Technical Report
Likelihood Methods for Estimating Model Discrepancy

Allan Seheult[*]

Department of Mathematical Sciences, Durham University,
Science Laboratories, Stockton Road, Durham DH1 3LE, UK

September 30, 2010

Abstract

This report describes methods for estimating a parameterised version of "model discrepancy" regarded as a stochastic representation of the difference between a computer simulator of a system and the system itself. Likelihood methods are illustrated on published examples.

Keywords: Computer Models, Emulation, Galaxy Formation, Hydrocarbon Reservoirs, Model discrepancy, Reification, Simulators.

1   Introduction

Mathematical models of complex physical systems, such as those for reservoirs, galaxy formation and climate are usually implemented as a computer code, often referred to as a "simulator". In this account, we consider inference for the inevitable discrepancy between the mathematical model and the physical system it purports to represent. We refer to this discrepancy as "model discrepancy'.

It is widely accepted that the uncertainties associated with both calibrating a mathematical model to observations on a physical system and prediction for a physical system should take account of model discrepancy. However, model discrepancy specification is challenging, drawing on expert knowledge of the physical system and any crucial simplifications assumed in the mathematical model of the system; see, for example, Craig et al. (1998).

While we include some discussion on model discrepancy specification in particular examples, we mostly focus attention on inference for model discrepancy when it has been partially specified. We develop likelihood and Bayesian methods and illustrate them on two reasonably substantive examples discussed elsewhere: (i) a model for galaxy formation proposed by Bower et al. (2006) for which Goldstein and

---

[*]email: a.h.seheult@durham.ac.uk; tel: +44 (0)191 334 3046; fax: +44 (0)191 334 3051.

Vernon (2009) describe a careful specification exercise of model discrepancy with the cosmologists, linked to an extensive analysis of model calibration using the notion of "implausibility", detailed for example in Craig et al. (1997) and (ii) a hydrocarbon reservoir model, calibrated in Craig et al. (1997) and used in Craig et al. (2001) to illustrate forecasting for computer models using Bayes linear methods; see Goldstein and Wooff (2007). We include a brief outline of a Bayes Linear approach to estimating model discrepancy which will be explored fully in a subsequent report.

2   Framework

In this section, we start by outlining the construction of an emulator of a simulator for a mathematical model of a physical system. We then model a relationship linking observations on the physical system and the simulator using the notion of a "best input" to the simulator. This allows us to combine observations on the physical system with an ensemble of runs on the simulator to infer about unknowns, such as the best input, predictions of the physical system and model discrepancy.

2.1   The emulator

Consider a deterministic mathematical model of a complex physical system implemented as computer code $f(\cdot)$, called a simulator, which we can evaluate as $f(x)$ at any allowable input x. As we cannot evaluate f at every allowable input, we specify prior beliefs about $f(x)$ for each x and update these beliefs using the results of well chosen simulator runs. These updated beliefs comprise the emulator of f. We proceed as follows.

We specify prior beliefs about the value of component $f_i(x)$ of the vector $f(x) = (f_1(x), \ldots, f_k(x))$ of k simulator outputs at any allowable input $x = (x_1, \ldots, x_r)$ as

$$f_i(x) = \sum_j \beta_{ij} \, g_j(x) + u_i(x) \tag{1}$$

where the components of the vector $g(x) = (g_1(x), \ldots, g_p(x))$ are p specified "regression" functions, which we will assume to have the same form for each output, the β_{ij} are pk unknown coefficients, and $u(x) = (u_1(x), \ldots, u_k(x))$ is a random vector with expectation zero and variance matrix Σ, the same for each input x. We can write these relationships in vector form as

$$f(x) = g(x)\beta + u(x) \tag{2}$$

where $\beta = (\beta_{ij})$ is a $p \times k$ matrix.

When we have specific prior beliefs about g, β and the process $u(x)$, these combine to give prior beliefs about $f(x)$, the "prior emulator". Such specific prior beliefs may arise from expert judgement or many runs on a "fast" approximation to $f(x)$, or a combination of both, as described, for example, in Craig et al. (1996).

We now run the simulator at n inputs to give n outputs which we assemble in the $n \times r$ matrix X and the $n \times k$ matrix F, respectively; and we refer to $S = (X, F)$ as an "ensemble" of simulator runs. We can now write

$$F = G\beta + U \tag{3}$$

where G is the $n \times p$ "model matrix" and U is an $n \times k$ matrix of "random errors". In this account, we make the following simplifying assumptions:

(i) Each row of U has the same variance matrix Σ

(ii) Each column of U has the same correlation matrix C, where for any two input vectors x_i and x_j, rows i and j of X, we choose the correlation between the corresponding outputs $f(x_i)$ and $f(x_i)$, rows i and j of F, to have the exponential form

$$C_{ij} = \exp\left[-(x_i - x_j)\Theta^{-2}(x_i - x_j)^\mathrm{T}\right] \tag{4}$$

where Θ is a diagonal matrix of "correlation lengths" $\theta = (\theta_1, \ldots, \theta_r)$, one for each input component.

(iii) U has a matrix normal distribution with probability density function

$$p(U) = \frac{\exp[-\frac{1}{2}\mathrm{trace}(\Sigma^{-1}U^\mathrm{T}C^{-1}U)]}{(2\pi)^{\frac{nk}{2}}|\Sigma|^{\frac{n}{2}}|C|^{\frac{k}{2}}} \tag{5}$$

(iv) The prior distribution for β and Σ has the so-called "non-informative" form

$$p(\beta, \Sigma|\theta) \propto |\Sigma|^{-\frac{k+1}{2}} \tag{6}$$

To "emulate" an unknown simulator value $f(x)$ at an input x, we start by considering the conditional distribution of $f(x)$ given the ensemble of runs S and θ, which standard calculations, such as those in Conti et al. (2009), show is the k-variate Student t-distribution

$$f(x)|S, \theta \sim T_k\left[g(x)\hat{\beta} + c(x)C^{-1}(F - G\hat{\beta}); l(x)\hat{\Sigma}; n - p\right] \tag{7}$$

with $n - p$ "degrees-of-freedom", where $\hat{\beta} = (G^\mathrm{T}C^{-1}G)^{-1}G^\mathrm{T}C^{-1}F$ is the multivariate generalised least squares estimate of β, $\hat{\Sigma} = (F - G\hat{\beta})^\mathrm{T}C^{-1}(F - G\hat{\beta})/(n - p)$ is the associated "unbiased" estimate of Σ, $c(x)$ is the vector of the n correlations of $u(x)$ with $u(x_1), \ldots, u(x_n)$, and the scalar function $l(x)$ is given by

$$l(x) = 1 - c(x)C^{-1}c(x)^\mathrm{T} + \left[g(x) - G^\mathrm{T}C^{-1}c(x)^\mathrm{T}\right]\left[G^\mathrm{T}C^{-1}G\right]^{-1}\left[g(x) - G^\mathrm{T}C^{-1}c(x)^\mathrm{T}\right]^\mathrm{T} \tag{8}$$

Note that the emulator exactly interpolates the ensemble S; that is, $l(x_i) = 0$ for every row x_i of X, so that the emulator variance is zero, and therefore $f(x_i) = g(x_i)\hat{\beta} + c(x_i)C^{-1}(F - G\hat{\beta})$ for each row x_i of X and the corresponding row $f(x_i)$ of F

If we emulate $f(x)$ at several inputs simultaneously, the conditional distribution is a matrix t distribution: see, for example, Kotz and Nadarajah (2004).

We use $p(f(x)|S, \hat{\theta})$ as the "emulator" for $f(x)$, where $\hat{\theta}$ is the REML estimate of θ obtained by maximising the log-likelihood function

$$L(\theta) = -\frac{1}{2}\left[(n - p)\log|\hat{\Sigma}| + k\log|C| + k\log|G^\mathrm{T}C^{-1}G|\right] \tag{9}$$

resulting in the likelihood that appears in the the posterior distribution of the correlation parameters in Kennedy and O'Hagan (2001) when using an improper prior distribution for $\log(\theta)$.

This generalises a result of Harville (1974) to multivariate regression with matrix normal errors. The hessian $L''(\hat{\theta})$ provides standard errors and approximate confidence intervals in the usual way. Experience suggests that it is better to re-parameterise in terms of $(\log \theta_1, \ldots, \log \theta_r)$, leading to a better quadratic approximation to the log-likelihood function and uncorrelated estimates; see Nagy et al. (2007a). When n is large compared to $\max\{k, p, r\}$, the posterior distribution of $(\log \theta_1, \ldots, \log \theta_r)$ will be approximately multivariate normal with mean vector $(\log \hat{\theta}_1, \ldots, \log \hat{\theta}_r)$ and precision matrix $-L''(\log \hat{\theta})$. Thus, while uncertainty in emulation, calibration and prediction should take account of the uncertainty in θ, this will not be pursued here: however, see, for example, Nagy et al. (2007b).

When some components of the input vector x are omitted in the prior specification for $f_i(x)$, we may choose to include a "nugget effect" in (1) which we assume to have expectation zero and variance $\delta \sigma_i^2$, where σ_i^2, the i-th diagonal element of Σ, is the variance of $u_i(x)$ and $0 \leq \delta \leq 1$; see, for example, Cumming and Goldstein (2009). With this formulation, we replace C by $\delta I + (1-\delta)C$ and $c(x)$ by $(1-\delta)c(x)$ in the above. Note that when δ is positive the emulator no longer interpolates the ensemble S.

3 Linking the simulator to reality

In what follows, we work with the so-called "best input" approach, in which the "system output" y is related to the simulator $f(\cdot)$ evaluated at a special input x^* by the additive relationship

$$y = f(x^*) + \varepsilon \tag{10}$$

where ε, called the "model discrepancy" and x^*, called the "best input" are such that $\varepsilon \perp\!\!\!\perp \{f, x^*\}$. Usually, but not always, we assume that $\mathrm{E}[\varepsilon] = 0$ and write $\mathrm{Var}[\varepsilon] = \Sigma_\varepsilon$. This formulation has been critically analysed by Goldstein and Rougier (2009) who replace it by the notion of a "reified" model. It is the discrepancy between the reified model and reality which they regard as model discrepancy. Moreover, the process which leads them to the reified model also acts as a basis for assessing the model discrepancy variance Σ_ε.

In practice, we have measured values z of the components of y or a reduced set of q linear features yH, so that

$$z = yH + e \tag{11}$$

where the measurement error term e is such that $e \perp\!\!\!\perp y$, $\mathrm{E}[e] = 0$ and $\mathrm{Var}[e] = \Sigma_e$ is usually assumed known.

In this account, we further assume that $e \sim N(0, \Sigma_e)$, $\varepsilon \sim N(0, \Sigma_\varepsilon)$ and, either $x^* \sim N(\mu_*, \Sigma_*)$ for "known" μ_* and Σ_*, or $x^* \sim U(a, b)$ for known a and b.

Inferences about the best input and prediction for the system are now based on z, S, the relationships (10) and (11) and the various normality assumptions.

4

4   Likelihood and Bayesian inference for model discrepancy

In this section, we consider how to estimate a parameterised specification of the variance $\Sigma_\varepsilon(\varphi)$ of model discrepancy ε using the emulator and system observations z. The parameter φ varies in a space with dimension less than k.

We choose likelihood as a basis for inference about φ, unless there are prior beliefs about φ, in which case we use its posterior distribution.

The likelihood for $l(\varphi)$ for φ is such that

$$l(\varphi) \propto \int p\left(z \mid S, \varphi, x^*\right) p\left(x^*\right) dx^* \tag{12}$$

where $p\left(x^*\right)$ is the prior distribution for x^*. The expectation and variance of the first distribution in the integrand can be computed as

$$\mathrm{E}[z \mid S, \varphi, x^*] = \mathrm{E}[z \mid S, x^*] = \mu(x^*)H \tag{13}$$

and

$$\mathrm{Var}[z \mid S, \varphi, x^*] = H^{\mathrm{T}}[\Sigma(x^*) + \Sigma_\varepsilon(\varphi)]H + \Sigma_e \tag{14}$$

where $\mu(x)$ and $\Sigma(x)$ are the emulator mean and emulator variance at input x; and for simplicity, we assume a Gaussian distribution for $p\left(z \mid S, \varphi, x^*\right)$.

The integral in (12), which gives the likelihood for any particular value of φ, is computed using numerical integration or by simulating from the prior distribution $p\left(x^*\right)$ for x^*. We can then proceed to compute the maximum likelihood estimate $\hat{\varphi}$ and confidence regions for φ using the Hessian of the log-likelihood function at $\hat{\varphi}$. However, it is perhaps preferable to regard the likelihood as a measure of "support" for each allowable value of φ, as in Edwards (1972). This likelihood approach is illustrated for the examples in Section 5.

If we are prepared to quantify our prior information about φ in terms of a prior distribution, then we may base inferences on its posterior distribution, computed using Bayes theorem in the usual way.

4.1   Inference for model discrepancy with fast simulators

When a simulator is fast to run, it is it's own emulator and equations (12), (13) and (14) reduce to (15), (16) and (17) below.

$$l(\varphi) \propto \int p\left(z \mid \varphi, x^*\right) p\left(x^*\right) dx^* \tag{15}$$

$$\mathrm{E}[z \mid \varphi, x^*] = \mathrm{E}[z \mid x^*] = f(x^*)H \tag{16}$$

$$\mathrm{Var}[z \mid \varphi, x^*] = H^{\mathrm{T}}\Sigma_d(\varphi)H + \Sigma_e \tag{17}$$

Since in this case $\mathrm{Var}[z \mid \varphi, x^*]$ is not a function of x^*, computation of the integral in (15) is simpler.

The first distribution in the integrand in both (12) and (15) may also be interpreted as a joint likelihood function for φ and x^*. Integration over x^* hides the potential for this joint likelihood surface to be multimodal, as there will often be fits to the observations for some choices of x^* with small variance and low correlation

across outputs and other fits with large variance and high correlation across outputs. However, as we assume there is a prior distribution for x^*, $l(\varphi)$ is the likelihood for φ.

5 Examples

5.1 Hydrocarbon reservoir

System: Craig et al. (2001) consider an active, mostly gas-producing hydrocarbon reservoir, comprising one mainly onshore field and three offshore fields, previously considered by Craig et al. (1997) in a case study on "history matching". Temporal series of bottom hole pressure at six onshore producing wells were available.

Model: A computer model of the reservoir, which was constructed by reservoir engineers using commercial software, includes reservoir structure, geometry, fault patterns and spatial distributions of permeability and porosity.

Inputs: Among the many simulator inputs, the focus was on seven permeability multipliers (range $[0.1, 10.0]$) for the seven regions into which the reservoir was divided, and 33 fault transmissibility multipliers (range $[0, 1]$). Previous experience and the engineer's judgements suggested that logarithms of the permeability multipliers are more suitable for use in statistical modelling. Each input was linearly transformed to vary over $\left[-\frac{1}{2}, \frac{1}{2}\right]$.

Outputs: Among the simulator outputs there were 34 bottom hole pressures distributed between seven wells through time. The corresponding observed bottom-hole pressures z, well numbers and times of measurement are as follows:

Pressure:

```
149.8 138.0 134.8 136.0 123.2 129.7 118.0 123.0 119.7 127.5 117.6
115.0 114.0 112.7 120.5 117.5 107.8 112.7 119.1 117.7 111.2 116.0
110.3 115.9 114.1 116.1 106.8 115.0 106.0  94.9 100.3  98.1 114.7
 94.4
```

Well number:

```
30 92 11 92 11 30 89 92 11 30 89 92 92 11 30 38 89
11 30 38 89 92 11 30 38 56 89 92 92 11 30 38 56 89
```

Measurement time:

```
3377 3377 3742 3742 4168 4199 4199 4199 4504 4504 4504 4504 5325
5326 5326 5326 5326 5721 5721 5721 5721 5721 5995 5995 5995 5995
5995 5995 7622 7640 7640 7640 7640 7640
```

Each run of the computer simulator at a specified set of inputs took an average of about 10 hours.

Design: 101 simulator evaluations were made in a latin hypercube design across the 40-dimensional input space. However, an analysis based on a 'coarsened' version of the simulator, with the same inputs and outputs but with larger grid blocks and

bigger time steps, with each run taking about only 3 minutes, showed that just 4 of the permeability inputs were active. Thus, X is 101×4 and F is 101×34.

Model Discrepancy: There were two sources of information to help specify Σ_ε. First, the reservoir engineer suggested that median absolute error of about 5% would be appropriate for each of the components of ε. Secondly, there was available what was judged to be a "best run" of the simulator in the history matching exercise from Craig et al. (1997). A simple analysis using the difference between simulator outputs at this best run and the actual field data z, suggested that there were temporal effects and well effects, a hypothesis supported by the engineer. These considerations lead to the model choice

$$\text{Cov}[\varepsilon_k, \varepsilon_\ell] = \sigma_1^2 \exp\left(-\theta_1(T_k - T_\ell)^2\right) + \sigma_2^2 \exp\left(-\theta_2(T_k - T_\ell)^2\right) I_{W_k = W_\ell} \qquad (18)$$

where the k-th output component comes from well W_k at time T_k, and I_P denotes the indicator function of the proposition P. Values were assigned to the four parameters by informal data analysis using a combination of guessing-and-simulating and variogram methods, resulting in $\sigma_1^2 = 25$, $\theta_1 = (6 \times 10^{-4})^2$, $\sigma_2^2 = 6$ and $\theta_2 = (2 \times 10^{-3})^2$. The resulting variances for the individual components of ε are smaller than, but of the same order of magnitude as those suggested by the engineer. A more sophisticated model for a larger-scale study might replace the well "effects" by spatial modelling.

Measurement Error: The engineer suggested a standard deviation of 3% of observed bottom hole pressure.

5.1.1 Inference for the example

We re-parameterised Σ_ε by setting $\varphi = (\log(\sigma_1), -0.5\log(\theta_1), \log(\sigma_2), -0.5\log(\theta_2))$ with specified values $(1.61, 7.42, 0.90, 6.22)$.

The maximum likelihood estimates of the φ components are $(2.17, 7.79, 1.60, 8.23)$ with corresponding hessian based standard errors $(0.49, 0.39, 0.37, 0.58)$. The first two components are uncorrelated with the last two components, while the correlation within each of these two pairs is about 40%. Thus, even though the first three specified components of φ are inside their nominal 95% confidence intervals, the fourth is well outside its interval, casting doubt on the overall specification of Σ_ε.

The structure of the specification of $\text{Var}[\varepsilon]$ and $\text{E}[\varepsilon]$ may too simple, possibly by not including actual distance between wells.

5.2 Galaxy formation

System: Current cosmology theories suggest the Universe began about 13 billion years ago and has been expanding ever since. However, observations imply there exists far more matter than the visible matter that makes up the stars and planets. The deficit is called "dark matter", and understanding its nature and how it has affected galaxy evolution is a major problem in cosmology.

Model: Cosmologists simulate galaxy formation from the beginning of the Universe in two parts. First, dark matter is simulated to determine early Universe mass fluctuation and subsequent growth into galaxies.. Second, the dark matter simulation

results are inputs into Galform which simulates interactions of gas cloud formation, radiative cooling, star formation and the effects of black holes. The first simulation is run on a space volume of $(1.63 \text{ billion light-years})^3$ which is divided into 512 sub-volumes which are independently simulated with Galform. Each Galform run takes between 20 and 30 minutes.

Inputs: Galform has 16 input parameters that cosmologists were interested in varying. However, only 10 of the inputs were judged to be "active". Inputs were linearly transformed to vary over $[-1, 1]$.

Outputs: Goldstein and Vernon (2009) focus on two outputs, the Bj and K band luminosity functions. The Bj band gives the number of young galaxies of a certain luminosity per unit volume, while the K band describes the number of old galaxies. We compare the average over 40 of the 512 sub-volumes of the logarithms of $k = 63$ representative outputs (34 from the Bj band and 29 from the K band) to the corresponding observational data z from the 2dFGRS galaxy survey; see Fig. 5.2.

Design: 2000 evaluations of Galform were made in a carefully chosen design across the 16-dimensional input space, following several "waves" of "history matching". Sixteen evaluations were "burnt" (the model failed to compute), so that X is 1984 × 16 and F is 1984 × 69. Fig. 5.2 shows the observed Bj and K luminosity curves with one of the 1984 Galform runs indicated.

Model Discrepancy: A leading cosmologist's opinion is that there is no overall bias of the model, so that $\text{E}[\varepsilon] = 0$. On the other hand, he identified two possible major physical defects of Galform: (i) the model may have too much (or too little) mass in the simulated universe, leading to the luminosity outputs all being too high (or too low), suggesting positive correlation between all luminosity outputs; and (ii) galaxies might age at the wrong rate, leading to more/less young galaxies and therefore less/more old galaxies, suggesting a smaller negative correlation between the Bj and K luminosity outputs. To respect the symmetries of these possible defects, $\text{Var}[\varepsilon]$ was parameterised as

$$
\Sigma_\varepsilon(a, b, c) = a
\begin{bmatrix}
1 & b & .. & c & .. & c \\
b & 1 & .. & c & . & c \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
c & .. & c & 1 & b & .. \\
c & .. & c & b & 1 & .. \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{bmatrix}
\tag{19}
$$

so that $\varphi = (a, b, c)$, where a is a variance and b and c are correlations. However, as the specifications of a, b and c were imprecise, the cosmologist gave upper and lower values, $\underline{a} = 1.41 \times 10^{-3}, \overline{a} = 5.66 \times 10^{-3}, \underline{b} = 0.4, \overline{b} = 0.8$ and $\underline{c} = 0.2, \overline{c} = b$.

Measurement Error: There are several contributions, including normalisation error, luminosity zero point error and an error correcting for galaxies being seen in the past and receding at different speeds. As these errors are well understood, it is reasonable to treat their overall effect as precisely specified uncertain quantities e with $\text{E}[e] = 0$ and known $\text{Var}[e] = \Sigma_e$.
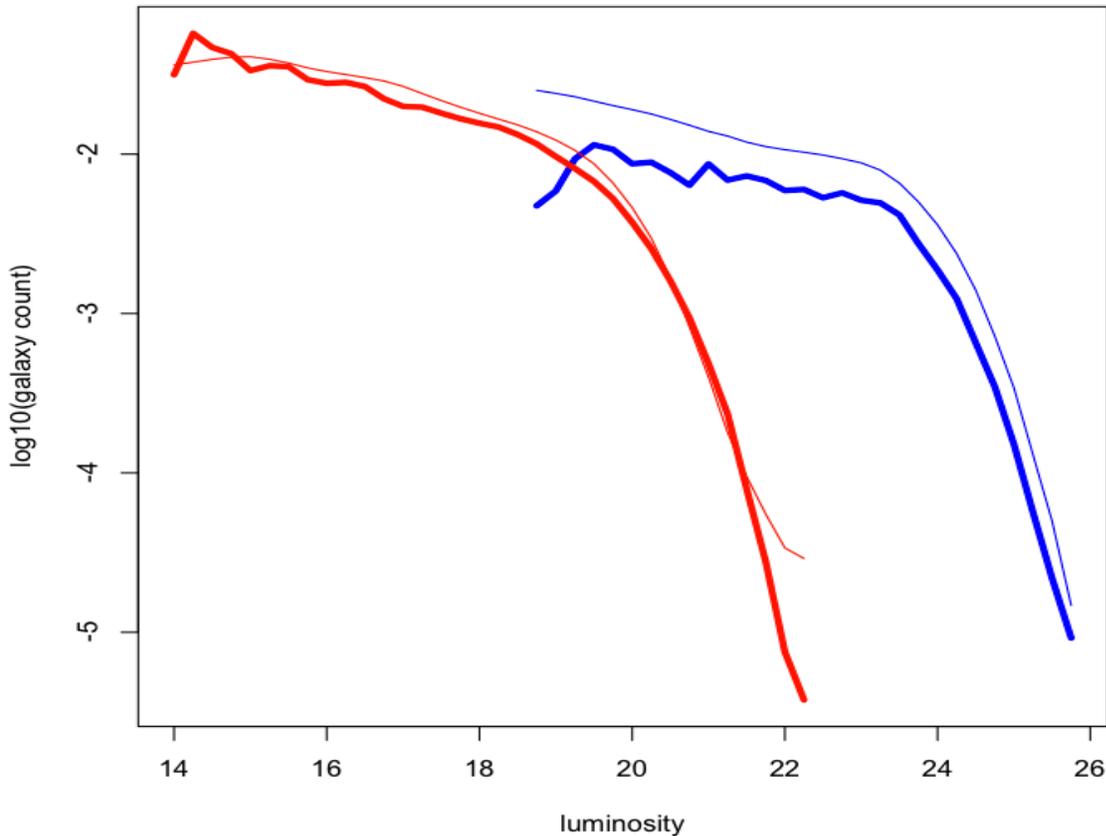
Figure 1: Bj (red) band and K (blue) band 2dFGRS galaxy survey observed luminosity functions (thick lines) and from a single run of Galform (thin lines).

5.2.1 Inference for the example

As the number of runs is "large", the fast version of likelihood given in (15), (16) and (17) was used: an analysis using an emulator with a cubic mean function in all 10 active inputs gave similar results.

We re-parameterised Σ_ε by setting $\varphi = (0.5 \log(a), \operatorname{arctanh}(b), \operatorname{arctanh}(c))$. However, as we might anticipate, likelihood inference for (a, b, c) is unreliable: estimation of correlation and variance in constant correlation models is problematical because correlation and variance are confounded; for example, it is particularly difficult to disentangle the difference between a small variance and a large correlation. We focus instead on the variance a of model discrepancy, the most important of the three parameters.

The correlations b and c were varied over the four combinations of their specified upper and lower bounds. Fig. 5.2.1 shows the likelihood analysis for the variance parameter a. It can be seen that:

(i) the maximum likelihood estimate $\hat{a}$ is inside cosmologist's uncertain interval;

(ii) the uncertain interval for a is inside the approximate 95% limits;

(iii) values of a outside its uncertain interval are supported;

(iv) uncertainty about a increases with increasing b, while holding c fixed;

(v) uncertainty about a increases with decreasing c, while holding b fixed.

Remark (iv) demonstrates the effect of confounding of correlation and variance, while (v) suggests the possibility that discrepancy variance may be different for the Bj and K bands.

An alternative discrepancy model for Galform

The foregoing analysis suggests that the within and between band equal correlation assumptions may be too strong a representation of the cosmologists beliefs. Indeed, estimation of these correlations is problematical.

We consider an alternative model where the correlations decay exponentially within bands and model the Bj and K separately.

The following alternative covariance specification uses the fact that each pair of adjacent observation are at luminosity 0.25 lumens apart.

$$\text{Cov}[\varepsilon_i, \, \varepsilon_j] = a e^{-\left(\frac{\lambda_i - \lambda_j}{\theta}\right)^2} \tag{20}$$

where a (as before) is the discrepancy variance, λ_i is the luminosity at observation z_i and θ controls the correlation between any pair of discrepancies ε_i and ε_j. Additionally, we assume, as usual, that $\text{E}[\varepsilon_i] = 0$. Notice that the correlation increases with increasing θ and decreases with $|\lambda_i - \lambda_j|$ increasing for any θ.

Bj band results

Applying the fast likelihood to the Bj observations, we found that $\hat{a} = 0.001$ with $[0.0001, 0.01]$ as an approximate 95% uncertainty interval, and $\hat{\theta} = 1.1$ lumens with 95% uncertainty interval $[0.47, 2.44]$.

The estimate of a though smaller than the cosmologists assessment which as above is inside the uncertainty interval.

The estimate of θ gives a correlation of 0.95 between adjacent discrepancies and ranges from 0.75 to 0.99 across its uncertainty interval: it is essentially *zero* for the discrepancies furthest apart.

The correlation $\text{Corr}[\hat{a}, \, \hat{\theta}]$ about 10%, so that it is reasonable to infer about a and θ separately.

K band results

The corresponding results for the K band are: $\hat{a} = 0.0285$ and $\hat{\theta} = 1.72$ with 95% intervals $[0.00522, 0.155]$ and $[0.95, 3.12]$, respectively.

The correlation between adjacent discrepancies is estimated to be 0.98 and ranges from 0.93 to 0.99 across its uncertainty interval and is essentially *zero* for the discrepancies furthest apart.

The correlation $\text{Corr}[\hat{a}, \hat{\theta}]$ here is more substantial at about 54%, suggesting sensibly that the variance increases with increasing correlation, lending support to the observation in item (iv) above. The log-likelihood based 95% support region for $(\log a, \log \theta)$ shown in Fig. 5.2.1 illustrates this correlation.

Note that a is estimated to be much larger than both the cosmologist's assessment and the corresponding Bj band estimate, lending support to the observation in item (v) above.
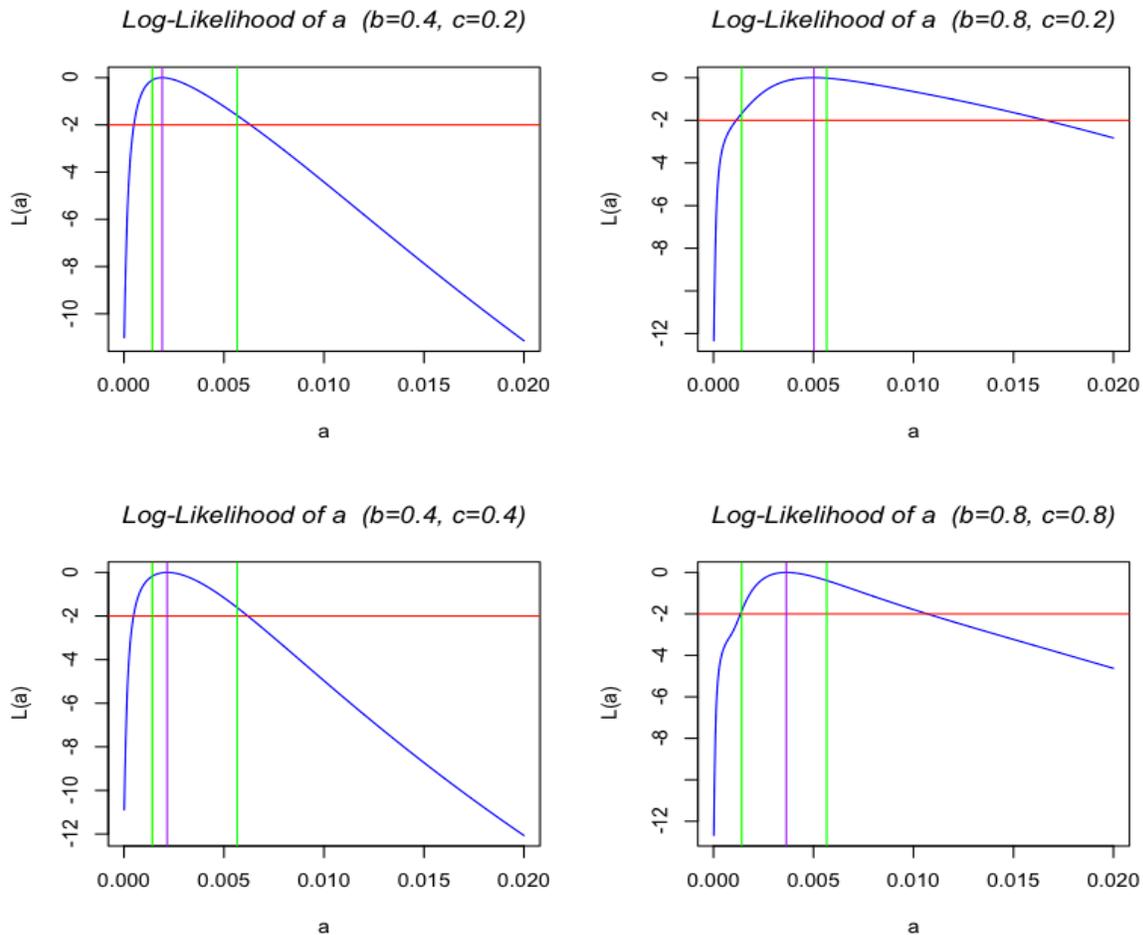
Figure 2: Log-likelihood (blue) of the variance a for each of the 4 combinations of upper and lower specifications for the correlations b and c, and the vertical lines indicate maximum likelihood estimates $\hat{a}$ (purple) and approximate 95% uncertainty intervals (green).

6 Point-wise estimation of discrepancy

Unfortunately, we are not always in a position to have an expert-elicited discrepancy model. In the absence of such expertise, empirical modelling of discrepancy over time or space, for example, can be problematical. Nevertheless, it is crucial for history matching, calibration and forecasting future system behaviour. We have recently had such an experience attempting to model and estimate temporal discrepancy
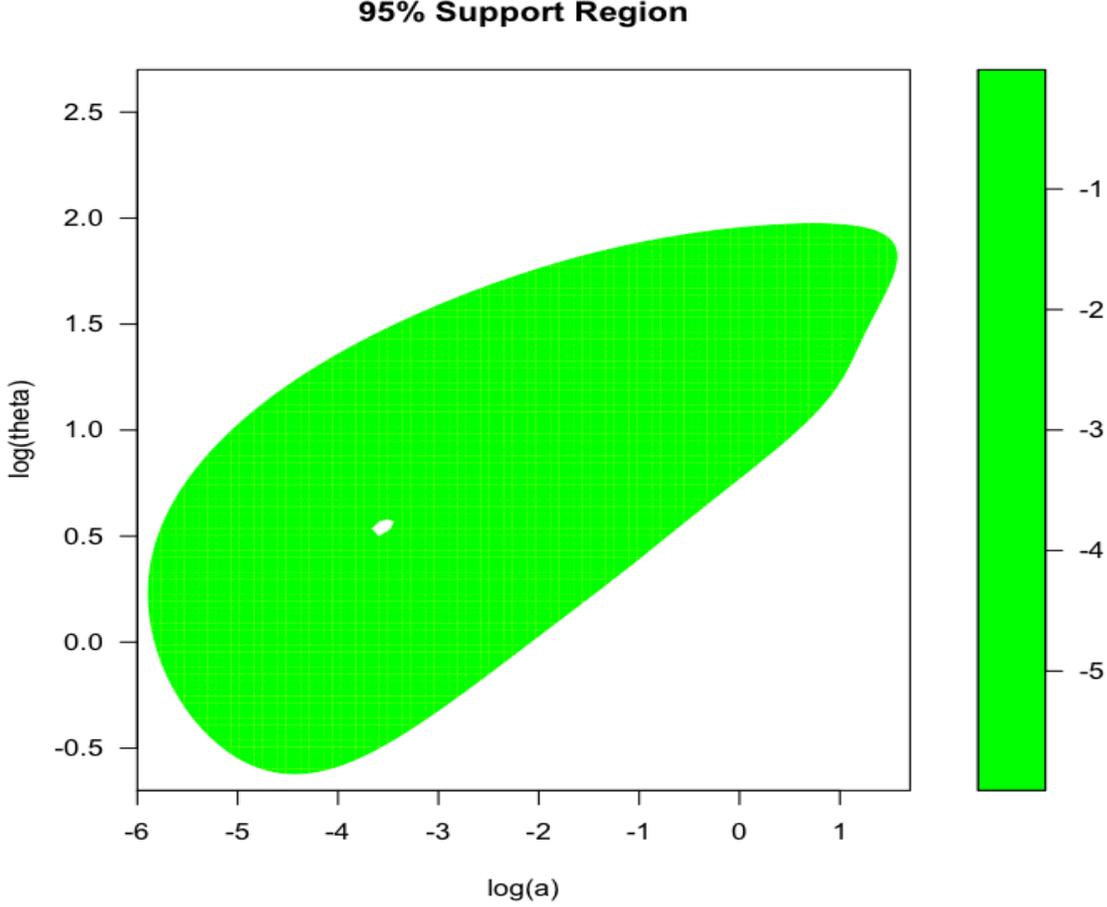
95% Support Region

Figure 3: Ninety-five percent support region for the logarithm of the variance a and the logarithm of the correlation length θ in the alternative discrepancy formulation for the Galform simulator: the maximum likelihood estimate is contained in the small white 0.5% support region.

using a stationary autoregression process to model a clearly non-stationary temporal discrepancy. While the model was estimable, the results were unusable for history matching and calibration simply because the variances resulting from naively fitting a stationary process to a non-stationary process were too large to learn about x^*.

To resolve the issue we developed a simple method to estimate the variance and expectation of discrepancy separately at each time-point. This proved useful for history matching and calibration but left unresolved the difficulty of modelling future model discrepancy to properly assess uncertainty of future system behaviour. An unexplored possibility would be to try to model the variance and expectation over time using the time point estimates.

The point-wise method needs a rough estimate f^* of $f(x^*)$, such as the mean $\bar{f}$ of the simulator runs at the time-point or the simulator run closest to z at the time-point.

With $z = f^* + \varepsilon + e$ implies $\mathrm{E}[z] = f^* + \mu_\varepsilon$ when allowing for a discrepancy bias μ_ε; and $\mathrm{Var}[z] = \sigma^2 \equiv \sigma_\varepsilon^2 + \sigma_e^2$, where σ_ε^2 is the discrepancy variance and σ_e^2 is the

measurement error variance. We now choose σ^2 and μ_ε so that

$$\frac{z - \mathrm{E}[z]}{\sqrt{\mathrm{Var}[z]}} = \frac{c - \mu_\varepsilon}{\sigma} = k \qquad (21)$$

where $c = z - f^*$ is the "empirical discrepancy", and typically $k = 3$. Squaring the condition in (21) re-expresses it as $c - \mu_\varepsilon)^2 = k^2\sigma^2$. Subject to this condition, we choose σ_ε^2 and μ_ε to minimise the mean squared error J of ε away from zero, which can be written:

$$J = \mathrm{E}[\varepsilon^2] = \mathrm{Var}[\varepsilon] + \mathrm{E}[\varepsilon]^2 = \sigma_\varepsilon^2 + \mu_\varepsilon^2$$

Subject to the condition $c - \mu_\varepsilon)^2 = k^2\sigma^2$ it is straightforward to show that J is minimised at:

$$\mu_\varepsilon = \frac{c}{1 + k^2} \quad \text{and} \quad \sigma = \frac{kc}{1 + k^2} \qquad (22)$$

Notice from (22) that (i) σ and μ_ε are always in $k{:}1$ ratio, (ii) $c = \mu_\varepsilon + k\sigma$ and (iii) the estimate of μ_ε is never *zero*.

The method worked for both history matching and calibration in the study where the issue first arose.

7 Bayes Linear estimation of model discrepancy

An interesting alternative simulation-based Bayes linear method for estimating φ is currently being investigated and will be reported elsewhere. Very briefly, simulated Bayes linear inference for φ proceeds as follows: (i) simulation to derive mean and covariance structures between z and x^*, which are used to identify the Bayes linear assessment $\hat{x}$ for x^* adjusted by z; (ii) evaluation of the hat run $\hat{f} = f(\hat{x})$, as in Goldstein and Rougier (2006); and (iii) simulation to assess the mean, variance and covariance structures across the squared components of the difference $z - \hat{f}$ and the components of φ, to carry out the corresponding Bayes linear update for φ.

8 Discussion

This report considers estimation for the discrepancy ε between a computer simulator of a complex physical system and the system itself.

The model discrepancy should be specified initially by the subject-matter experts. This is our starting point. Usually $\mathrm{E}[\varepsilon]$ and $\mathrm{Var}[\varepsilon]$ are specified, often with an added Gaussian assumption. $\mathrm{E}[\varepsilon]$ is usually specified to be zero.

When $\mathrm{Var}[\varepsilon]$ is specified to be a function of parameters φ, we use a likelihood or Bayesian analysis to learn about φ. Likelihood methods are illustrated on two reasonably substantial examples.

A important criticism of the likelihood approach is that integration over x^* to obtain the likelihood for φ hides the potential for joint likelihood surface $l(x^*, \varphi)$ to be multimodal, as there will often be fits to the observations for some choices of x^* with small variance and low correlation across outputs and other fits with large variance and high correlation across outputs. If there is a joint prior distribution for x^* and φ, the same objection may be levied at the marginal posterior distributions. Exploration of these issues, particularly from a Bayes Linear perspective, will be reported elsewhere.

Acknowledgements

References

Bower, R. G., Benson, A. J., Malbon, R., Helly, J. C., Frenk, C. S., Baugh, C. M., Cole, S., and Lacey, C. G. (2006), "The broken hierarchy of galaxy formation," *Monthly Notices of the Royal Astronomical Society*, 370, 645–655.

Conti, S., Gosling, J. P., Oakley, J. E., and O'Hagan, A. (2009), "Gaussian process emulation of dynamic computer codes," *Biometrika*, 96, 663–676.

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

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1996), "Bayes linear strategies for history matching of hydrocarbon reservoirs," in *Bayesian Statistics 5*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford, UK: Clarendon Press, pp. 69–95.

— (1997), "Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments," in *Case Studies in Bayesian Statistics*, eds. Gatsonis, C., Hodges, J. S., Kass, R. E., McCulloch, R., Rossi, P., and Singpurwalla, N. D., New York: Springer-Verlag, vol. 3, pp. 36–93.

— (1998), "Constructing partial prior specifications for models of complex physical systems," *Applied Statistics*, 47, 37–53.

Cumming, J. A. and Goldstein, M. (2009), "Bayes linear uncertainty analysis for oil reservoirs based on multiscale computer experiments," in *Handbook of Bayesian Analysis*, eds. OHagan, A. and West, M., Oxford, UK: Oxford University Press.

Edwards, A. W. F. (1972), *Likelihood*, Cambridge (expanded edition, 1992, Johns Hopkins University Press, Baltimore): Cambridge University Press.

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

— (2009), "Reified Bayesian modelling and inference for physical systems (with Discussion)," *Journal of Statistical Planning and Inference*, 139, 1221–1239.

Goldstein, M. and Vernon, I. (2009), "Bayes linear analysis of imprecision in computer models, with application to understanding the Universe," in *6th International Symposium on Imprecise Probability: Theories and Applications.*

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

Harville, D. (1974), "Bayesian inference for variance components using only error contrasts," *Biometrika*, 61, 383–385.

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

Kotz, S. and Nadarajah, S. (2004), *Multivariate t Distributions and Their Applications*, New York: Cambridge University Press.

Nagy, B., Loepky, J. L., and Welch, W. J. (2007a), "Correlation parameterization in random fuction models to improve normal approximation of the likelihood or posterior," *Report 229, University of British Colombia Technical.*

— (2007b), "Fast Bayesian Inference for Gaussian Process Models," *Report 230, University of British Colombia Technical.*