Notes on statistical modelling for complex systems

Jonty Rougier

Department of Mathematics
University of Bristol

Ver. 0.5, compiled August 17, 2009

Abstract

These notes are an outline of the concepts, terms, and notation used in the statistical treatment of model-based inference for complex systems. They are a synthesis of a wide and rapidly-developing literature, and only the simplest situations are covered. Having said that, the 'all-gaussian' approach that is illustrated here requires only minor generalisations to be state-of-the-art in current statistical practice. References are not given, but there is Further Reading at the end.

Please reference these notes as: J.C. Rougier, 2009, Notes on statistical modelling for complex systems, ver. 0.5, unpublished.

KEYWORDS: SIMULATOR, PARAMETRIC AND STRUCTURAL UNCERTAINTY, 'BEST INPUT' APPROACH, DISCREPANCY, 'ALL GAUSSIAN' CASE, CALIBRATION, MODEL CRITICISM, CALIBRATED PREDICTION, BIAS CORRECTION, EMULATION

1 Two types of uncertainty

The process of modelling a complex system involves abstraction, which has the effect of introducing ambiguity in the precise meaning of model parameters and model outputs, even when these things are clearly defined in the underlying system. The ambiguity is increased if the model has to be further simplified in order to be solved, as is the case for models that start out as differential equations. While this may or may not affect the qualitative features of the model, it definitely affects the way in which the outcome of model evaluations is representative of the system. If model evaluations are used to make inferences about the system, then this ambiguity is represented as uncertainty about the system, given a set of evaluations of the model. Since the word 'model' is heavily overloaded, from now on I'll refer to the computer code that is evaluated as the 'simulator'. Furthermore, it is helpful to restrict 'parameters' to be those quantities that are common to many different simulators for the same system. Typically these would be the coefficients in the underlying mathematical equations, and would exclude initial and boundary values, and forcing.

A simple but powerful statistical framework is used to represent the relationship between the simulator and the system. Formally, we identify two sets of uncertain quantities. First, there are the system values, which we represent as Y. Y could be a huge collection of quantities; it is represented as a majuscle letter following the general statistical convention that uncertain quantities are written as majuscule letters, while particular instances are written as minuscule letters. The second set of uncertain quantities are the 'correct' values of the simulator parameters, which we represent as θ^*. In statistics, parameters are often represented using θ, but we need to distinguish here between an arbitrary parameter value and the 'correct' parameter value, hence the star in the latter case. The issue of whether there is such a thing as a 'correct' parameter value in an imperfect simulator is subtle, but there is no doubt that almost all of scientific practice using simulators proceeds on this basis (from now on I'll drop the scare quotes around 'correct').

Now we have two sets of uncertain quantities, that must be probabilistically related, if we think that evaluations of the simulator are informative about the system. To represent this relationship we would specify the joint *probability distribution function*

$$F_{Y,\theta^*}(y,\theta) \triangleq \Pr(Y \leq y, \theta^* \leq \theta)$$

where '$\triangleq$ denotes 'defined as'. This notation for the distribution function is unambiguous but also a bit clumsy. Usually, statisticians would write simply $F(y, \theta)$. The distribution function is the primitive concept in probability, in the sense that every random quantity possess a distribution function, and every function satisfying the basic properties of a distribution function corresponds to a random quantity.[1] If Y and θ^* are both 'absolutely continuous' random quantities, then the distribution function can be differentiated to give the *probability density function (PDF)* $f_{Y,\theta^*}(y, \theta)$, shortened to $f(y, \theta)$, with the property that, for any set $\mathcal{A}$ in the domain of Y and θ^*

$$\Pr\big((Y, \theta^*) \in \mathcal{A}\big) = \int_{\mathcal{A}} f(y, \theta)\, \mathrm{d}(y, \theta).$$

We will treat Y and θ^* as absolutely continuous from now on, for simplicity.

The PDF of Y and θ^* can be factorised, using the notion of *conditional probability*. If A and B are two uncertain propositions then the conditional probability of A given B, written $\Pr(A \mid B)$, is defined as

$$\Pr(A \mid B) \triangleq \frac{\Pr(A, B)}{\Pr(B)}.$$

Interpretively, $\Pr(A \mid B)$ can be thought of as 'the probability that A is true supposing that B were true'.[2] It is a hypothetical, because we do not actually know whether B is true or not. According to the definition of conditional probability, we can write

$$f(y, \theta) = f(y \mid \theta)\, f(\theta), \tag{1}$$

where the final term is the shortened form of $f_{\theta^*}(\theta)$. In words, the PDF of $(Y = y, \theta^* = \theta)$ can be expressed as the product of the PDF of $Y = y$ supposing that $\theta^* = \theta$, and the PDF of $\theta^* = \theta$.

[1] If X is a random quantity, then the properties of $F_X(x)$ are (i) $F_X(-\infty) = 0$, (ii) $F_X(\infty) = 1$, (iii) F_X is non-decreasing, and (iv) F_X is right-continuous. This final condition is slightly technical, but not important for us.

[2] Note that this is an interpretation which serves well in practice. But, fundamentally, the conditional probability is nothing more or less than its definition.

The two PDFs on the righthand side represent two different sources of uncertainty. The marginal PDF $f(\theta)$ represents our uncertainty about the correct value of the simulator parameters. We refer to this as *parametric uncertainty*. The conditional PDF $f(y \mid \theta)$ represents our uncertainty about the system, supposing that we knew (somehow) that the correct value of the simulator parameters happened to be θ. We refer to this as *structural uncertainty*.

Note that the evaluations of the simulator are tucked away inside the structural uncertainty; this will become clear in section 2.

2 The special case of a deterministic simulator

Most simulators of complex systems are currently *deterministic*, which is to say that they return an identical value every time they are evaluated at the same parameter value. I hope that this will change because I believe we can get much richer representations of structural error through using *stochastic* simulators, especially where the simulator is dynamical. However, in these notes I stick with the predominant case (with a slight digression at the end of this section).

We write the simulator as $g : \theta \mapsto y$, read as 'g maps points in the space containing θ to points in the space containing y'. In fact, g does not have to map θ into the same space as y. Typically, though, simulators are built to understand Y, and therefore the value $g(\theta)$ will 'look like' y.

Now we can construct a hierarchy of simulator quality, starting with the best. First, there is the *perfect simulator*. In this case there is no structural uncertainty, so that

$$f(y \mid \theta) \propto \delta\big(y - g(\theta)\big) \tag{2}$$

where δ is the Dirac delta function, which is 1 when the argument equals zero, and zero otherwise. In other words, if we knew the correct parameter, there would be no further uncertainty, i.e. $Y = g(\theta^*)$. Any analysis that chooses to ignore structural uncertainty is effectively assuming a perfect simulator—this is a gross mistake for complex systems. It would be compounded by also assuming that there is no parametric uncertainty, which would be to choose

$$f(\theta) \propto \delta(\theta - \theta_0) \tag{3}$$

for some specified value θ_0.

Second, there is the *best input* approach, in which we choose to treat $g(\theta^*)$ as representing all the information about Y that there is in θ^*. This is typically a restriction on (1) because g may well be a many-to-one function, in which case $g(\theta^*)$ would contain less information than θ^*. The *gaussian* instantiation of the structural error in the best input approach is

$$f\big(y \mid \theta\big) = \phi\big(y; g(\theta), \Sigma\big) \tag{4}$$

where ϕ is the multivariate gaussian (or 'normal') PDF with specified mean $g(\theta)$ and variance Σ. This is one example from a class of representations of structural uncertainty in which $Y - g(\theta^*)$ is probabilistically independent of θ^*. The difference $Y - g(\theta^*)$ is termed the simulator *discrepancy*, and Σ is the *discrepancy variance*. The gaussian best input approach is currently the 'state of the art' in quantifying structural error, and it already involves the considerable challenge of specifying the discrepancy variance. But ignoring Σ by setting it to zero is a return to the perfect simulator.

Note that the use of a gaussian PDF in (4) may necessitate a transformation of some of the components of y and $g(\theta)$; e.g. logs of strictly positive quantities, logits for proportions and so on. Some care has to be taken to ensure that the variance matrix Σ, which applies on the transformed scale, is an appropriate measure of structural uncertainty on the original scale.

At the next step of generalisation, we retain the simple gaussian distribution form of structural uncertainty, but allow Σ to be affected by the value of θ^*. Thus instead of specifying a discrepancy variance matrix, we specify a discrepancy variance *function*, $\Sigma(\theta)$, and write

$$f\big(y \mid \theta\big) = \phi\big(y; g(\theta), \Sigma(\theta)\big) \tag{5}$$

This generalisation provides one way of treating stochastic simulators within a deterministic framework. It also shows us how we might derive a plausible value for Σ in situations where the system itself tightly constrains the acceptable values of Y. Suppose that our simulator is represented as

$$h(\theta, \omega) \tag{6}$$

where ω is a random vector that we can take to comprise independent standard gaussian quantities, without loss of generality. This is a stochastic simulator, because each evaluation at the same θ gets a different realisation for ω. We can linearise this simulator around $E(\omega)$, which is zero, to give

$$h(\theta, \omega) \approx h(\theta, \mathbf{0}) + \nabla_\omega h(\theta, \mathbf{0})\, \omega \tag{7}$$

where $\nabla_\omega h$ is the Jacobian matrix of first derivatives with respect to ω. If we are happy to assert that $Y = h(\theta^*, \omega)$, i.e. that uncertainty about ω captures our structural uncertainty, and happy with the first-order approximation, then the deterministic simulator is $g(\theta) = h(\theta, \mathbf{0})$, and the variance function is $\Sigma(\theta) = \{\nabla_\omega h(\theta, \mathbf{0})\}\{\nabla_\omega h(\theta, \mathbf{0})\}^T$.

Overall, though, a good entry-level representation of the system uncertainty that follows from model limitations is given by the gaussian best input approach, which requires, in addition to the simulator itself, a specification of the discrepancy variance Σ.

3 Simulator calibration

Simulator calibration involves learning about θ^* on the basis of observations about the system. This presupposes a functional relationship between the value θ^* and the distribution of the observations, which we denote Z when they are thought of as uncertain (e.g. before being made), and z^{obs} when they take the specific observed values. Calibration is simplified by making the plausible choice that the system value Y is statistically sufficient for Z, so that the joint distribution of Z and Y given θ^* factorises as

$$f(z, y \mid \theta) = f(z \mid y, \theta) f(y \mid \theta) = f(z \mid y) f(y \mid \theta), \tag{8}$$

where the first equality is always true and the second embodies the sufficiency property of Y for Z.[3]

The situation becomes simpler still if we adopt the gaussian best input approach, and also choose to model the conditional distribution of Z given Y as a multivariate gaussian distribution

$$f(z \mid y) = \phi(z; Hy, T) \tag{9}$$

where H is a specified matrix that represents a linear mapping from y to z (which can account for missing observations, or observations which are the means of subsets of system values), and T is a specified variance matrix which represents measurement error. I will refer to the combination of gaussian best input approach, sufficiency of Y for Z, and gaussian distribution for $Z \mid Y$ as the *all-gaussian* case. In the all-gaussian case we can integrate Y out of $f(z, y \mid \theta)$, and so find an explicit form for the *likelihood function* of θ^*,

$$L(\theta) \triangleq f_{Z|\theta^*}(z^{\text{obs}} \mid \theta) = \phi\big(z; Hg(\theta), H\Sigma H^{T} + T\big), \tag{10}$$

where the definition is general, and the equality shows the form of the likelihood function in the all-gaussian case.

In general, the likelihood function computes the probability density of the observations, conditional on the correct value of the simulator parameters. It is effectively a scoring function in which better-fitting values of the parameters get higher scores. The point about any such scoring function is that it needs to account for the uncertainty that exists between the simulator outputs and the observed system values.

In the all-gaussian likelihood function in (10) we treat the discrepancy variance Σ as specified (likewise T, but this is usually less of a problem). It *is* possible to treat Σ as unknown, and to be learnt about along with θ^*, but this is statistically demanding, unless there are already strong judgements about Σ. I tend to take a fairly hard line here. If all you care about is your simulator, then you are not a scientist but an artist. Scientists care about

[3] One should note here that this is another notational abuse. Statisticians write $f(z \mid y)$ to show that $f(z \mid y, \theta) = f(z \mid y, \theta')$ for all θ and θ'.

what their simulators can tell us about the system, and they are obliged to think about how inaccurate their simulators might be. In the gaussian best input approach, this thinking in quantified in Σ: the diagonal elements represent the accuracy of the simulator, while the off-diagonal elements represent the degree to which the simulator will be systematically wrong, over sets of components of the system vector. It is not straightforward to turn one's thinking about simulator inaccuracy into a specification of Σ, and this is an area where consulting a statistician will definitely help. But even a crude assessment of Σ is going to be better than setting Σ equal to zero.

A note on computation. The likelihood in the all-gaussian case needs to be computed carefully. First, find the Choleski decomposition of the variance, namely the unique upper-triangular matrix Q for which $Q^T Q = H \Sigma H^T + T$. This calculation only has to be done once. Then, for each θ, find the vector

$$w(\theta) \triangleq Q^{-T} \big(z^{\text{obs}} - H g(\theta) \big) \tag{11}$$

by back-substitution. Finally, compute the log likelihood

$$\ell(\theta) \triangleq \log L(\theta) = \Big(\sum_{i=1}^{n} w_i(\theta)^2 + n \log \pi + 2 \sum_{i=1}^{n} \log q_{ii} \Big) / (-2) \tag{12}$$

where n is the number of observations in z^{obs}, $w_i(\theta)$ is the ith component of $w(\theta)$, and q_{ii} is the ith term on the diagonal of Q. The only term that involves θ is the one in $w_i(\theta)$, and so the log-likelihood can be simplified to

$$\ell(\theta) = c - \sum_{i=1}^{n} w_i(\theta)^2 / 2 \tag{13}$$

where c can be neglected.

This representation of the log-likelihood allows us to interpret a common choice for the cost function when finding a fitted value for θ^*, in the case where all the simulator outputs are of the same type. Very often the analyst will minimise the sum of the squared distances between the simulator output and the observations, for which H would be the identity matrix. This approach would be implied by $\Sigma + T$ being a diagonal matrix, because then Q would be diagonal, and $w(\theta) \propto z^{\text{obs}} - g(\theta)$. Effectively, both Σ and T would have to be diagonal. It is quite plausible that T is diagonal, if all observations are measured with the same type of (unbiased) instrument. It is also fairly plausible that all of the diagonal elements of Σ would be the same, if the simulator quality is judged not to depend on the indexing of the output vector (e.g. not to depend on the location or time index). But setting all the off-diagonal elements of Σ to zero represents a judgement that the simulator is never systematically wrong over a subset of output components. This is highly implausible for simulators of complex systems, where simulator errors are expected to persist in similar outputs. What then happens is that the observations appear more numerous than they actually are, and the likelihood function is too peaked.

3.1 Maximum likelihood (ML)

The *Frequentist* approach to inference avoids making probabilistic statements about θ^*, preferring to locate *all* of the uncertainty in the sampling behaviour of Z. Therefore there is no $f(\theta)$ beyond basic boundaries for θ^*; instead a less structured approach is taken in which θ^* is simply 'unknown'. This might be very attractive in situations where it is hard to specify $f(\theta)$, but it has its own complications. Inferences about θ^* using z^{obs} are judged on their hypothetical properties under many independent realisations of Z, supposing that the statistical model underlying the likelihood function is true. This means that we cannot make probabilistic assertions about θ^*, but we *can* make probabilistic assertions about the sampling behaviour of estimates of θ^* based on Z. These assertions are typically in the form of *confidence sets* (see below).

Point estimation. Sometimes all one wants is a point estimate for θ^*. For example, one wants a better estimate of θ^* to plug into the simulator than that arrived at purely by introspection.

One approach to finding a point estimate for θ^* is to maximise the log-likelihood function, giving

$$\hat{\theta}_{\text{ML}} = \operatorname*{argmax}_{\theta} \ell(\theta). \tag{14}$$

Note that this type of estimate presupposes that the simulator g is quite cheap to evaluate, because the process of numerically maximising ℓ requires many evaluations at different candidate values of θ. When the simulator is expensive to evaluate, we must use an emulator, discussed in section 5.

Model criticism. The value $\hat{\theta}_{\text{ML}}$ may or may not be a good estimate of θ^*: just because it gives a maximum does not mean that the resulting simulator evaluation is a good fit to the observations. 'Model criticism' assesses this, but it is always hampered in practice by the fact that we do not know θ^*. If $\hat{\theta}_{\text{ML}}$ is to be used as a plug-in value for θ^*, though, it is reasonable to expect that the statistical model at $\hat{\theta}_{\text{ML}}$ will do a good job of matching the observations z^{obs}. Note that this is *model* criticism: what we are critiquing is the statistical model for Z, for which the simulator is only a part, albeit an important one. It is nonsense to 'validate' a simulator, since the simulator must be wrong, and will be discovered to be wrong if the observations are sufficiently informative. If we want to assess the simulator's performance using system observations, we must make an explicit allowance for the simulator's limitations.

Suppose, for illustration, that we are in the all-gaussian case. If our statistical model is a good one, and $\hat{\theta}_{\text{ML}}$ is a good estimate of θ^*, then we would expect the components of $w(\hat{\theta}_{\text{ML}})$ from (11) to be approximately independent standard gaussian random quantities. Therefore one diagnostic for model criticism is a histogram of these quantities, or a QQ plot. Probably this will look bad; it is, after all, a difficult job to construct a simulator for a complex system, and then to quantify its structural uncertainty in the form of a variance matrix. The individual standardised components

$$\frac{z_i^{\text{obs}} - [Hg(\hat{\theta}_{\text{ML}})]_i}{\sqrt{[H\Sigma H^{T} + T]_{ii}}} \tag{15}$$

are not independent, but they can show where the mis-fitting is worst. Then it is probably a good idea to adjust Σ accordingly. This type of informal revision of judgements alarms the purists, who worry about double-counting, but to me it seems like a perfectly reasonable step, and also a simple approximation to more complicated inferences in which Σ is also learnt from z^{obs}.

This revision of Σ will affect the maximum likelihood value. So a refit of the $\hat{\theta}_{\text{ML}}$ might also be a good idea, followed by another round of model criticism. This process could be iterated, but one iteration is probably about as much double-counting as the observations can stand.

Parameter uncertainty. The justification of ML as a good way of estimating an uncertain parameter relies mainly on asymptotic results that hold under certain conditions.[4] It is not clear that these conditions will hold in a simulator of a complex system, and nor is it obvious that asymptotic results apply. Unfortunately, these same conditions and results are also the basis for estimating the imprecision of $\hat{\theta}_{\text{ML}}$, in the form of confidence sets for θ^*. Taking the optimistic view, though, the level set

$$\Big\{\theta : \ell(\theta) \geq \ell(\hat{\theta}_{\text{ML}}) - k_p(1-\alpha)/2\Big\} \tag{16}$$

describes an asymptotic $100(1-\alpha)\%$ confidence set for θ^*, where $k_p(1-\alpha)$ is the $1-\alpha$ quantile of the χ^2_p distribution, and p is the number of components in θ.[5]

There is an equivalence between confidence sets and hypothesis tests: the 95% confidence set for θ^* is also the largest set of null hypotheses that are not rejected by the observations z^{obs}. Thus this type of confidence set provides a simple way to test a specific hypothesis about θ^* against a two-tailed alternative. By fiddling with α it is also possible to find the p-value

[4] Most importantly, g must be a smooth function of θ, and θ must be a continuous parameter defined on an open set.

[5] **Confidence sets.** A $100(1-\alpha)\%$ confidence set has the property that, when many such sets are considered (over many different experiments), at least $100(1-\alpha)\%$ of them will contain the correct parameter value, no matter what the correct value might be, providing that the statistical model is correct in every case. Typically we take $\alpha = 5\%$ and compute the 95% confidence set.

of this test. A very small p-value would be suggestive, but, bearing in mind that there are conditions underlying the confidence set that maybe satisfied only approximately for g and for the amount of information in z^{obs}, it would be a mistake to over-interpret hypothesis tests (or, indeed, the confidence sets).

The level set in (16) is a not-necessarily-convex subset of p-dimensional parameter space, and so it is impossible to visualise beyond $p = 2$ or 3. Therefore we need a method for projecting this into a lower-dimensional subset of the components of θ^*. Caution is needed here, because there are likely to be strong dependencies among the components of θ^* if the effect of one component can be somewhat offset by others; this shows up as a diagonal ridge in the likelihood function, and a similar feature in the confidence set. With that *caveat* in mind, lower-dimensional confidence sets can be found by the method of *profile likelihood.* Suppose we wanted to consider just the first component of θ^*. The profile log-likelihood function is

$$\ell(\theta_1) \triangleq \sup_{\theta_2,\dots,\theta_p} \ell(\theta) \tag{17}$$

i.e. for each value θ_1 we find the maximum value of the log-likelihood in the other $p-1$ components.[6] Asymptotic confidence sets can then be constructed using the same approach as in (16), i.e.

$$\Big\{\theta_1 : \ell(\theta_1) \geq \ell(\hat{\theta}_{\text{ML}}) - k_1(1-\alpha)/2\Big\} \tag{18}$$

describes an asymptotic $100(1-\alpha)\%$ confidence set for the first component of θ^*. In the case $\alpha = 5\%$, we have $k_1(0.95)/2 \approx 2$, which makes for a useful rule-of-thumb.

3.2 The Bayesian approach

In the Frequentist approach, probability is taken to be limiting frequency, and the statistical model that describes the probability distribution of Z

[6] Or, to put it another way, we project by maximising over the unwanted components. By construction, $\sup_{\theta_1} \ell(\theta_1) = \ell(\hat{\theta}_{\text{ML}})$.

for a specified value of θ^* is a description of what might happen if we were to sample Z in many worlds quite similar to this one. If this seems quite convoluted, it is because complex systems do not generate the type of independent and identically distributed observations under which the Frequentist approach is more compelling. In the Bayesian approach, probability is a description of uncertainty, and does not need to be underpinned by the notion of frequency. Therefore the likelihood function becomes a description of our uncertainty about Z supposing $\theta^* = \theta$, and we can also describe our parametric uncertainty by $f(\theta)$.

Bayes's Theorem allows us to compute the posterior distribution $f(\theta | z^{\text{obs}})$ in a mechanical way from the prior and the likelihood,

$$f(\theta \mid z^{\text{obs}}) \propto q(\theta) \triangleq L(\theta) \times f(\theta). \tag{19}$$

Generic methods for Bayesian updating are now mainly sampling-based (so-called Monte Carlo methods), and are effective for posterior distributions known up to a normalising constant (i.e. only the function q in (19) is required). Therefore the result of a Bayesian calibration is usually not a posterior PDF for θ^*, but a sample from the posterior distribution.

Point estimation. If a point-estimate is required, then the posterior mode is a good choice,

$$\hat{\theta}_{\text{B}} \triangleq \operatorname*{argmax}_{\theta} f(\theta \mid z^{\text{obs}}) \equiv \operatorname*{argmax}_{\theta} q(\theta), \tag{20}$$

where '$\equiv$' denotes 'equivalent to'. This can be estimated from the posterior sample as long as the value of $q(\theta)$ is kept for each member of the sample. If the prior is rectangular (independent components and marginally uniform) then the posterior mode is equal to the ML estimate, otherwise it takes into account the additional information present in the prior. In other statistical situations the posterior *mean* is advocated, but this can be a risky choice when the likelihood function has an irregular form, as is often the case for complex systems. There is always a risk that the posterior mean might correspond to a parameter value with low posterior probability density—this is not what one wants for a plug-in.

Model criticism. For model criticism, the Bayesian plug-in $\hat{\theta}_{\mathrm{B}}$ can be assessed in the same way as the ML plug-in, $\hat{\theta}_{\mathrm{ML}}$. We expect it to perform worse in this assessment, because it is less sensitive to the data than $\hat{\theta}_{\mathrm{ML}}$, but it should not do too much worse, unless there is strong information in the prior that conflicts with the observations. Typically this would involve probabilistic information about θ^*, e.g. information from previous studies that some component of θ^* is more likely to be small than large. In this case, a large divergence between the two estimates and/or a large difference in the model criticism diagnostics may reflect that z^{obs} is an unusual realisation of Z (so we would not want to fit too closely to z^{obs}). As a general rule it is always a good idea to see how the prior and the likelihood contribute to the posterior. The situation to avoid is where they do not overlap very much, in which case the posterior is determined by tail behaviour in L and f_{θ^*} that we are less confident about.

Parameter uncertainty. This can be assessed directly from the posterior sample, using standard descriptive techniques. For example, the posterior probability that θ^* falls into the region $\mathcal{T}$ is estimated as the proportion of the sample that fall into $\mathcal{T}$. Note that this is a probability statement directly about θ^*; such things are not meaningful in the Frequentist approach.

4 Calibrated prediction

Calibrated prediction is making statements about Y on the basis of z^{obs}. We already have a crude approach to calibrated prediction, which is to plug the point estimate of θ^* into the simulator, which then gives a point prediction for Y. This point estimate may be either ML or Bayesian, so we denote it as $\hat{\theta}$, and the point prediction is $\hat{y} \triangleq g(\hat{\theta})$. More generally, the distribution $f_{Y|\theta^*}(y \mid \hat{\theta})$ gives a distribution for Y. In the case of the gaussian best input approach, this distribution is multivariate gaussian with mean $\hat{y}$ and variance Σ.

There are two limitations to this plug-in approach. First, it does not account for the fact that $\hat{\theta}$ is not an exact estimate of θ^*. Second, it does not account fully for the impact of z^{obs} on Y. Both of these issues are addressed in the Bayesian approach to calibrated prediction, which is illustrated here for the all-gaussian case. There is no equivalent Frequentist approach.

In the Bayesian approach, we start by writing down the density function that we require, and then we use the probability calculus to render this in a form that is calculable using the conditional and marginal PDFs that we have specified. In our case, what we would like is $f_{Y|Z}(y \mid z^{\text{obs}})$, and then the probability calculus tells us

$$f_{Y|Z}(y \mid z^{\text{obs}}) = \int f_{Y|\theta^*,Z}(y \mid \theta, z^{\text{obs}})\, f_{\theta^*|Z}(\theta \mid z^{\text{obs}})\, \mathrm{d}\theta. \tag{21}$$

Basically, we introduce θ^* and then integrate it out again, using what is known as the 'law of total probability'. The second term in the integrand is the posterior distribution of θ^*, described this in section 3.2. So we assume that we have a mechanism for generating a sample of m values from this distribution, and write the predictive distribution for Y as the approximation

$$f_{Y|Z}(y \mid z^{\text{obs}}) \approx m^{-1} \sum_{i=1}^{m} f_{Y|\theta^*,Z}(y \mid \theta^{(i)}, z^{\text{obs}}) \tag{22}$$

for $\theta^{(i)} \sim f_{\theta^*|Z}(\theta \mid z^{\text{obs}})$. This Monte Carlo approximation is based on the 'weak law of large numbers': it is consistent (the error goes to zero as $m \to \infty$), and its accuracy can be approximated using confidence intervals from the gaussian distribution, using the Central Limit Theorem.

Here it is clear that uncertainty about θ^* is taken into account. The plug-in approach, in which $\hat{\theta}$ is used in place of θ^*, would collapse the sum in (22) to a single value, $f_{Y|\theta^*,Z}(y \mid \hat{\theta}, z^{\text{obs}})$, which is equivalent to replacing $f_{\theta^*|Z}(\theta \mid z^{\text{obs}})$ in (21) with the Dirac function $\delta(\theta^* - \hat{\theta})$. This replacement would be a reasonable approximation only when the calibration distribution was highly concentrated.

There remains a problem, however. In general, the PDF $f(y \mid \theta, z)$ is not directly available, and so (22) is not helpful. The very attractive feature of the all-gaussian approach is that in this case it *is* available. In fact, $(Y, Z) \mid \theta^*$ is multivariate gaussian, and it follows that $Y \mid Z, \theta^*$ is also multivariate gaussian. Therefore we can approximate the modal value of $f_{Y|Z}(y \mid z^{\text{obs}})$ by maximising (22) over y; this would serve as a good point-estimate of Y. Slightly more complicated techniques can be used to find the probability that Y lies in any specified set $\mathcal{Y}$.

To show how the Bayesian calibrated prediction approach addresses the second limitation, suppose for simplicity that the calibration distribution is highly concentrated around $\hat{\theta}$, and consider, for illustration, the mean of the distribution $f_{Y|\theta^*,Z}(y \mid \hat{\theta}, z^{\text{obs}})$. This is

$$E(Y \mid \hat{\theta}, z^{\text{obs}}) = g(\hat{\theta}) + \Sigma H^T \big(H \Sigma H^T + T\big)^{-1} \big(z^{\text{obs}} - H g(\hat{\theta})\big) \qquad (23)$$

according to the standard result for multivariate gaussian conditioning. Clearly this is the plug-in value $\hat{y} = g(\hat{\theta})$ *plus an additional term.* This additional term carries the information from z^{obs} to Y that is not captured by $\hat{\theta}$. So, for example, if some components of Z are also components of Y, then knowing z^{obs} allows us to know Y, modulo measurement error. Likewise, if there are judged to be systematic mis-predictions in the simulator (represented by non-zero off-diagonal components in Σ), then knowing how the simulator mis-predicts Z, as shown in $z^{\text{obs}} - Hg(\hat{\theta})$, allows us to correct $g(\hat{\theta})$. From this point-of-view, the calibrated prediction approach is ‘bias correcting’.

5 Emulators

An emulator is a statistical framework for predicting the simulator output at untried values of the parameters. An emulator is needed if it is not possible to evaluate the simulator as much as we would like, usually because the simulator is slow to evaluate, or the number of parameters is large, or both. Emulators are also required if the analyst does not have control over the simulator evaluations, but must use an ensemble of evaluations prepared for some other purpose. Emulators can be used in both the ML and the Bayesian approach. Usually, though, they are constructed on Bayesian principles.

The most popular emulators in statistics are those constructed from Gaussian Processes (GPs). These can efficiently handle large numbers of multivariate simulator outputs, and provide a flexible framework for incorporating extra information about the how the simulator output is expected to change in response to changes in the parameter value. There are 'out-of-the-box' settings for GPs which are widely advocated, but for important problems I would definitely recommend a hand-crafted emulator with lots of expert input.

A GP emulator is built from a carefully chosen collection of simulator evaluations: selecting those evaluations is an exercise in *experimental design*. The standard approach, in the absence of information about which simulator parameters are important, is to use a maximin latin hypercube for the design, possibly in several stages. The first stage is a *screening* stage, in which the important parameters are identified—these are termed the *active* parameters. Subsequent stages ensure that the sub-space of the active parameters is well-spanned, and that evaluations are available at a range of inter-point spacings, to learn about the different scale lengths of the simulator's response.

The output from a GP emulator is a *mean function*, denoted $\mu(\theta)$, and a *variance function*, denoted $\Upsilon(\theta)$. At parameter values where the simulator has been run, the mean function is equal to the actual simulator outcome, and the variance function is a matrix of zeros. Generally, the mean function is a point-estimate of what the simulator will return when evaluated at θ, and the variance function gives the uncertainty.

Other emulation frameworks have also been proposed, but those not based on GPs have difficulty in incorporating expert judgements, handling large numbers of outputs, and making realistic appraisals of uncertainty. GP emulators tend to exploit smoothness in the response of the simulator to changes in the parameters, and so when the simulator response changes abruptly, GP emulators can struggle (usually by returning large uncertainties). In this case other approaches such as neural networks or random forests should perform better, but they have their own problems.

The emulator output is incorporated into the likelihood function. In the all-gaussian case with a GP emulator, this becomes

$$L(\theta) = \phi\big(z; H\mu(\theta), H(\Upsilon(\theta) + \Sigma)H^T + T\big), \tag{24}$$

i.e. the mean function has replaced the simulator evaluation, and the variance function has appeared alongside the discrepancy variance. Note that (24) is a highly convenient representation of the likelihood function which depends crucially on everything being gaussian, conditional on θ^*. The only slight complication is that θ now occurs in both the mean and the variance. This affects the calculation of the log-likelihood, because the Choleski decomposition is $Q(\theta)^T Q(\theta) = H(\Upsilon(\theta) + \Sigma)H^T + T$, and this must be recomputed for every θ. The log-likelihood is

$$\ell(\theta) = c' - \sum_{i=1}^{n} w_i(\theta)^2/2 - \sum_{i=1}^{n} \log q_{ii}(\theta) \tag{25}$$

where $w(\theta) = Q(\theta)^{-T}\big(z^{\text{obs}} - H\mu(\theta)\big)$. This is (13), plus a term that penalises large variances.[7] Otherwise, the inferences proceed as before.

Further reading

General references for probability theory are, roughly in increasing order of difficulty, Ross (1988), Rice (1994), DeGroot and Schervish (2002), Grimmett and Stirzaker (2001). The second and third of these also contain second-year undergraduate level material (UK) in the theory of statistical inference, necessary to understand the properties of Maximum Likelihood estimators and the construction of confidence sets. This is also covered in Wasserman (2004).

The general treatment of model limitations within the 'best input' approach is covered in Goldstein and Rougier (2004) and Goldstein and Rougier (2009). The second reference introduces the notion of 'reified modelling' of the relationship between the simulator and the system, which extends the

[7] A little thought indicates the necessity of this penalty term. Otherwise the log likelihood can be maximised by heading for an area of the parameter space where the emulator variance is large, because this will minimise the $\sum_i w_i(\theta)^2$ term.

best input approach, and allows for multiple simulators of the same system. The discrepancy $Y - g(\theta^*)$ is sometimes referred to as a 'bias' term. It has been an explicit feature of statistical modelling since Craig *et al.* (1997). Some authors advocate a detailed approach to specifying the discrepancy variance Σ (Craig *et al.*, 2001), but the more mainstream approach is to parameterise it and then to learn the parameters in a fully-Bayesian approach (Kennedy and O'Hagan, 2001), or using an estimation method such as REML (see, e.g. Santner *et al.*, 2003).

Calibration is an *inverse problem*, and there is a huge literature on this, largely pragmatic but some probabilistic. One entry-point is Tarantola (2005). The term 'calibrated prediction' was coined by Kennedy and O'Hagan (2001), who provide a fully-Bayesian treatment. Goldstein and Rougier (2006) provide a review and a Bayes linear treatment (the 'hat run' approach), suited to large and expensive simulators. Craig *et al.* (1997) provide an alternative calibration approach in which 'bad' choices of θ^* are ruled out.

The theory and practice of simulation-based inference is covered in Robert and Casella (1999) or Evans and Swartz (2000); Liu (2001) has a more application-oriented treatment. Gelman *et al.* (2003) contains practical advice, although the type of statistical model is different, being based on notions of *exchangeability* which are not so pertinent here.

The use of emulators goes back to Sacks *et al.* (1989) and Currin *et al.* (1991), the latter is the first Bayesian treatment using Gaussian Processes. Reviews of experimental design and emulator construction are given by Koehler and Owen (1996, mainly experimental design) and the book by Santner *et al.* (2003, lots on emulator construction). O'Hagan (2006) provides an accessible overview, Rougier *et al.* (2009a) an application with some 'emulator philosophy', and Rougier (2008b) gives the complete treatment for multivariate outputs. The MUCM consortium (Managing Uncertainty in Complex Models) is a UK Research Council project to develop the practice of model-based inference, and is developing a large on-line resource for Gaussian Process emulation and related issues, see `http://mucm.group.shef.ac.uk/index.html`. Gaussian Processes are also popular in the machine learning community, and Rasmussen and Williams (2006) is a very useful reference (freely available online).

Murphy *et al.* (2004) is an early exercise in Bayesian **climate prediction**, as discussed in Rougier (2004). The all-gaussian approach is described by Rougier (2007), not including an emulator. Gaussian Process emulators (for a scalar output) are developed in Rougier and Sexton (2007), Murphy *et al.* (2007), and Rougier *et al.* (2009b); the latter for combining evaluations from two climate simulators. A fully-Bayesian treatment of a climate simulator calibration including a Gaussian Process emulator (although it is somewhat hidden) is given by Sansó *et al.* (2008), and robustly critiqued by Rougier (2008a). Non-Gaussian Process emulators have been tried, such as neural networks (Sanderson *et al.*, 2008a,b).

References

P.S. Craig, M. Goldstein, J.C. Rougier, and A.H. Seheult, 2001. Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, **96**, 717–729.

P.S. Craig, M. Goldstein, A.H. Seheult, and J.A. Smith, 1997. Pressure matching for hydrocarbon reservoirs: A case study in the use of Bayes Linear strategies for large computer experiments. In C. Gatsonis, J.S. Hodges, R.E. Kass, R. McCulloch, P. Rossi, and N.D. Singpurwalla, editors, *Case Studies in Bayesian Statistics III*, pages 37–87. New York: Springer-Verlag. With discussion.

C. Currin, T.J. Mitchell, M. Morris, and D. Ylvisaker, 1991. Bayesian prediction of deterministic functions, with application to the design and analysis of computer experiments. *Journal of the American Statistical Association*, **86**, 953–963.

M. H. DeGroot and M.J. Schervish, 2002. *Probability and Statistics*. Reading, Mass.: Addison-Wesley Publishing Co., 3rd edition.

M. Evans and T. Swartz, 2000. *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, 2003. *Bayesian Data Analysis*. Chapman & Hall, 2nd edition.

M. Goldstein and J.C. Rougier, 2004. Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM Journal on Scientific Computing*, **26**(2), 467–487.

M. Goldstein and J.C. Rougier, 2006. Bayes linear calibrated prediction for complex systems. *Journal of the American Statistical Association*, **101**, 1132–1143.

M. Goldstein and J.C. Rougier, 2009. Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*, **139**, 1221–1239. With discussion.

G.R. Grimmett and D.R. Stirzaker, 2001. *Probability and Random Processes*. Oxford University Press, 3rd edition.

M.C. Kennedy and A. O'Hagan, 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B*, **63**, 425–450. With discussion, pp. 450–464.

J.R. Koehler and A.B. Owen, 1996. Computer experiments. In S. Ghosh and C.R. Rao, editors, *Handbook of Statistics, 13: Design and Analysis of Experiments*, pages 261–308. North-Holland: Amsterdam.

J.S. Liu, 2001. *Monte Carlo Strategies in Scientific Computing.* New York: Springer.

J.M. Murphy, B.B.B. Booth, M. Collins, G.R. Harris, D.M.H. Sexton, and M.J. Webb, 2007. A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philosophical Transactions of the Royal Society, Series A*, **365**, 1993–2028.

J.M. Murphy, D.M.H. Sexton, D.N. Barnett, G.S. Jones, M.J. Webb, M. Collins, and D.A. Stainforth, 2004. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772.

A. O'Hagan, 2006. Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety*, **91**, 1290–1300.

C.E. Rasmussen and C.K.I. Williams, 2006. *Gaussian Processes for Machine Learning.* MIT Press. Available online at `http://www.GaussianProcess.org/gpml/`.

J.A. Rice, 1994. *Mathematical Statistics and Data Analysis.* Wadsworth Publishing Co Inc, 2nd edition.

C.P. Robert and G. Casella, 1999. *Monte Carlo Statistical Methods.* New York: Springer.

S. Ross, 1988. *A First Course in Probability.* New York: Macmillan, 3rd edition.

J.C. Rougier, 2004. Brief comment arising re: "Quantification of modelling uncertainties in a large ensemble of climate change simulations" by Murphy et al (Nature, 2004). Unpublished, available at `http://www.maths.bris.ac.uk/~mazjcr/commentMurphyetal.pdf`.

J.C. Rougier, 2007. Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change*, **81**, 247–264.

J.C. Rougier, 2008a. Discussion of 'Inferring climate system properties using a computer model' by Sansó *et al. Bayesian Analysis*, **3**(1), 45–56.

J.C. Rougier, 2008b. Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics*, **17**(4), 827–843.

J.C. Rougier, S. Guillas, A. Maute, and A. Richmond, 2009a. Expert knowledge and multivariate emulation: The Thermosphere-Ionosphere Electrodynamics General Circulation Model (TIE-GCM). Forthcoming in *Technometrics*, currently available at `http://www.maths.bris.ac.uk/~mazjcr/TIEGCM.pdf`.

J.C Rougier and D.M.H. Sexton, 2007. Inference in ensemble experiments. *Philosophical Transactions of the Royal Society, Series A*, **365**, 2133–2143.

J.C. Rougier, D.M.H. Sexton, J.M. Murphy, and D. Stainforth, 2009b. Analysing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments. *Journal of Climate*, in press, DOI: 10.1175/2008JCLI2533.1.

J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn, 1989. Design and analysis of computer experiments. *Statistical Science*, **4**(4), 409–423. With discussion, pp. 423–435.

B.M. Sanderson, R. Knutti, T. Aina, C. Christensen, N. Faull, D.J. Frame, W.J. Ingram, C. Piani, D.A. Stainforth, D.A. Stone, and M.R. Allen, 2008a. Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. *Journal of Climate*, **21**, 2384–2400.

B.M. Sanderson, C. Piani, W.J. Ingram, D.A. Stone, and M.R. Allen, 2008b. Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics GCM simulations. *Climate Dynamics*, **30**, 175–190.

B. Sansó, C. Forest, and D. Zantedeschi, 2008. Inferring climate system properties using a computer model. *Bayesian Analysis*, **3**(1), 1–38. With discussion, pp. 39–62.

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

A. Tarantola, 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial Mathematics.

L.A. Wasserman, 2004. *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer-Verlag Inc.