairly.

In response to this situation a new subquestion was added: Q_2 : "What is the probability that alternative low carbon technologies will also be promoted fairly?" Associated with this question is the state of nature parameter θ_2 – the probability that alternative technologies would be promoted fairly: $\theta_2 = 1$ if alternative technologies would be promoted fairly, and $\theta_2 = 0$ if not. The investigator's prior for this parameter was $p(\theta_2 = k) = \frac{1}{2}$ for both options $k = 0, 1$.

The discussion was resumed in a second iteration. Experimental design d_2 was designed to first elicit an answer to Q_2 (the value of θ_2); then elicit the answer to Q_1 (the value of θ_1) conditional on each possible answer to Q_2. The content of the second discussion resulted in data y_{d_2}. These data

were first used to construct the posterior distribution $p(\theta_2|y_{d_2}, y_{d_1})$ thus providing an answer to Q_2. Secondly, the conditional posterior distribution $p(\theta_1|\theta_2, y_{d_2}, y_{d_1})$ was constructed for each of the two scenarios $\theta_2 = 0$ or 1. Finally, the marginal posterior for θ_1, the ultimate goal of the exercise, and the answer to Q_1, was calculated by

$$\begin{aligned} p(\theta_1|y_{d_2}, y_{d_1}) &= p(\theta_1|\theta_2 = 0, y_{d_2}, y_{d_1})p(\theta_2 = 0|y_{d_2}, y_{d_1}) \\ &\quad + p(\theta_1|\theta_2 = 1, y_{d_2}, y_{d_1})p(\theta_2 = 1|y_{d_2}, y_{d_1}) \end{aligned} \tag{61}$$

While the true answer to this overall question is unknown (and may not exist in any objective sense), the experts stated that they felt better able to estimate the probabilities elicited in design d_2 than those in design d_1. This may explain why results were significantly more consistent (the inter-expert variance was significantly reduced) using the second approach (Figure 5), as subjective biases that occur in highly uncertain situations are reduced (Bond et al. 2007, 2012; Curtis 2012). What is more, the experts' overall tendency was to promote CO_2 storage under design d_1 (estimates of θ_1 are on average greater than 50% in Figure 5a) whereas they slightly prefered not to promote CO_2 storage under design d_2 (estimates of θ_1 in Figure 5b are on average lower than 50%). This illustrates how critical the design of experiments is to the results of such elicitation experiments. Moreover this example demonstrates the manner in which a realisation on the part of the interrogator during the interrogation can allow that interrogator to fruitfully change the course of the inquiry by posing a new question and collecting data under a new experiemental design.

This example shows how a common, discursive elicitation method to solve a problem can be represented within the interrogation schema. The additional advantage of the schema is that it provides a basis to optimise such elicitation sessions in future: elicitation methods such as that of Polson & Curtis (2010) which provide quantitative, probabilistic outcomes based on parameterisations of a problem, fit perfectly within the interrogation schema. This would allow, for example, the experimental design – the choice of which data to elicit from the experts (which verbal questions to pose) – to be optimised in each iteration (e.g., Coupé & Van der Gaag 1997; Curtis & Wood 2004). Alternatively, the schema could be used to divide the experts into two groups in order to assess different parameters, or to provide independent estimates of the same parameters. Methods similar to those in the asset portfolio example above could be used to decide how many experts should be in each group, or which parameters each group should estimate.

The above elicitation example illustrates the critical role of *inspiration* within interrogation problems. In iteration 1, the focus was not initially on factors controlling whether the experts perceived other technologies to be adequately promoted. Hence, parts of their background knowledge seemed irrelevant, and were represented by Ω_1. However, it became clear during the course of that iteration

that these were relevant factors; hence in iteration 2 some of the knowledge in Ω_1 was used to re-parameterise the problem to construct the new parameterisation θ_2. Other aspects of their knowledge in Ω_1, such as whether they perceived politicians to give sufficient weight to alternative technologies were then considered relevant. Since this knowledge was irrelevant in iteration 1 and explicitly relevant in iteration 2, it was therefore within set Ω_1 but not Ω_2, and that transfer of information out of Ω_1 is the mathematical representation of inspiration.

In that example, the overall goal of the investigation did not change between iterations (the ultimate goal was still to estimate θ_1). Within interrogation theory, the extent to which iteration k of the elicitation achieved the overall goal might be assessed by calculating the inter-expert variation of estimates of distribution $p_k(\theta_1)$. This could be represented by a supra-utility $\mathcal{U} = -Var[\theta_1|\mathbf{y}]$ where the variance is calculated across experts. The high variance of results from iteration 1 represented a low supra-utility, inspiring the investigator to change the questions Q in iteration 2. The results from iteration 2 exhibit far lower inter-expert variation, producing a higher supra-utility.

Varying questions is an important flexibility in many interrogation scenarios, not least because this potentially allows the interrogation system, including inspirational transitions, to be semi-automated. For example, consider the area of machine learning. In the case of robotics, autonomous systems may need to interrogate their surroundings by sensing and moving, then update all aspects of the subsequent problem to be solved, based both on their findings and on their overall objectives (Ramamoorthy et al. 2013). Consider a robot with 'senses' of directional sight and directional hearing, with an overall objective (embodied more precisely within its supra-utility): "Arrive at a source of a particular specified sound, while conserving as much battery power as possible". Initially the relevant question might be, "What is the optimal direction to move most directly towards the sound source?" Upon solving the initial Design Problem, the robot decides to spend power on collecting data by both hearing and visualising its environment (the experimental design). It performs that experiment and based on the data it solves the Decision Problem of choosing an optimal answer (a direction), and moves towards the sound. After a sequence of such steps, the robot senses that it approaches a barrier, on the other side of which the sound appears to originate. The relevant interrogation question at that point may change to, "Which direction of motion is most likely to result in circumnavigating the barrier, given the robot's limited motion capabilities?" Again a Design Problem would be solved to decide which data to collect, and some information in Ω_k (grey-scale values in a previously stored photograph) that was thought to be irrelevant now has a specific interpretation (a barrier) and hence becomes relevant – so is not included in Ω_{k+1}. After solving the resulting Decision Problem to estimate an optimal answer, the robot moves in one direction until it reaches the end of the barrier.

In the subsequent iteration it decides to perform an acoustic-only experiment to estimate the new

direction of the sound source. Upon collecting the acoustic data and solving the source location inverse problem it becomes clear that there are in fact at least *two* distinct sources of the sound. Based on the supra-utility, the question is then updated to, "Which of the sources can be investigated with minimal power consumption?" On solving the Design Problem, the robot may design a visual experiment to identify objects that may be responsible for each sound source, then solve the Decision Problem to decide which object could be approached with minimal power consumption (cost). Thereafter the interrogation might change the question back to, "What is the optimal direction to move towards the sound source?", similarly to the initial iteration. The robot eventually arrives at a noise source and stops, having achieved its objective.

An interrogation system that is formalised, numerically implementable, and which allows all of questions, parameters and experiments to be updated sequentially, is thus crucial for certain applications, illustrating the importance of the theory presented herein. The supra-utility is key to being able to make choices of appropriate questions at each iteration of the algorithm in Figure 2. Unexpected occurrences during interrogations at one iteration (e.g., finding two noise sources instead of one) may entirely change the most relevant questions at the next. This concept of 'relevance' must be defined in terms of the overall objective or *raison d'etre* (above, of the robot), and this is embodied in the supra-utility.

The decision about whether to stop a sequential interrogation may be controlled by many forms of stopping criteria. In the example above, the stopping criteria may either be that the robot has reached the sound source, or that battery power is too low to continue safely. In the elicitation or interrogation of a human (in the common-parlance use of the word interrogation), a stopping criterion may embody a trade-off between the fatigue that humans experience after several iterations making them more prone to errors, against the value of additional knowledge that is expected to be gained in the next iteration. Whatever the circumstances, the criterion used must depend on whether the overall interrogational objective has been achieved (utility maximised), and this is also embodied within the supra-utility.

Interrogation theory therefore provides a useful overarching context for fields of inversion, design and decision theory. These theories may thus be applied more efficiently, since interrogation theory focusses experimental and inversion effort only on the parts of data and solutions that are relevant to questions posed. However, interrogation theory also extends these fields theoretically, particularly by the introduction and formalisation of (i) the target space which embodies relevant answers to our particular questions, (ii) the changing state of knowledge in sequential studies, (iii) the potential to consider different parameterisations of information, and (iv) the overall *raison d'etre* of the interrogation procedure which allows for inspirational changes in the questions posed. How each of these

should be implemented in the case of specific problems in different areas of application will no doubt be resolved in future studies.

ACKNOWLEDGMENTS

The authors thank Debbie Polson and Siobhan Prise for their leading contributions to the elicitation experiment described in the Discussion, and Ian Main for suggesting the rock physics example used to illustrate model discrimination. This work was carried out during RA's sabbatical stay at the University of Edinburgh, UK, and in part during AC's visits to ETH Zurich, Switzerland.

REFERENCES

Bodin, T. & Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm, *Geophysical Journal International*, **178**, 1411–1436.

Bond, C. E., Gibbs, A. D., Shipton, Z. K., & Jones, S., 2007. What do you think this is?: 'Conceptual uncertainty' in geoscience interpretation, *GSA Today*, **17**(11), 4–10.

Bond, C. E., Lunn, R. J., Shipton, Z. K., & Lunn, A. D., 2012. What makes an expert effective at interpreting seismic images?, *Geology*, **40**(1), 75–78.

Boukharov, G. N., Chanda, M. W., & Boukharov, N. G., 1995. The Three Processes of Brittle Crystalline Rock Creep, *Int. J. Rock Mech. Min. Sci. & Geomech. Abstr.*, **32**(4), 325–335.

Brockmann, H. J. & Dawkins, R., 1979. Joint nesting in a digger wasp as an evolutionarily stable preadaptation to social life, *Behaviour*, **71**, 203–245.

Burnham, K. P. & Anderson, D. R., 2002. *Model selection and multi-model inference: a practical information-theoretic approach*, Springer, New York, 2nd edn.

Candela, A., Thompson, D., Dobrea, E. N., & Wettergreen, D., 2017. Planetary Robotic Exploration Driven by Science Hypotheses for Geologic Mapping, Proceedings of the IEEE/RSJ IROS.

Chaloner, K. & Verdinelli, I., 1995. Bayesian Experimental Design: A Review, *Statistical Science*, **10**, 273–304.

Coupé, V. M. H. & Van der Gaag, L. C., 1997. Supporting Probability Elicitation by Sensitivity Analysis, in *Lecture Notes in Computer Science: Knowledge Acquisition, Modeling and Management*, vol. 1319, pp. 335–340, Springer.

Curtis, A., 2004a. Theory of model-based geophysical survey and experimental design. Part A: Linear Problems, *The Leading Edge*, **23**(10), 997–1004.

Curtis, A., 2004b. Theory of model-based geophysical survey and experimental design. Part A: Nonlinear Problems, *The Leading Edge*, **23**(10), 1112–1117.

Curtis, A., 2012. The science of subjectivity, *Geology*, **40**, 95–96.

Curtis, A. & Lomax, A., 2001. Prior information, sampling distributions and the curse of dimensionality, *Geophysics*, **66**, 372–378.

Curtis, A. & Wood, R., 2004. Optimal elicitation of probabilistic information from experts, in *Geological Prior Information*, vol. 239, pp. 127–145, Geological Society London, Special Publication.

Dawkins, R., 2015. *Brief Candle in the Dark: My Life in Science*, Bantam Press, London.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D., 2013. *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, FL, 3rd edn.

Maurer, H., Curtis, A., & Boerner, D., 2010. Recent advances in optimized geophysical survey design, *Geophysics*, **75**(5), 75A177–75A194.

Mosegaard, K. & Tarantola, A., 1995. Monte carlo sampling of solutions to inverse problems, *Journal of Geophysical Research: Solid Earth*, **100**(B7), 12431–12447.

Myung, J. I. & Pitt, M. A., 2009. Optimal Experimental Design for Model Discrimination, *Psychological Review*, **116**, 499–518.

Polson, D. & Curtis, A., 2010. Dynamics of uncertainty in geological interpretation, *Journal of the Geological Society*, **167**, 510.

Polson, D. & Curtis, A., 2015. Assessing individual influence on group decisions in geological carbon capture and storage problems, in *Collaborative knowledge in scientific research networks*, chap. 4, pp. 55–75, IGI Books.

Polson, D., Curtis, A., & Vivalda, C., 2012a. *Environmental Risk of Carbon Capture and Storage*, vol. 20 of **Advances in Environmental Research**, chap. 8, pp. 181–196, Nova, Hauppauge NY.

Polson, D., Curtis, A., & Vivalda, C., 2012b. The Evolving Perception of Risk During Reservoir Evaluation Projects for Geological Storage of CO_2, *International Journal of Greenhouse Gas Control*, **9**, 10–23.

Ramamoorthy, S., Mahmud, M. M. H., Rosman, B., & Kohli, P., 2013. Latent-Variable MDP Models for Adapting the Interaction Environment of Diverse Users, Tech. Rep. http://wcms.inf.ed.ac.uk/ipab/autonomy/publications/interaction.pdf, Technical Report, School Of Informatics, University of Edinburgh.

Sambridge, M., Gallagher, K., Jackson, A., & Rickwood, P., 2006. Transdimensional inverse problems, model comparison and the evidence, *Geophysical Journal International*, **167**, 528–542.

Snieder, R., 1998. The role of nonlinearity in inverse problems, *Inverse Problems*, **14**, 387–404.

Tarantola, A., 2005. *Inverse Problem Theory and methods for model parameter estimation*, Society for Industrial and Applied Mathematics, Philadelphia.

Vehtari, A. & Ojanen, J., 2012. A survey of Bayesian predictive methods for model assessment, selection and comparison, *Statistics Surveys*, **6**, 142–228.

Young, G. A. & Smith, R. L., 2005. *Essentials of statistical inference*, Cambridge University Press, Cambridge, UK.