Statistical inference
Part 6
Perspectives on inference

Michael Goldstein
Durham University

Perspectives

In this section, we examine aspects of the assumptions and methods of statistical inference.

Firstly, we consider the meaning of the parameters in our models.
(We will use ideas of exchangeability.)

Then we look at the logic underlying the inferences within our models and suggest complementary approaches.

Finally, we consider practical approaches for dealing with problems of scientific inference for large and complex models.

Epistemic and aleatory uncertainty

Uncertainty is often considered to be divided into two basic types.

Epistemic uncertainty relates to our lack of knowledge.

This could be reduced by receipt of further information.

Aleatory uncertainty relates to intrinsic chance variation in the system.

This cannot be resolved except by direct observation.

Much of science can be viewed as the attempt to move uncertainty from the aleatory to the epistemic form, where it can be better understood and, possibly, reduced.

Uncertainty in a Bayesian analysis

For example, suppose that we spin a coin.

We are uncertain about the outcome because

(i) we don't know the underlying probability that the coin will land heads

[This is epistemic uncertainty]

(ii) even if we did know the underlying probability of heads, we would not know the outcome of the individual spin.

[This is aleatory uncertainty.]

Bayesian representation

In a typical Bayesian analysis:

Aleatory uncertainty is expressed through the likelihood function for the data given the population parameters,

Epistemic uncertainty is expressed through the prior distribution over the parameters.

This division is helpful for thinking about our uncertainty about a physical system. However, all that we observe are individual measurements of real things. The parametric forms that we introduce are models whose relationship to real world uncertainty remains to be established.

Within the subjectivist approach, there is a precise meaning for this relationship, rooted in the judgement of **exchangeability**.

Models and exchangeability

Exchangeability allows us to construct parametric statistical models purely on the basis of natural uncertainty statements about observable random quantities.

In many cases, the argument shows that we have no choice but to behave as though we consider that we are

(i) sampling from a parametric model
(the aleatory uncertainty)

(ii) given the true but unknown values of some population distribution
(the epistemic uncertainty).

Therefore, exchangeability offers a real world meaning for uncertainty quantification.

Exchangeable coin spins

Suppose that we make a series of coin spins $X_1, X_2, ...$ where $X_i = 1$ if the ith toss is a head, and $X_i = 0$ otherwise.

We say that the coin spins are **exchangeable** (to you) if, for each m, your joint probability distribution of any collection of m of the spins, $X_{i_1}, ...X_{i_m}$ say, is the same.

So, for example, you would assign the same probability distribution for (X_1, X_2, X_3) as you would for (X_5, X_7, X_9)

For the sequence of coin spins, you may make the judgement of exchangeability directly, rather than deducing it from some formal model.

Exchangeability: definition

This is a special case of the following general definition.

Definition

A sequence $(Y_1, Y_2, \ldots)$ *of random vectors* $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im})$ *takes values in some space* Ω*.*

The sequence is said to be **exchangeable** *if, for each* n*, the joint probability distribution of each subcollection of* n *quantities* $(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_n})$ *is the same.*

The representation theorem for coin spins

$X_1, X_2, \ldots$ is an infinite exchangeable sequence of coin spins.

W_n is the number of heads in the first n spins

There is a probablity measure $F(q)$ on [0,1] such that, for each k, n, you must assign the probabilities for W_n as follows:

$$\mathrm{P}(W_n = k) = \int_0^1 \binom{n}{k} q^k (1-q)^{(n-k)} \, dF(q)$$

This is de Finetti's theorem for an infinite exchangeable sequence of binary outcomes.

In many cases, this simplifies to the form

$$\mathrm{P}(W_n = k) = \int_0^1 \binom{n}{k} q^k (1-q)^{(n-k)} f(q) dq$$

Interpreting the representation theorem

$$\mathrm{P}(W_n = k) = \int_0^1 \binom{n}{k} q^k (1-q)^{(n-k)} f(q) dq$$

The theorem shows that the judgement of exchangeability, alone, is sufficient to ensure that your beliefs about the sequence are just as if you consider that

(i) there is a true but unknown quantity q, for which you have prior distribution $f(q)$.

(ii) given the value of q, you view the sequence as a series of independent Bernoulli trials each with probability q.

(iii) having observed a sequence of n spins, and obtaining k heads, your posterior beliefs for the next m spins, W_m, are found by Bayes theorem as

$$\mathrm{P}(W_m = j|k, n) = \int_0^1 \binom{m}{j} q^j (1-q)^{(m-j)} f(q|k, n) dq$$

Adding coin flips to a bucket

Here is the informal idea underlying the proof of the representation theorem.

Consider the following thought experiment.

Imagine spinning the coin N times (N large).

Each time that you spin heads, add a counter marked H to a bucket.

Each time that you spin tails, add a counter marked T.

Suppose that the bucket has M counters marked H, (N-M) counters marked T.

From exchangeability, your probability that a single spin of the coin lands heads is the same as your probability that a single pick of a counter from the bucket is marked H.

Informal version of representation theorem

Your uncertainty about the single spin is therefore the combination of

(i) your uncertainty about the value of M/N and

(ii) your uncertainty about picking a H counter for given value of M/N.

In this representation

(i) corresponds to your prior distribution for q, the probability of heads.

(ii) corresponds to your binomial likelihood for the spin outcome, given q.

The argument is similar for m spins of the coin with $m << N$.

The technical part of the proof is checking that, as $N \to \infty$, the above construction converges to the probability specification given in the theorem.

The general form for the exchangeability representation

The coin spinning example is a special case of the generalization given by Hewitt and Savage, which is as follows.

Theorem *Let* $(Y_1, Y_2, \ldots)$ *be an infinite exchangeable sequence of random quantities with values in* Ω*. Then there exists a probability measure* F *on the set of probability measures* $Q(\Omega)$ *on* Ω*, such that, for each* n*, and subsets* $A_1, \ldots, A_n$ *of* Ω*,*

$$P(Y_1 \in A_1, \ldots, Y_n \in A_n) = \int Q(A_1) \ldots Q(A_n) dF(Q),$$

F is the limiting distribution of the empirical measure, i.e., the probability assigned to any set A by F is given by the limit of the probability assigned to the proportion of the first N trials whose outcome is in A.

For discussion and references see:

Goldstein (2013) Observables and Models: exchangeability and the inductive argument, in Bayesian Theory and Applications Paul Damien (ed.) et al. Oxford

On foundations

Now let’s discuss how the foundations of statistical inference relate to real world inference.

Here are two quotes from the hugely important book on statistical foundations, “The foundations of statistics” by L.J. Savage (1954).

“It is often argued academically that no science can be more secure than its foundations, and that, if there is controversy about the foundations, there must be even greater controversy about the higher parts of the science.”

“It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel.”

Let’s consider some real world contexts for this.

Thermohaline shutdown

Global ocean circulation is driven by winds and the exchange of heat and water vapour at the sea surface.

Wind driven surface currents head polewards from the equator, cooling all the while.

Having lost much of its heat, the surface water becomes so salty that it is dense enough to sink.

The return flow occurs at the bottom of the North Atlantic.

The water masses transport heat energy around the globe, which has a large impact on the climate of our planet.

Some atmosphere-ocean models show a slowdown of thermohaline circulation in simulations of the 21st century with the expected rise in greenhouse gases.

[This is due to a combination of effects which reduce the density of surface waters, which makes it harder for them to sink.]

The meaning of an uncertainty analysis

Statistical analysis, for example in climate research, results in a collection of uncertainty statements.

What do these uncertainty statements mean?

This quote from the BBC web-site is typical:

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

This means what exactly?

RAPID-WATCH

What are the implications of RAPID-WATCH observing system data and other recent observations for estimates of the risk due to rapid change in the MOC? In this context risk is taken to mean the probability of rapid change in the MOC and the consequent impact on climate (affecting temperatures, precipitation, sea level, for example). This project must:

* contribute to the MOC observing system assessment in 2011;
* investigate how observations of the MOC can be used to constrain estimates of the probability of rapid MOC change, including magnitude and rate of change;
* make sound statistical inferences about the real climate system from model simulations and observations;
* investigate the dependence of model uncertainty on such factors as changes of resolution;
* assess model uncertainty in climate impacts and characterise impacts that have received less attention (eg frequency of extremes).

The project must also demonstrate close partnership with the Hadley Centre.

Policy implications of uncertainties related to climate change

The Climate Change Act, 2008, constructed a legally-binding long-term framework for the UK to cut greenhouse gas emissions and a framework for building the UK's ability to adapt to a changing climate.

The Act requires a UK-wide climate change risk assessment (CCRA) that must take place every five years.

Also a national adaptation programme (NAP), setting out the Government's objectives, proposals and policies for responding to the risks identified in the CCRA.

The CCRA, and thus the NAP, drew heavily on the uncertainty analysis for future climate outcomes, published in 2009 by the Met Office as the UK Climate Projections UKCP09.

Uncertainty in climate projections

1.1.1 What do we mean by probability in UKCP09?

It is important to point out early in this report that a probability given in UKCP09 (or indeed IPCC) is not the same as the probability of a given number arising in a game of chance, such as rolling a dice. It can be seen as the relative degree to which each possible climate outcome is supported by the evidence available, taking into account our current understanding of climate science and observations, as generated by the UKCP09 methodology. If the evidence changes in future, so will the probabilities.

Subjective probability is a measure of the degree to which a particular outcome is consistent with the information considered in the analysis (i.e. strength of the evidence) ... Probabilistic climate projections are based on subjective probability, as the probabilities are a measure of the degree to which a particular level of future climate change is consistent with the evidence considered. In the case of UKCP09, a Bayesian statistical framework was used, and the evidence comes from historical climate observations, expert judgement and results of considering the outputs from a number of climate models, all with their associated uncertainties.

Uncertainty in practice

Many important areas of application are based on large modelling activities linked to complex data sets, involving many sources of uncertainty.

Our understanding of uncertainty is crucial for linking our model analysis with statements about the real physical systems.

Different sources of uncertainty may be quantified and analysed using different techniques.

However, all of these uncertainties must be combined into an overall uncertainty judgement about the physical system.

The natural way in which this combination can be understood is as a joint collection of uncertainty statements made by experts in the subject matter area, which leads naturally to an overall Bayesian formulation.

Subjectivist Bayes

In the subjectivist Bayes view, any probability statement is the uncertainty judgement of a specified individual, expressed on the scale of probability.

Therefore, we should speak not of the probability of rapid climate change, but instead of Anne's probability or Bob's probability of rapid climate change.

The aim of a subjectivist Bayes analysis is to develop a clear and logically well-founded methodology to support expert individuals in making such informed uncertainty judgements.

These should be made for reasons which are transparent and can be sufficiently well documented that the reasoning can be critically assessed by similarly knowledgeable experts.

If different expert individuals may reasonably reach different uncertainty judgements, then such differences are part of the analysis and should be explored and documented.

Objectivity and subjectivity:definition

Let’s consider the meaning of the terms. Here is how the The Internet Encyclopedia of Philosophy explains the distinction:

“Objective judgment or belief” refers to a judgment or belief based on objectively strong supporting evidence, the sort of evidence that would be compelling for any rational being.

A subjective judgment would then seem to be a judgment or belief supported by evidence that is compelling for some rational beings (subjects) but not compelling for others.

[Objectivity, D.H. Mulder,The Internet Encyclopedia of Philosophy,http://www.iep.utm.edu/]

Objectivity and subjectivity

This explanation corresponds to how the term is commonly used and understood and is why it is valued as a gold standard in science and elsewhere.

A common aim for a scientific Bayesian analysis is to ascertain whether the data is sufficiently convincing that any reasonable assignment of prior judgments would lead to roughly the same conclusions.

Calling such an outcome an “objective Bayesian analysis” is appropriate because the term “objective” attaches to the judgements that we reach and thus to the claims that we may make for the results of the analysis.

Bayes analysis as a model

Most careful discussions of foundations recognise the element of abstraction within the Bayesian formalism by invoking such considerations as the inferential behaviour of “perfectly rational individuals” or the “small worlds” account of Savage.

Such accounts effectively recognise Bayesian analysis as a model for inferential reasoning.

Just as climate scientists study climate by means of climate models, Bayesian statisticians study uncertainty judgements by means of Bayesian models.

These models are a crucial source of information and insight, but to treat the model inference as identical to, rather than as informative for, our actual inferences is to make the same mistake as it would be to conflate the climate model with climate itself.

Critical issues for Bayes analysis

$$\mathrm{P}(B|D) = \frac{\mathrm{P}(D|B)\mathrm{P}(B)}{\mathrm{P}(D)}$$

Issues

[1] We may not feel able to specify all of the quantities on the RHS.
(In most problems, there will be a huge number to specify.)

[2] The calculations to get to the LHS can be very complicated, so we may be unsure that they have been done correctly.

[3] Suppose that the problem is simple enough that we can specify the quantities on the RHS, and calculate precisely the quantity on the LHS. What does $\mathrm{P}(B|D)$ mean?

Why conditioning?

The foundational argument for using Bayes theorem is some version of the following.

Suppose that you specify a joint probability distribution for a collection of random quantities.

Suppose that you also write down a rule for changing your probabilities for some of the quantities, as a function of the numerical values of the remaining quantities.

If this rule for changing your probabilities is not Bayes theorem, then you can be made a sure loser, in the sense of accepting a sequence of bets constructed in such a way as to guarantee sure loss.

Conditional and posterior judgements

This is not a demonstration that beliefs should change by conditioning.

What it does is to eliminate non-Bayesian rules for updating beliefs in the class of rules based exclusively on currently specified beliefs and the values of the observables within the model.

What relevance do such conditional probabilities hold for your actual posterior probabilities in the real world, when you do learn of event outcomes?

A subjectivist framework

We consider two events A and B.

We make current judgements $\mathrm{P}(A), \mathrm{P}(B), \mathrm{P}(A|B)$.

At future time t we intend to observe whether B occurs, and then to revise our assessment for $\mathrm{P}(A)$ as $\mathsf{P}_t(A)$.

At this moment, we are unsure as to what else we may observe by t or even what inferential method we will follow to make the revised judgement.

Question What is the relationship between the prior values that we have specified and the posterior probabilities that we will subsequently assess?

Temporal sure preference

Suppose that you must choose between two random penalties, J and K. Suppose that at some future time the values of J and K will be revealed, and you will pay the penalty that you have chosen.

You have a **sure preference** for J over K at (future) time t, if you know now, as a matter of logic, that at time t you will not express a strict preference for penalty K over penalty J.

The temporal consistency principle that we impose is that future sure preferences are respected by preferences today. We call this the **temporal sure preference principle**, as follows.

The Temporal Sure Preference (TSP) principle *Suppose that you have a sure preference for J over K at (future) time t. Then you should not have a strict preference for K over J now.*

Conditional and posterior probabilities

We can show from this preference that now you must make the specification

$$P_t(A) = P(A|\mathcal{B}) + \mathcal{R}$$

where $P(A|\mathcal{B})$ is the conditioning of A on the partition $\mathcal{B}$ = (B, B^c), namely

$$P(A|\mathcal{B}) = P(A|B)B + P(A|B^c)B^c$$

where B, B^c are the indicator functions for the corresponding events,

and R is a further random quantity with

$$E(R) = E(R|B) = E(R|B^c) = 0.$$

Discussion

This corresponds to the interpretation that we have for mathematical models of physical systems.

A model forecast is useful in giving us a “mean forecast”, with associated variance, which reduces, but does not eliminate, our uncertainty about future system behaviour.

$$P_t(A) = P(A|\mathcal{B}) + \mathcal{R}$$

We may view $P(A|\mathcal{B})$ as providing a mean forecast for our future judgements, while the residual quantity R expresses the uncertainty in this mean forecast.

Informally, the larger the variance of $P(A|\mathcal{B})$ as compared to the variance of R, the more informative a formal Bayes inference based on conditioning on B is considered to be for the actual posterior judgement on A.

Generalisation

Bayesian analysis is a model for belief specification and revision with many good properties and a meaningful foundational interpretation.

However, it is not the only model with these properties and interpretation.

Rather it is a special case of a more general formulation, in which expectation, rather than probability is the primitive description of uncertainty.

That formulation may be better suited to the actual abilities and limitations of individuals in making meaningful prior specifications.

This formulation is termed Bayes linear statistics (as expectation is linear).

The Bayes linear approach

The Bayes linear approach combines prior uncertainties with observational data, using **expectation** rather than **probability** as the primitive for the theory (see de Finetti “Theory of Probability”, Wiley, 1974).

This distinction is of particular relevance in complex problems with too many sources of information for us to be comfortable in making a meaningful full joint prior probability specification of the type required for a Bayesian analysis.

The Bayes linear approach is similar in spirit to a full Bayes analysis, but is based on a simpler approach to prior specification and analysis, and so offers a practical methodology for analysing partially specified beliefs for large problems.

Belief adjustment

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

We intend to observe $\boldsymbol{D}$ in order to improve our assessments of belief over $\boldsymbol{B}$.

The *adjusted* expectation, $\mathrm{E}_{\boldsymbol{D}}(B_i)$ for B_i given $\boldsymbol{D}$ is the linear combination $\boldsymbol{a}_i^T\boldsymbol{D}$ minimising

$$\mathrm{E}((B_i - \boldsymbol{a}_i^T\boldsymbol{D})^2)$$

over choices of $\boldsymbol{a}_i$.

To find the adjusted expectation, we must specify prior mean vectors and variance matrices for $\boldsymbol{B}$ and $\boldsymbol{D}$ and a covariance matrix between $\boldsymbol{B}$ and $\boldsymbol{D}$.

(We make these specifications directly.)

Adjusted mean and variance

The adjusted expectation vector, $\mathrm{E}_{\boldsymbol{D}}(\boldsymbol{B})$, for $\boldsymbol{B}$ given $\boldsymbol{D}$, is evaluated as

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

The *adjusted variance matrix* for $\boldsymbol{B}$ given $\boldsymbol{D}$, denoted by $\mathrm{Var}_{\boldsymbol{D}}(\boldsymbol{B})$, is evaluated as

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

Note that if $\boldsymbol{D}$ comprises the indicator functions for the elements of a partition, i.e. where each D_i takes value one or zero and precisely one element D_i will equal one, then adjusting by $\boldsymbol{D}$ is equivalent to probabilistic updating.

Foundational interpretation for Bayes linear analysis

The temporal sure preference principle implies that your actual posterior expectation, $\mathrm{E}_T(\boldsymbol{B})$, at time T when you have observed $\boldsymbol{D}$, satisfies two relations

$$\boldsymbol{B} = \mathrm{E}_T(\boldsymbol{B}) + \boldsymbol{S}$$
$$\mathrm{E}_T(\boldsymbol{B}) = \mathrm{E}_{\boldsymbol{D}}(\boldsymbol{B}) + \boldsymbol{R},$$

where $\boldsymbol{S}$, $\boldsymbol{R}$ each have, a priori, zero expectation and are uncorrelated with each other and with $\boldsymbol{D}$.

Therefore, adjusted expectation is a prior inference for your actual posterior judgments, which resolves a portion of your current variance for $\boldsymbol{B}$.

For full treatment of this approach (foundations and methods), see
Bayes linear Statistics: Theory and Methods, 2007, Wiley
Michael Goldstein and David Wooff

Models and physical systems: some examples

Systems biology Models of activity at the cellular level are used to make inferences about the behaviour of the biological organism.

Oil reservoirs An oil reservoir simulator is used to manage assets associated with the reservoir, in order to develop efficient production schedules, etc.

Natural Hazards Floods, volcanoes, tsunamis and so forth, are all studied by large computer simulators.

Disease modelling Agent based models are used to study interventions to control infectious diseases.

Energy planning Simulators of future energy demand and provision are key components of planning for energy investment.

Climate change Large scale climate simulators are constructed to assess likely effects of human intervention upon future climate behaviour.

Galaxy formation The study of the development of the Universe is carried out by using a Galaxy formation simulator.

The science in each is completely different. However, the underlying methodology for handling uncertainty is the same.

Sources of Uncertainty

(i) parametric uncertainty (each model requires a, typically high dimensional, parametric specification)
(ii) condition uncertainty (uncertainty as to boundary conditions, initial conditions, and forcing functions),
(iii) functional uncertainty (model evaluations take a long time, so the function is unknown almost everywhere)
(iv) stochastic uncertainty (either the model is stochastic, or it should be),
(v) solution uncertainty (as the system equations can only be solved to some necessary level of approximation).
(vi) structural uncertainty (the model only approximates the physical system),
(vii) measurement uncertainty (as the model is calibrated against system data all of which is measured with error),
(viii) multi-model uncertainty (usually we have not one but many models related to the physical system)
(ix) decision uncertainty (to use the model to influence real world outcomes, we need to relate things in the world that we can influence to inputs to the simulator and through outputs to actual impacts. These links are uncertain.)

General form of problem

We have a collection of observations z on the real world values y of a physical system. This is the system history.

We have a model $f(x)$ for the system.

[This is often implemented as a computer simulator.]

The model inputs are the parameter collection x

(plus other stuff like decision choices and forcing functions that we suppress to simplify notation)

The model outputs $f(x)$ are the assessment of the system history.

(plus other stuff which may be relevant and useful).

General issues

Scientific inferences about real world quantities raises three (at least!) general issues as compared to our treatment of statistical inference so far.

1. We know that our model is not correct. However, we still wish to use it to make statments about the real world.
2. There are no true values of the model parameters so that inferential methods which rely on there being a true but unknown value for the parameters have no logical foundation.
3. Typically the model input and output spaces are large and model evaluations are very time consuming.

   [This is a technical issue but it is hugely important in complicating our ability to address the above two issues.]

Questions of interest

Very often, modellers do not wish to know what are the true but unknown values of their model parameters.

Instead, they want to know if their models are capable of matching the real world outcomes that they have observed, by appropriate choices of parameters.

If so, they also wish to know the range of possible parameter choices which match system data, and the implications for future systems forecasts and related issues.

History matching

History matching is the problem of finding the collection $C(z)$ of **all** choices x^* for which $f(x^*)$ is “**near**” observed system history z.

Usually we wish to know whether it is possible to history match at all.

If $C(z)$ is empty, this typically identifies a problem with the model or the data.

In such cases, we aim to identify the conflicts which prevent a full history match.

More generally, the shape of the set $C(z)$ identifies the constraints on the parameter space that are imposed by the data.

History matching uses

Often, $f(x)$ will also contain outputs for unobservable quantities of interest. For example, these may be future system outcomes.

The collection of evaluations $f(x^*), x^* \in C(z)$ shows the range of model forecasts which are consistent with observed history.

[We must then turn these evaluations into actual real world system forecasts.]

If the model $f(.)$ takes inputs which may help control future outputs, then evaluation of effective control over the range of inputs in $C(z)$ identifies which are the safest control strategies and whether more data is needed before controls are introduced.

When is history matching difficult?

We want to “history match” system behaviour y by model behaviour $f(x)$.

Suppose y and x are high dimensional and/or $f(x)$ is a complex function which is expensive to evaluate for a single choice of x.

Then solving $y = f(x)$ for x is a complex, ill-posed inverse problem, which is generally difficult.

Finding all solutions for which $f(x)$ is “near” y, increases the complexity.

We don’t know y but only system observations z, which further adds to the difficulty.

And we still have to address the difference between the model and reality.

Function emulation

Uncertainty analysis, for high dimensional problems, is particularly challenging if $f(x)$ is expensive, in time and computational resources, to evaluate for any choice of x.

In such cases, f must be treated as uncertain for all input choices except the small subset for which an actual evaluation has been made.

Therefore, we must construct a description of uncertainty about the value of $f(x)$ for each x.

Such a representation is often termed an **emulator** of the model.

Function emulation

The emulator of a model both contains

(i) an approximation to the model and

(ii) an assessment of the likely magnitude of the error of the approximation.

Unlike the original model, the emulator is fast to evaluate for any choice of inputs.

This allows us to explore model behaviour for all physically meaningful input specifications.

Form of the emulator

We may represent beliefs about component f_i of f, using an emulator:

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

Global Variation

$\{\beta_{ij}\}$ are unknown scalars,

g_{ij} are known deterministic functions of x, (for example, polynomials)

Local Variation

$u_i(x)$ is a stationary stochastic process, with (for example) correlation function

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

(or a more complex version of this form).

[$u(x)$ might be second order stationary or a full Gaussian process.]

Emulation methods

We fit the emulators, given a collection of carefully chosen model evaluations, using our favourite statistical tools - generalised least squares, maximum likelihood, Bayes (linear) - supported by expert judgement.

For each output, $f_i(x)$ say, identify a collection of 'active' inputs, $x_{A(i)}$ say, which are most important in driving variation in that output.

Fit the emulator

$$f_i(x) = \sum_j \beta_{ij} g_{ij}(x_{A(i)}) + u_i(x)$$

and decompose the local residual $u_i(x)$ as the sum of one term involving $x_{A(i)}$, and one term involving all of the other inputs (possibly just a nugget).

It is often easier first to fit the global form and then fit the local form to the residuals.

Fitting the emulator

$$f_i(x) = \sum_j \beta_{ij} g_{ij}(x_{A(i)}) + u_i(x)$$

Assess the scientific plausibility of each emulator

(appropriate choice of active variables and form of global model)

Use careful diagnostics to test the validity of our emulators,

(for example, assessing the reliability of the emulator for predicting the model at new evaluations).

If the model is stochastic, then we might emulate the mean and the variance of each output,

Discussion

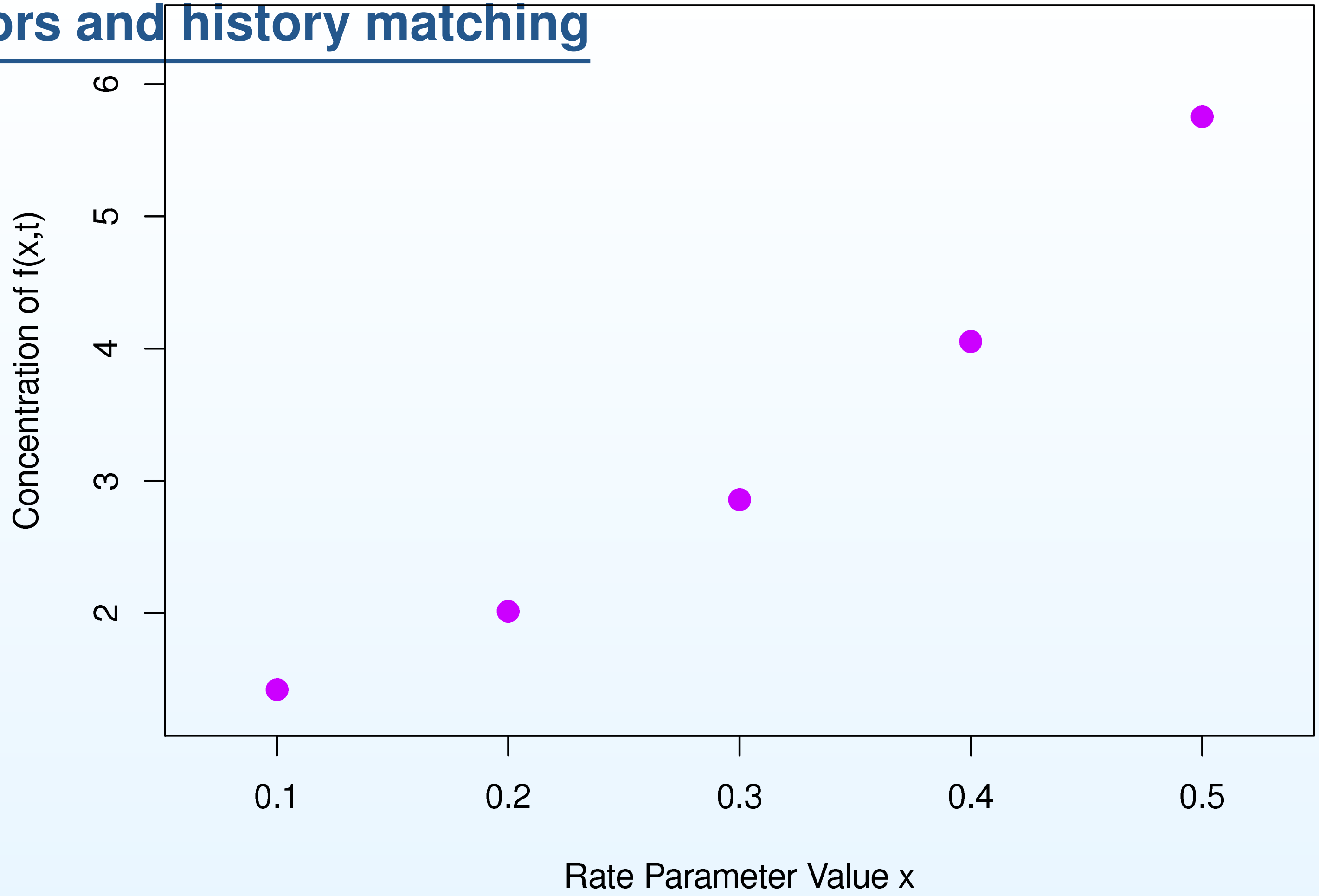
Emulation is an example of fitting a purely statistical model to a representation of a deterministic phenomenon.

In no sense is the model correct. Our criteria for a good emulator are

(i) is the emulator a good representation of our uncertainty about the function?

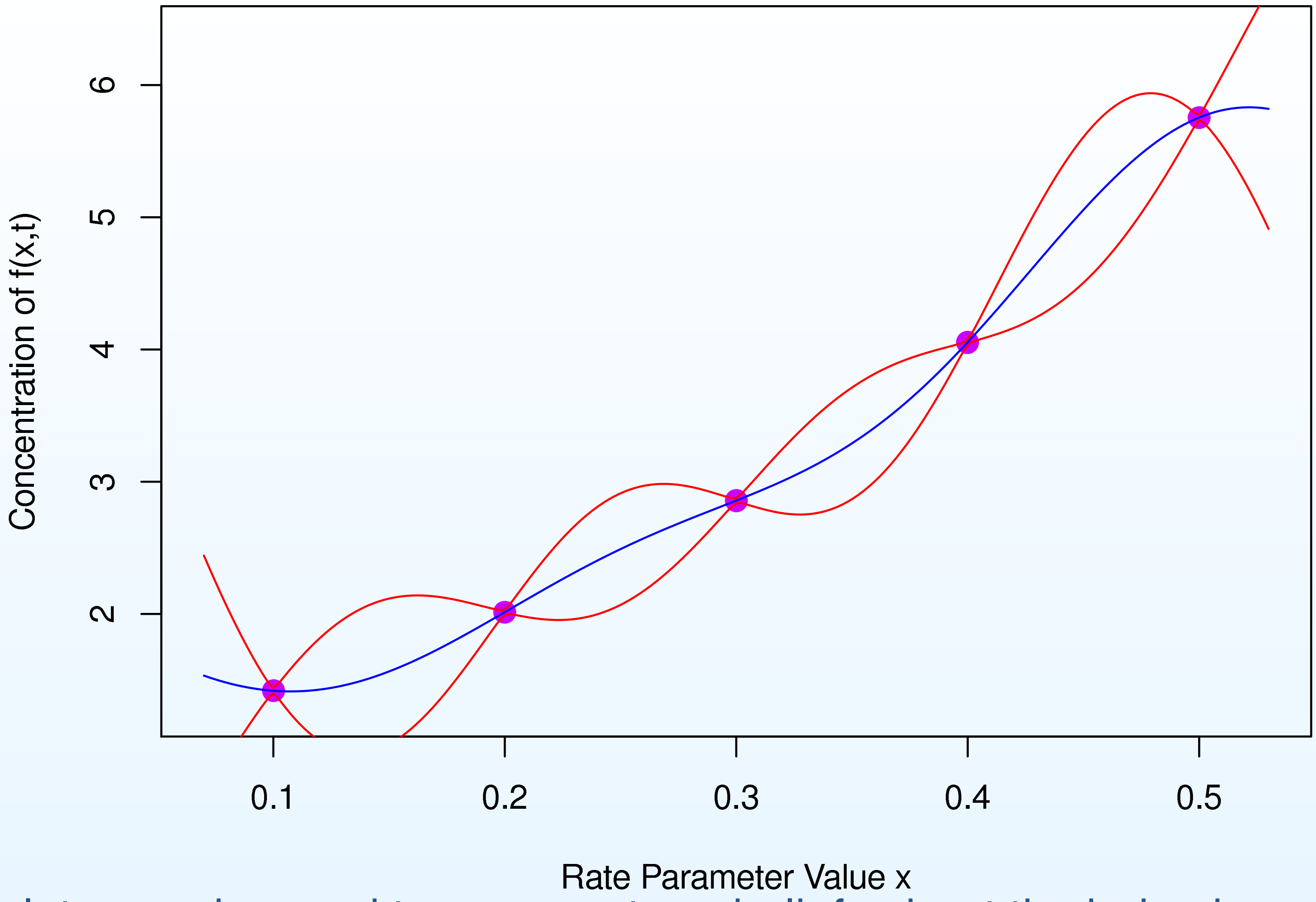
(ii) does the emulator pass careful diagnostic testing?

Emulators and history matching



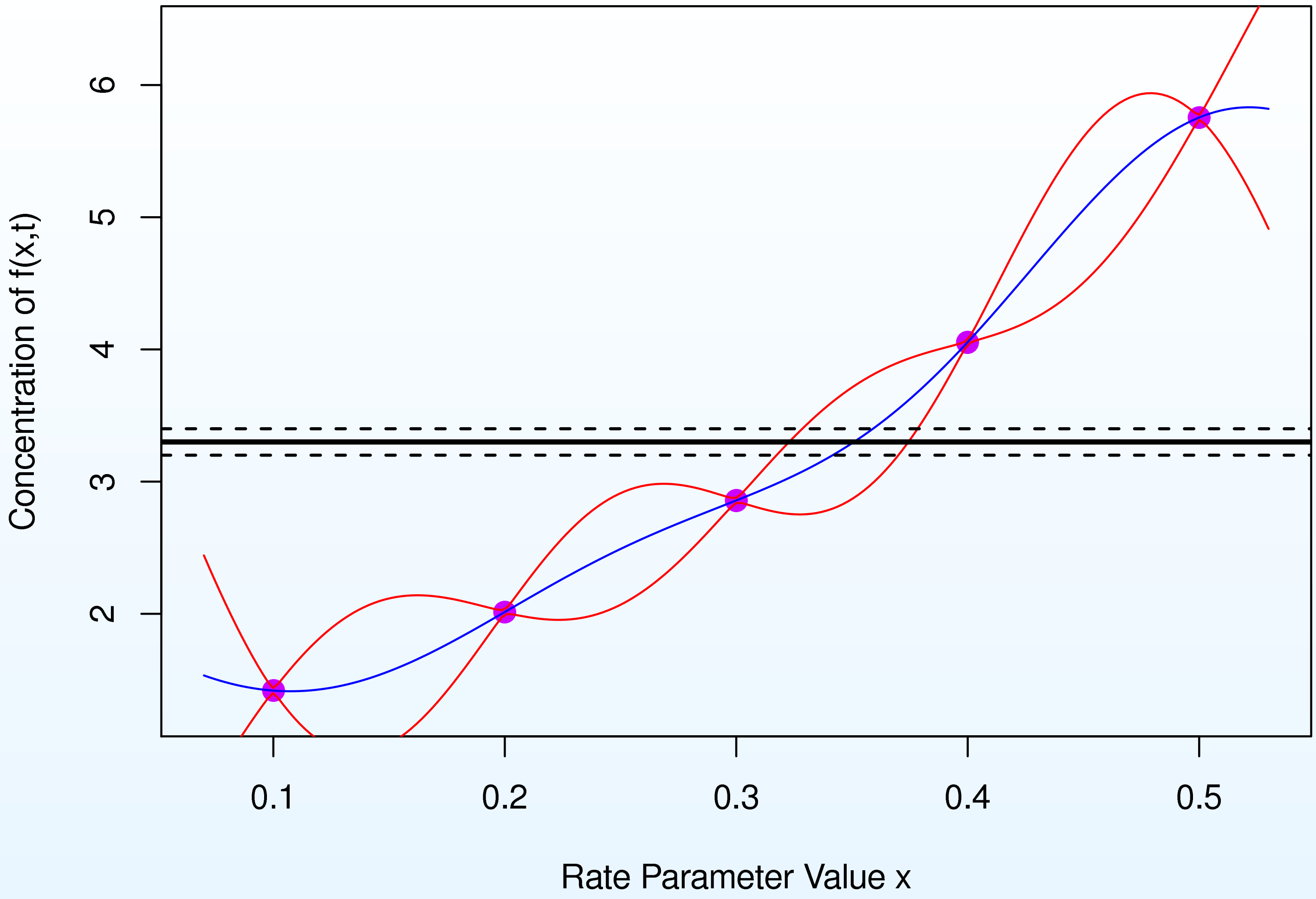
We have made five evaluations of the function $f(x)$.

Suppose that we now build an emulator for $f(x)$ based on these five points.



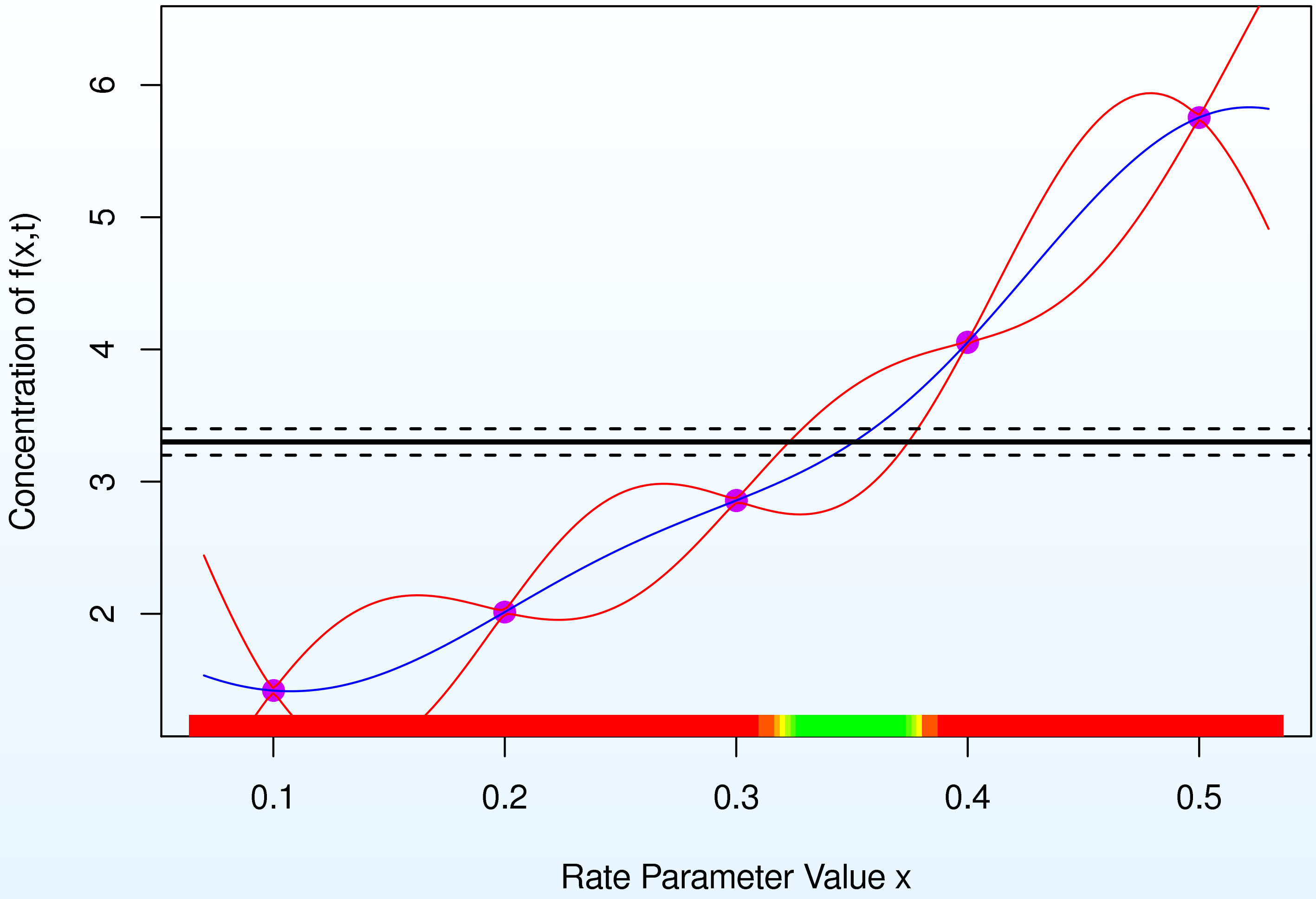
The emulator can be used to represent our beliefs about the behaviour of the model at untested values of x, and is fast to evaluate.

It gives both the expected value of $f(x)$ (the blue line) along with a credible interval for $f(x)$ (the red lines) representing uncertainty about the model's behaviour.

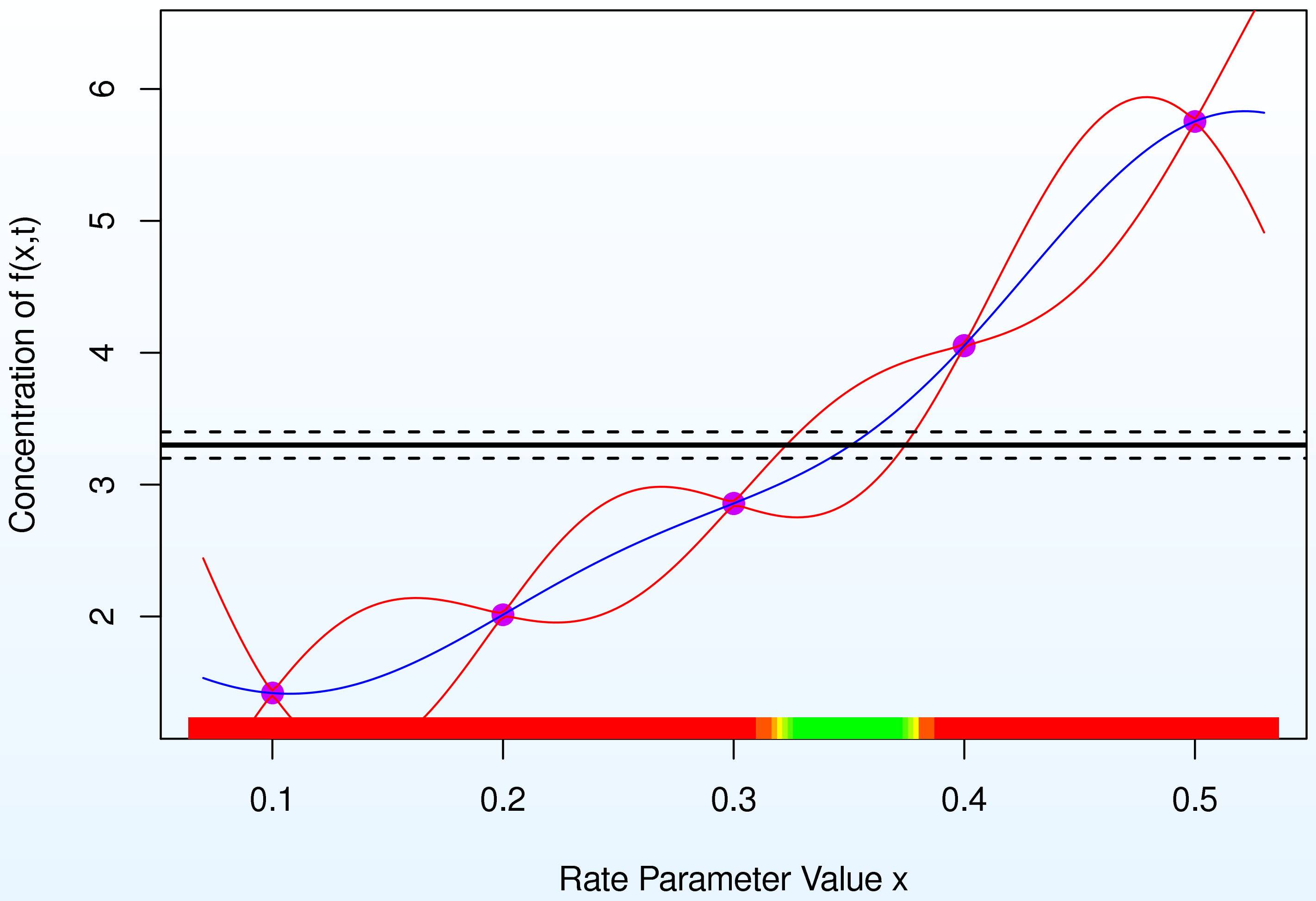


Suppose that we have an observation (represented by the black line with observational errors bounds).

Comparing the emulator to the observed measurement we can identify the set of x values which are “not inconsistent” with this data.

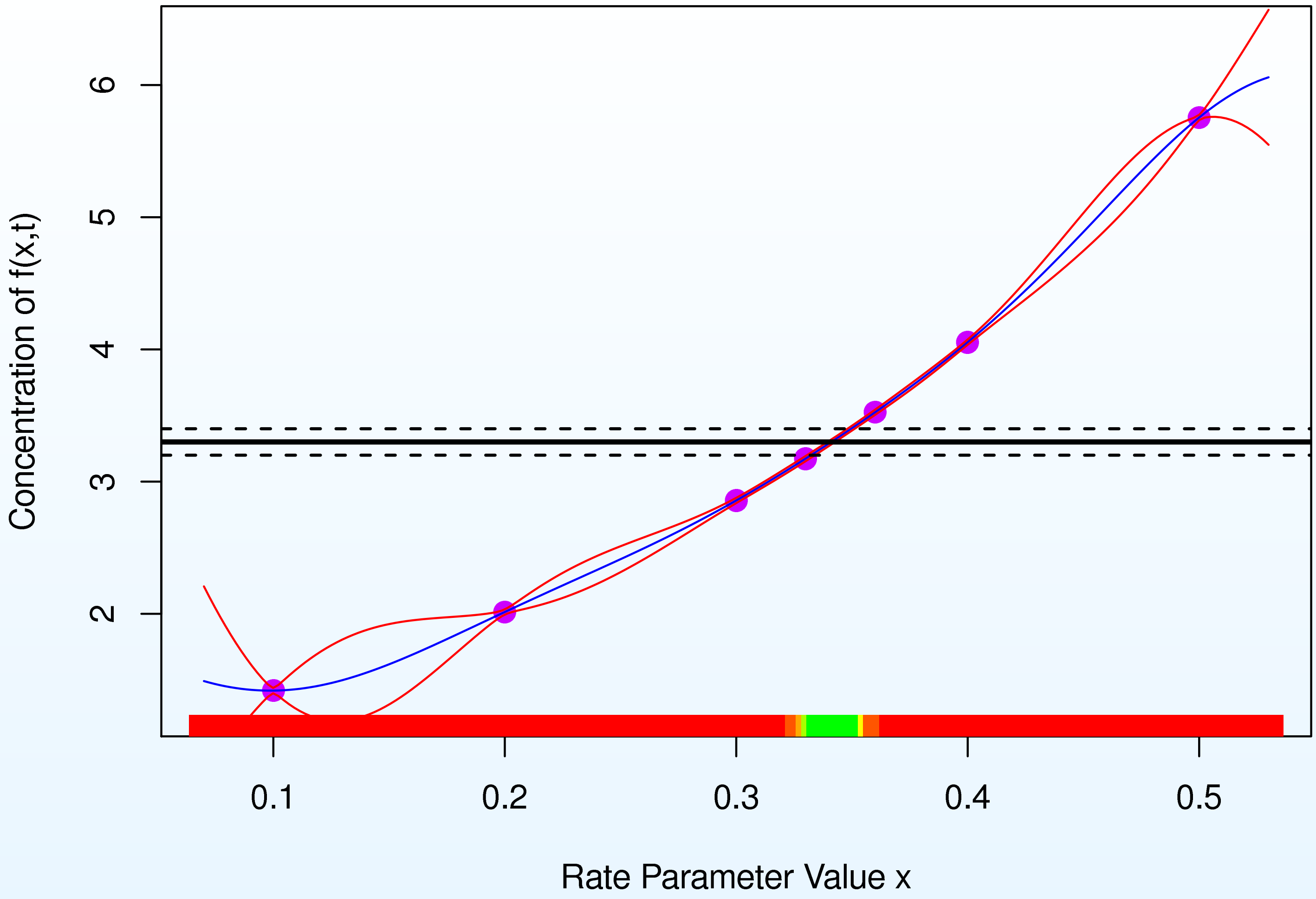


Comparing the emulator to the observed measurement we have identified the set of x values (the green values) which "match" the observed history, when we take into account all of the uncertainties (here, measurement and emulator error).



We now remove all of the implausible x values (the red values) and resample and re-emulate within the green region.

We perform a 2nd iteration or wave of runs to improve emulator accuracy. The runs are located only at non-implausible (green/yellow) points.



Now the emulator is more accurate than the observation, and we can identify the set of all x values of interest.

Note that this process is iterative and that we only need careful emulation in the region close to the output match.

Basic limitations of models

A model is a description of the way in which

system properties (the inputs to the model)

affect system behaviour (the output of the model).

This description involves two basic types of simplification.

(i) we approximate the properties of the system (as these properties are too complicated to describe fully and anyway we don’t know them)

(ii) we approximate the rules for finding system behaviour given system properties (because of necessary mathematical simplifications, and because we do not fully understand the relationships which govern the process).

[Statistical models face similar issues.]

Structural discrepancy

These approximations do not invalidate the modelling process.

Problems only arise when we forget these simplifications and confuse the analysis of the model with the corresponding analysis for the system itself.

Structural discrepancy is the difference between the model output at appropriately chosen input values and the real world system values the model purports to represent.

Structural discrepancy assessment should form a central part of the problem analysis.

Types of structural discrepancy

We may distinguish two types of model discrepancy.

(i) **Internal discrepancy**

Any aspect of discrepancy we can assess by direct experiments on the model.

(ii) **External discrepancy**

This arises from the inherent limitations of the modelling process.

Internal discrepancy

We may assess aspects of internal discrepancy by, for example

varying parameters/forcing functions held fixed in the standard analysis,

we may add random noise to the state vector which the model propagates,

we may allow parameters to vary over time/space.

We assess internal discrepancy by
(i) carrying out detailed experiments to determine discrepancy variance for certain input choices,

(ii) using emulation to extend the variance assessment over the input space.

M. Goldstein and N. Huntley (2017) Bayes linear emulation, history matching and forecasting for complex computer simulators, in The Handbook of Uncertainty Quantification, Ghanem, Higdon, Owhad (eds), Springer

External discrepancy

External discrepancy arises from the inherent limitations of the modelling process. It is determined by a combination of expert judgements and statistical estimation.

The simplest way to incorporate external discrepancy is to add an extra component of uncertainty to the model outputs. For example we may introduce, say, 10% additional error to account for structural discrepancy.

Better is to consider what we know about the limitations of the model, and build a probabilistic representation of additional features of the relationship between system properties and behaviour. We cannot evaluate the improved model, but we can emulate it.

Emulating external discrepancy

For example, if the emulator for the model is

$$f(x) = \sum_j \beta_j g_{ij}(x) + u(x)$$

then our emulator for the improved form might be

$$f^r(x) = \sum_j \beta_j^r g_{ij}(x) + u^r(x)$$

where the elements of the model form act as priors for the improved form.

M. Goldstein and J.C.Rougier (2009). Reified Bayesian modelling and inference for physical systems (with discussion), JSPI, 139, 1221-1239

The uncertainty representation

We want to judge whether input x produces output $f(x)$ which is near system value y. We don't observe y but see z.

We might judge that $z = y + e$ where e has zero mean and variance σ_e^2.

We don't judge the model to be a perfect representation of reality, so we introduce a structural discrepancy for example viewing the relation between y and $f(x)$ at an acceptable choice of x as $y = f(x) + \epsilon$ with variance σ_ϵ^2.

We don't know $f(x)$ for most values of x, so we use the emulator expectation $\mathrm{E}(f(x))$ with variance σ_f^2.

History matching by implausibility

We use an ‘implausibility measure’ $I(x)$ based on a probabilistic metric such as

$$I(x) = \frac{(z - \mathrm{E}(f(x)))^2}{\mathrm{Var}(z - \mathrm{E}(f(x)))} = \frac{(z - \mathrm{E}(f(x)))^2}{\sigma_e^2 + \sigma_f^2 + \sigma_\epsilon^2}$$

Large values of $I(x)$ suggest it is implausible that $f(x)$ is a good match to y

(for example, using Pukelsheim’s 3 sigma rule).
Inputs x with large $I(x)$ are unlikely to be appropriate choices.

The implausibility calculation can be performed over collections of outputs.

We can reject parts of the input space based on maximum implausibility or a vector version of implausibility based on Mahalobis distance.

Having identified a non-implausible region of the input space, we resample the reduced region, refit the emulators and repeat the analysis, continuing until we identify the region of acceptable matches.

Example: HIV modelling

This case study was based on a research project that explored HIV transmission in Uganda.

The model used, Mukwano, is a dynamic, stochastic, individual based computer model that simulates the life histories of hypothetical individuals (births, deaths, sexual partnership formation and dissolution and HIV transmission, modelled using time-dependent rates).

Each individual is represented by a number of characteristics, such as gender, date of birth, HIV status, level of sexual activity, concurrency level.

The behavioural inputs take different values in each of three calendar time periods. This enables sexual behaviour to vary over time.

Twenty behavioural and two epidemiologic inputs were varied for this study.

Example: empirical data

The empirical data were collected from a rural general population cohort in South-West Uganda. The cohort was established in 1989 and currently consists of the residents of 25 villages.

Every year, demographic information on the cohort is updated, the population is tested for HIV, and a behavioural questionnaire is conducted.

In this study, there are 18 simulator outputs with calibration targets and limits for what constitutes an acceptable match.

These include male and female population sizes,

male and female HIV prevalences at three time points.

outputs that check that the behavioural features of the model matched the empirical data.

The run time for a single simulation for the study varies between 10 minutes and 3 hours.

Example: references

Full details of example are in the paper:

Ioannis Andrianakis , Ian R. Vernon, Nicky McCreesh, Trevelyan J. McKinley, Jeremy E. Oakley, Rebecca N. Nsubuga, Michael Goldstein, Richard G. White

(2015) Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda,
PLOS Computational Biology.

More careful and detailed treatment in

Ioannis Andrianakis , Ian R. Vernon, Nicky McCreesh, Trevelyan J. McKinley, Jeremy E. Oakley, Rebecca N. Nsubuga, Michael Goldstein, Richard G. White

(2017) Efficient history matching of a high dimensional individual based HIV transmission model"
in SIAM/ASA Journal on Uncertainty Quantification.

which applies a development of the same ideas to a much larger version of the model (96 inputs, 50 outputs).

Example: HIV emulation and history matching

Emulation 220 point maximin Latin hypercube (100 replications at each point to average out stochasticity) 20 point Latin hypercube for validation set.

Logit transforms for outputs in [0,1].

Global fit used a cubic polynomial in the inputs.

All emulators validated in first wave except for 2 (HIV prevalence, male and female, 2007) which were emulated in later waves.

History matching 5.5×10^8 points were drawn from a 22 dimensional uniform distribution in [0,1] and the implausibility was evaluated for each one of them.

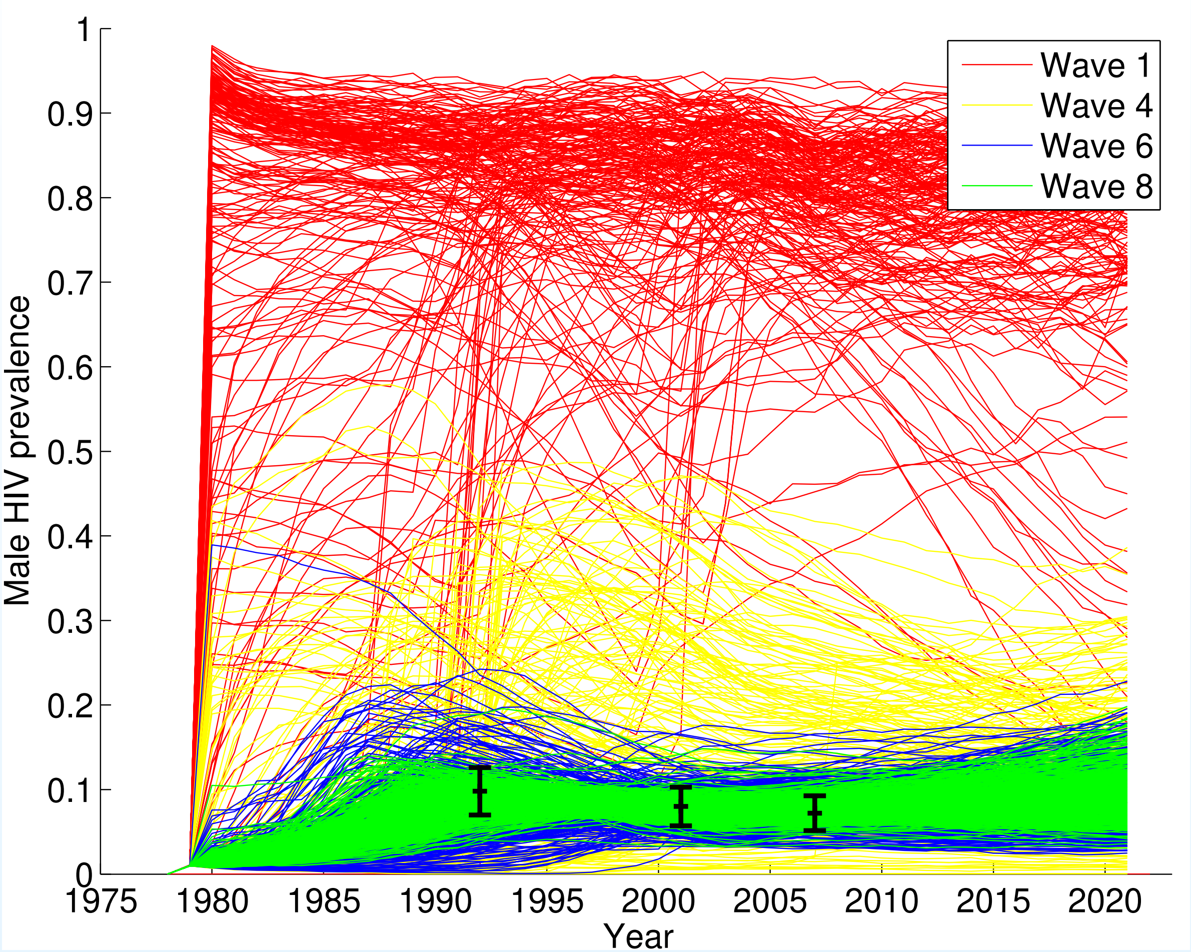
Initial implausibility criterion was the maximum implausibility

Evaluation was done on a 240 node cluster, completed in less than 5 minutes.

Only 21644 passed the implausibility test, implying that the volume of non-implausible space at this wave is of size 4×10^{-5} of the original input space.

This process was repeated through 10 waves.

History matching for the case study



After 10 waves, we have reduced the space to about 10^{-11} of original space. Around 65% of the simulator evaluations in the final space give runs with acceptable matches to the historical data.

The HMER package

HMER (history matching and emulation in R) is a system developed under a collaboration, with Wellcome Trust funding, between

Andrew Iskauskas, Michael Goldstein, Ian Vernon (Durham)

Nicky McCreesh, Danny Scarponi, Richard White (LSHTM)

TJ McKinley (Exeter)

building on a previous collaboration funded by MRC.

The package is customised for epidemic models, but the underlying methodology is fully general.

The HMER package

The package is available from CRAN.

You can find detailed documentation at

https://github.com/andy-iskauskas/hmer

The project web-page, which has lots of support material is at

https://hmer-package.github.io/website/

The programme has been extensively tested.

For example history matching a complex deterministic model for the country-level implementation of tuberculosis vaccines to 114 countries, fitting to 9–13 target measures, by varying 19–22 input parameters.

105 countries were successfully matched (i.e. producing many parameter choices which match history)

The remaining 9 countries revealed evidence of model or data misspecification.