Inference for Model Error

Allan Seheult*

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

June 25, 2009

Abstract

This report sketches Bayesian methods for (i) "testing" a full specification of the expectation and variance of "model error", regarded as a stochastic representation of the discrepancy between a computer model of a system and the system itself, and (ii) estimating a parameterised version of such a specification. The methods are illustrated on three published examples.

Keywords: Computer Models, Emulation, Galaxy Formation, Hydrocarbon Reservoirs, Model Error, Reification, Thermohaline Circulation.

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 and diagnostics for the inevitable discrepancy between the mathematical model and the physical system it purports to represent. We refer to this discrepancy as "model error" or "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 error. However, model error specification is challenging, drawing on expert knowledge of (a) the physical system and (b) any crucial simplifications assumed in the mathematical model; see, for example, Craig et al. (1998).

While we include some discussion on model error specification in particular examples, we mostly restrict attention to two situations: inference for model error when it has been partially specified and diagnostics when it has been fully specified. However, there is not always a clear distinction, and both may arise in a study. We develop methods for both situations and illustrate them on three reasonably substantive examples discussed elsewhere: (i) a model for thermohaline circulation in

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

the Atlantic Ocean proposed by Zickfeld et al. (2004) and used to illustrate the ideas developed in Goldstein and Rougier (2006) and Goldstein and Rougier (2009); (ii) a model for galaxy formation proposed by Bower et al. (2006) for which Goldstein and Vernon (2009) describe a careful specification exercise of model error 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 (iii) 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).

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 the relationship between observations on the physical system and the computer simulator using the notion of a "best input" to the simulator. We then see how we can combine the 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 error.

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 be the same 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

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

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 $-L''(\hat{\theta})$. 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 $0 \le \delta \le 1$; 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 Relating 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 an input x^* by the additive relationship

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

where ε, called the "model error" or "model discrepancy" and x^*, called the best input are such that $\varepsilon \perp\!\!\!\perp \{f, x^*\}$. 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 error. Moreover, the process which leads them to the reified model also acts as a basis for assessing the model error 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 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 *Diagnostics for model error*

In this section, we consider to what extent a full specification of the distribution of the model error ε, for example, Gaussian with specified expectation $\mathrm{E}[\varepsilon]$ (usually zero) and specified variance Σ_ε, is supported by the system observations z.

A natural approach is to compare the posterior distribution $p(\varepsilon \mid z, S, \Sigma_\varepsilon)$ with the completely specified prior distribution $p(\varepsilon \mid \Sigma_\varepsilon)$ for ε. In principle, we evaluate the posterior distribution as

$$p(\varepsilon \mid z, S, \Sigma_\varepsilon) = \int p(\varepsilon \mid x^*, z, S, \Sigma_\varepsilon)\, p(x^* \mid z, S, \Sigma_\varepsilon)\, dx^* \tag{12}$$

The first term in the integrand is Gaussian, but the second term, the posterior distribution of x^*—the "calibration distribution"—can be very difficult to compute, usually has "thick tails", and can be multi-model. Thus, the full posterior distribution is not readily available.

Another related approach is to simulate from the joint posterior distribution of ε and x^*

$$p(\varepsilon, x^* \mid z, S, \Sigma_\varepsilon) = p(\varepsilon \mid x^*, z, S, \Sigma_\varepsilon)\, p(x^* \mid z, S, \Sigma_\varepsilon) \tag{13}$$

by first simulating x^* from the calibration distribution and then simulating ε from the first distribution on the right-hand-side of (13). This leads to a sample-based estimate of the posterior distribution of ϵ which can be used to assess its prior specification. Again, this approach is limited by our ability to simulate exactly from the calibration distribution.

Two approximations to simulating from the calibration distribution have been explored. The first approximation is to simulate x^* from a convenient distribution with location and dispersion determined by the Bayes linear adjustment of the expectation and variance of x^* by z: see Goldstein and Rougier (2006). Similarly, the second approximation is to simulate from a convenient distribution with location and dispersion determined by the maximum likelihood estimate $\hat{x}^*$ of x^* (based on z) and the Hessian of the log-likelihood at $\hat{x}^*$; see, for example, Rougier (2009). We focus here on the second approximation. The first approximation will be included in a follow-up *technical report* in conjunction with other Bayes linear methods for model error estimation and diagnostics.

The first distribution $p(\varepsilon \mid x^*, z, S, \Sigma_\varepsilon)$ on the right-hand-side of (13) is Gaussian with expectation

$$\mathrm{E}[\varepsilon \mid x^*, z, S, \Sigma_\varepsilon] = \mathrm{E}[\varepsilon] + \mathrm{Cov}[\varepsilon,\, z]\mathrm{Var}[z]^{-1}(z - \mathrm{E}[z]) \tag{14}$$

and variance

$$\mathrm{Var}[\varepsilon \mid x^*, z, S, \Sigma_\varepsilon] = \mathrm{Var}[\varepsilon] - \mathrm{Cov}[\varepsilon,\, z]\mathrm{Var}[z]^{-1}\mathrm{Cov}[z,\, \varepsilon] \tag{15}$$

where the conditioning by x^*, S and the prior specification Σ_ε, common to all terms on the right-hand-sides of (14) and (15), has been suppressed. Straightforward

calculations show that these terms are given by $\mathrm{Cov}[\varepsilon,\, z] = \Sigma_\varepsilon H^{\mathrm{T}}$, $\mathrm{E}[z] = \mu(x^*)H$ and $\mathrm{Var}[z] = H^{\mathrm{T}}[\Sigma(x^*)+\Sigma_\varepsilon]H+\Sigma_e$, where $\mu(x)$ and $\Sigma(x)$ are the emulator mean and emulator variance at input x, respectively; and $\mathrm{E}[\varepsilon]$ and $\mathrm{Var}[\varepsilon] \equiv \Sigma_\varepsilon$ are specified.

The calibration distribution on the right-hand-side of (13) is such that

$$p\left(x^* \,|\, z, S, \Sigma_\varepsilon\right) \propto p\left(z \,|\, x^*, S, \Sigma_\varepsilon\right) p\left(x^*\right) \tag{16}$$

where $p\left(x^*\right)$ is the prior distribution for x^* and $p\left(z \,|\, x^*, S, \Sigma_\varepsilon\right)$, a Gaussian distribution with expectation $\mu(x^*)H$ and variance $H^{\mathrm{T}}[\Sigma(x^*) + \Sigma_\varepsilon]H + \Sigma_e$, is the likelihood for x^* given observations z; and, as above, $\mu(x)$ and $\Sigma(x)$ are the emulator mean and emulator variance at input x.

If we assume the prior distribution is "flat" where the likelihood is "concentrated", as is often the case, the calibration distribution is approximately proportional to the likelihood. Furthermore, we may choose to approximate the calibration distribution by a Gaussian distribution with expectation equal to the maximum likelihood estimate $\hat{x}^*$ and variance equal to the negative inverse of the Hessian matrix of the log-likelihood $L(x^*)$ evaluated at $\hat{x}^*$. However, as we expect the calibration distribution to have "thick tails" we examine robustness of diagnostic indications using multivariate t-distributions with different degrees-of-freedom and the same location and scale as for the Gaussian approximation; see Rougier (2009) for details. While these approximations do not account for any multi-modality, we believe that the main features of diagnostics will not be affected to any great extent.

Diagnostics: These are based on simulation from the approximate posterior distribution $p\left(\varepsilon \,|\, z\right)$ of ε given z, where S and the specification Σ_ε has been suppressed in the conditioning. There are many choices, but obvious diagnostics are the posterior distribution of the quadratic form

$$Q = \left(\varepsilon - \mathrm{E}[\varepsilon]\right) \mathrm{Var}[\varepsilon \,|\, z]^{-1} \left(\varepsilon - \mathrm{E}[\varepsilon]\right)^{\mathrm{T}} \tag{17}$$

and the variance "ratio"

$$\mathrm{Var}[\varepsilon]^{-1}\mathrm{Var}[\varepsilon \,|\, z] \tag{18}$$

The first diagnostic, a Mahalanobis-type distance, provides a summary check on the prior specification of $\mathrm{E}[\varepsilon]$, typically zero. The specification will be judged acceptable if the distribution is close to a chi-squared distribution with k degrees-of-freedom. Thus, a crude check is to compute the tail probability for the posterior expectation of Q in (17), using a chi-squared distribution with k degrees-of-freedom. When the diagnostic indicates against the prior specification of $\mathrm{E}[\varepsilon]$, we may inspect the posterior distributions of the components of ε to examine the contra-indications. Alternatively, denoting the Choleski decomposition of $\mathrm{Var}[\varepsilon \,|\, z]$ by R, we can examine the posterior distributions of the components of $[\varepsilon - \mathrm{E}[\varepsilon]]R^{-1}$, which are uncorrelated, unit variance, random quantities which have zero expectation when $\mathrm{E}[\varepsilon]$ has been "correctly" specified.

The second diagnostic in (18) indicates the extent to which we learn about $\mathrm{Var}[\varepsilon]$ from the systems observation z. If $\mathrm{Var}[\varepsilon]^{-1}\mathrm{Var}[\varepsilon \,|\, z]$ is "close" to the $k \times k$ identity, the eigen-values of $\mathrm{Var}[\varepsilon]^{-1}\mathrm{Var}[\varepsilon \,|\, z]$ will be "close" to unity and we learn very little about $\mathrm{Var}[\varepsilon]$ from z. On the other hand, when the variance ratio differs from the

identity, small eigen-values point to those linear combinations of the components of ε about which the data z has been most informative. We illustrate application of both diagnostics to the examples in Section 6.

5 *Inference for model error*

In this section, we consider how to estimate a parameterised specification of the variance $\Sigma_\varepsilon(\varphi)$ of model error 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

$$l(\varphi) \propto \int p\left(z \mid S, \varphi, x^*\right) p\left(x^*\right) dx^* \tag{19}$$

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

$$\mathrm{E}[z \mid S, \varphi, x^*] = \mathrm{E}[z \mid S, x^*] = \mu(x^*)H \tag{20}$$

and

$$\mathrm{Var}[z \mid S, \varphi, x^*] = H^{\mathrm{T}}[\Sigma(x^*) + \Sigma_\varepsilon(\varphi)]H + \Sigma_e \tag{21}$$

where, as before, $\mu(x)$ 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 (19), 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 two of the examples in Section 6. If there is information about φ which can be expressed in the form of a prior distribution, its posterior distribution can be computed in the usual way and used to infer about φ.

6 *Examples*

6.1 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'. Bottom hole pressure at six onshore producing wells at various times 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 7 permeability multipliers (range $[0.1, 10.0]$) for the 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. The inputs were 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 7 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, 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 the '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

$$\mathrm{Cov}[\varepsilon_k, \varepsilon_\ell] = \sigma_1^2 \exp\left(-\theta_1 (T_k - T_\ell)^2\right) + \sigma_2^2 \exp\left(-\theta_2 (T_k - T_\ell)^2\right) I_{W_k = W_\ell} \qquad (22)$$

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.

6.1.1 Diagnostics for the example

The REML estimates of the correlation lengths are $\theta = (9.1, 1.4, 0.62, 4.9)$ and the nugget multiplier was chosen to be $\delta = 0.01$. The regression terms used for the emulator were $x_2, x_3, x_4, x_6, x_3^2, x_4^2, x_3x_4, x_4x_6$, which correspond to the maximal model used by Craig et al. (2001), who use different regressions for each output.

The maximum likelihood estimate of x^* and Hessian based variance matrix are $\hat{x}^* = (0.157, 0.014, -0.052, 0.119)$ and

```
 0.0263 -0.0023  0.0056 -0.0141
-0.0023  0.0078 -0.0010  0.0019
 0.0056 -0.0010  0.0060 -0.0046
-0.0141  0.0019 -0.0046  0.0225
```

Recall that Σ_ε is specified as

$$\mathrm{Cov}[\varepsilon_k, \varepsilon_\ell] = \sigma_1^2 \exp\left(-\theta_1(T_k - T_\ell)^2\right) + \sigma_2^2 \exp\left(-\theta_2(T_k - T_\ell)^2\right) I_{W_k = W_\ell} \qquad (23)$$

where $\sigma_1^2 = 25$, $\theta_1 = (6 \times 10^{-4})^2$, $\sigma_2^2 = 6$ and $\theta_2 = (2 \times 10^{-3})^2$.

To investigate this specification, 5000 simulations from the joint posterior distribution for ε and x^* in (13), using the Gaussian approximation to the calibration distribution, based on $\hat{x}^*$ and its Hessian based variance, gave a highly significant Mahalanobis P-value (1.386×10^{-6}), suggesting that the specification $\mathrm{E}[\varepsilon] = 0$ is in serious doubt. This doubt is further supported by the following mostly large negative ratios of the posterior expectations of ε to their posterior standard deviations.

```
-1.79 -2.30 -2.49 -2.89 -2.97 -3.50 -3.01 -3.70 -3.12 -3.54 -3.00
-3.88 -3.22 -3.67 -2.74 -4.08 -3.67 -3.43 -2.22 -3.88 -3.09 -2.38
-2.75 -1.84 -3.10 -2.04 -2.63 -1.80  0.83 -0.50 -0.03 -0.35  1.29
-0.20
```

The estimated posterior density of the model error for bottom hole pressure at time 5326 for well 38, for example, corresponding to the ratio of -4.08 above, depicted in Fig. 6.1.1, indicates very clearly how its prior distribution has been modified by the data.

Additionally, the estimated posterior expectations of the unit variance, orthogonal components εR^{-1}, where R is the Choleski decomposition of Σ_ε, shown below, indicate several "significant" directions.

```
-1.72 -1.49 -1.27 -1.17 -1.32 -1.84 -0.35 -1.26 -0.31 -0.62 -0.01
-0.75 -0.16 -0.46 -0.13 -1.69 -0.18 -0.03  0.67 -1.51  0.87  1.09
 0.04  0.58 -0.63  0.36  0.03  0.89  3.17 -0.25  0.17  0.27  2.50
-0.30
```

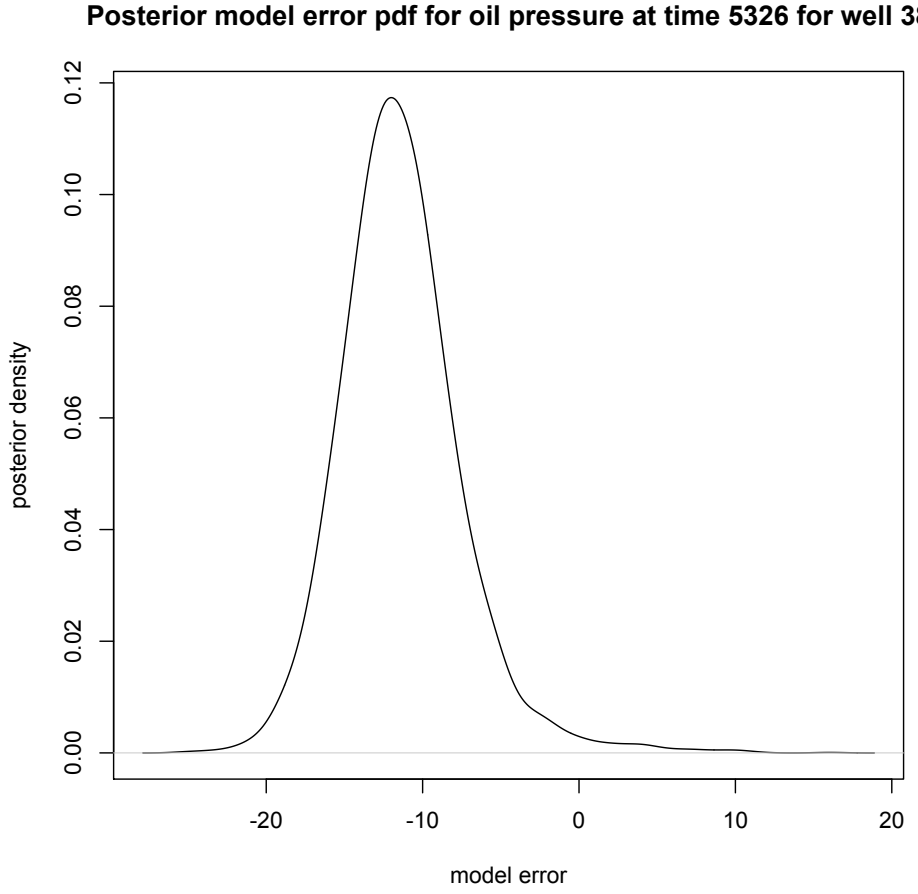

Figure 1: Simulated posterior distribution of model error component ε_{16} for the reseervoir data.

The following ratios of posterior standard deviations of ε to their prior standard deviations indicate we have learnt a lot about Σ_ε.

```
0.56 0.53 0.61 0.55 0.63 0.54 0.67 0.59 0.64 0.56 0.67 0.62 0.57 0.55
0.58 0.52 0.59 0.53 0.56 0.50 0.58 0.56 0.56 0.59 0.54 0.60 0.58 0.58
0.70 0.66 0.69 0.64 0.68 0.68
```

The small eigen-values of $\mathrm{Var}[\varepsilon]^{-1}\mathrm{Var}[\varepsilon \mid z]$ shown below are a further indication that we have learnt a lot about certain linear combinations of the components of ε.

```
1.08 0.99 0.97 0.96 0.93 0.92 0.91 0.89 0.86 0.85 0.83 0.76 0.74 0.72
0.70 0.70 0.68 0.67 0.66 0.64 0.63 0.60 0.58 0.58 0.57 0.55 0.54 0.52
0.51 0.46 0.42 0.32 0.15 0.06
```

For comparison, 5000 simulations from the joint posterior distribution for ε and x^*, using a multivariate t-dstribution with 3 degrees-of-freedom as a thicker tailed approximation to the calibration distribution, gave a less significant Mahalanobis P-value (4.644×10^{-5}). Furthermore, the following corresponding ratios of posterior expectations of ε to their posterior standard deviations, ratios of posterior standard deviations of ε to their prior standard deviations, and the eigen-values of $\mathrm{Var}[\varepsilon]^{-1}\mathrm{Var}[\varepsilon \mid z]$ are all similar to those with the Gaussian approximation, but more conservative, as we might expect.

```
-1.51 -1.91 -1.81 -2.18 -2.12 -2.62 -2.05 -2.58 -2.19 -2.56 -2.03
-2.68 -2.18 -2.45 -1.92 -2.94 -2.46 -2.28 -1.51 -2.74 -1.98 -1.59
-1.87 -1.24 -2.18 -1.37 -1.72 -1.22  0.69 -0.29  0.08 -0.16  1.03
-0.05
```

```
0.63 0.64 0.78 0.70 0.85 0.70 0.93 0.81 0.87 0.74 0.96 0.86 0.81 0.80
0.78 0.70 0.84 0.76 0.78 0.67 0.84 0.79 0.78 0.80 0.72 0.82 0.83 0.80
0.95 0.91 0.95 0.89 0.96 0.94
```

```
1.99 1.10 1.01 0.98 0.95 0.94 0.92 0.90 0.90 0.88 0.87 0.80 0.77 0.75
0.73 0.72 0.70 0.69 0.66 0.65 0.63 0.62 0.62 0.61 0.60 0.58 0.56 0.54
0.54 0.50 0.49 0.32 0.17 0.09
```

This example benefits from having as many as 34 observation z compared with just the 4 components of φ, which is not the case in the next two examples.

6.1.2 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.

Table 1: Simulator inputs and ranges

x_1	$\tau_1 = T_2^*$	°C	0	10
x_2	$\tau_2 = T_1^* - T_2^*$	°C	0	5
x_3	$\tau_3 = T_3^* - T_1^*$	°C	0	10
x_4	Γ	$\mathrm{W\,m^{-2}\,{}^\circ C^{-1}}$	10	70
x_5	K	Sv	5,000	100,000

While diagnostics for Σ_ε evaluated at the maximum likelihood estimate of φ appear to be a little more supportive than those for the initial specification, it is the structure of the specification of $\mathrm{Var}[\varepsilon]$ and $\mathrm{E}[\varepsilon]$ which seems to be in doubt, possibly by not including actual distance between wells.

6.2 Thermohaline circulation

System: Thermohaline circulation in the Atlantic Ocean [THC] is the mechanism by which heat is drawn up from the tropics towards the western seaboard of Europe. There is concern about the effect of global warming on THC, because changing temperature and precipitation patterns will alter temperature and salinity. One important quantity to predict is the amount of freshwater re-distribution in the Atlantic that would cause THC shutdown (which could significantly lower the temperature of the western seaboard of Europe), using temperature and salinity of the Atlantic, and current THC size.

Model: We use a model of Zickfeld et al. (2004) [*ZSR*], as implemented by Goldstein and Rougier (2006) *GR* and Goldstein and Rougier (2009) to illustrate their ideas. The *ZSR* model of the Atlantic is a four-compartment system, where each compartment is described by its volume and its depth. The equilibrium state vector comprises a temperature T_i and salinity S_i for each compartment Freshwater re-distribution is modelled by two parameters F_1 and F_2, and atmospheric temperature forcing by the three parameters T_1^*, T_2^* and T_3^*. The key quantity is the rate of meridional overturning m, the water flow-rate through the compartments, a proxy for THC which tends to be bigger when the 'northern Atlantic' is colder and more salty than the 'southern Atlantic'. Large values of F_1 make the 'south Atlantic' more salty, and tend to reduce m.

Inputs: Five of the parameters of the *ZSR* model T_1^*, T_2^*, T_3^*, Γ and K were treated as uncertain and the others fixed at nominal values. *GR* reparameterise the three temperatures as $\tau_1 = T_2^*$, $\tau_2 = T_1^* - T_2^*$ and $\tau_3 = T_3^* - T_1^*$, leading to simulator inputs $x = (\tau_1, \tau_2, \tau_3, \Gamma, K)$ with ranges given in Table 6.2. The inputs were linearly transformed to vary over $[0, 1]$.

Outputs: The 7 components of the simulator $f(x)$ are

$$f_i(x) = \begin{cases} T_i(x) & i = 1, 2, 3 \\ \Delta S_{21} = S_2(x) - S_1(x) & i = 4 \\ \Delta S_{32} = S_3(x) - S_2(x) & i = 5 \\ m(x) & i = 6 \\ F_1^{\text{crit}}(x) & i = 7 \end{cases} \tag{24}$$

GR follow *ZSR*, who calibrate their model to data from the CLIMBER-2 intermediate complexity coupled ocean/atmosphere simulator. This simulator provides values for the three equilibrium temperatures $T_1 = 6$, $T_2 = 4.7$ and $T_3 = 11.4$, the salinity differences $S_2 - S_1 = -0.15$ and $S_3 - S_2 = 0.25$, and the equilibrium overturning $m = 22.6$, which together comprise the six components of z, one less than for the simulator, namely, $F_1^{\text{crit}}(x)$, so that H is a 6×7 matrix, not the usual identity matrix.

Design: 30 evaluations of the *ZSR* model were made in a Latin hypercube design across the 5-dimensional input space. Thus, X is 30×5 and F is 30×7.

Model Discrepancy: *GR* in Goldstein and Rougier (2009) set the expectation of ε to zero, and for Σ_ε choose a mostly-diagonal matrix with individual component standard deviations 0.84, 0.84, 0.84, 0.075, 3.30, 0.044 for $T_1, T_2, T_3, \Delta S_{21}, \Delta S_{32}, m, F_1^{\text{crit}}$, respectively. They also included a correlation of -0.5 between the two salinity differences in ε (components four and five) to account for the shared salinity term in compartment 2. The overall specification results from their judgements based on a careful reification exercise concerning the relationship between three emulators and the system. However, no climate experts were involved in the specifications.

Measurement Error: As z is the output of the simulator CLIMBER-2 there is no measurement error; that is, $e = 0$, exactly.

6.2.1 Diagnostics for the example

The REML estimate of the single correlation length is $\theta = 0.788$, and the regression terms for the emulator were chosen to be x_1, x_2, x_3, x_4, x_5, in accord with Goldstein and Rougier (2009).

The maximum likelihood estimate of x^* and the Hessian based variance are $\hat{x}^* = (0.351, 0.554, 0.518, 0.458, 0.916)$ and

```
 0.0048 -0.0026 -0.0039  0.0044 -0.0064
-0.0026  0.0069 -0.0001 -0.0071  0.0027
-0.0039 -0.0001  0.0111  0.0001  0.0089
 0.0044 -0.0071  0.0001  0.0301 -0.0129
-0.0064  0.0027  0.0089 -0.0129  0.0464
```

Recall that Σ_ε is specified to have standard deviations 0.84, 0.84, 0.84, 0.075, 3.30, 0.044 and a single correlation of -0.5 between two salinity differences.

To investigate this specification, 5000 simulations from the joint posterior distribution for ε and x^* in (13), using the Gaussian approximation to the calibration

distribution, based on $\hat{x}^*$ and its Hessian based variance, gave a non-significant Mahalanobis P-value (0.3039), suggesting that the specification $\mathrm{E}[\varepsilon] = 0$ is not in doubt. This suggestion is further supported by the following small ratios of the posterior expectations of ε to their posterior standard deviations.

```
-0.26 0.01 0.12 -0.85 0.62 0.47 0.01
```

The following ratios of posterior standard deviations of ε to their prior standard deviations are an indication that little has been learnt about the specification of Σ_ε.

```
0.74 0.67 0.98 0.85 0.84 1.00 1.01
```

The single small eigen-value of $\mathrm{Var}[\varepsilon]^{-1}\mathrm{Var}[\varepsilon \mid z]$ below is a further indication that we have learnt mostly about only one linear combination of the components of ε.

```
1.05 1.05 0.97 0.86 0.77 0.68 0.07
```

For comparison, 5000 simulations from the joint posterior distribution for ε and x^*, using a multivariate t-dstribution with 3 degrees-of-freedom as a thicker tailed approximation to the calibration distribution, gave a similar Mahalanobis P-value (0.3259). The following corresponding ratios of posterior expectations of ε to their posterior standard deviations, ratios of posterior standard deviations of ε to their prior standard deviations, and the eigen-values of $\mathrm{Var}[\varepsilon]^{-1}\mathrm{Var}[\varepsilon \mid z]$ all give a conservative confirmation of the indications from using the Gaussian approximation to $p(x^* \mid z)$; namely, that we have a learnt mostly about only one linear combination of the components of ε.

```
-0.17 -0.03 0.05 -0.76 0.54 0.49 0.03

 1.32 1.00 1.68 0.89 0.88 1.17 1.00

 3.09 2.28 1.38 1.01 0.83 0.75 0.19
```

6.3 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 the model Galform which models 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 17 input parameters that cosmologists were interested in varying. Expert judgements on the impact of these inputs on the luminosity functions lead Goldstein and Vernon (2009) [GV] to vary only 8 of them, while taking into account the possible effects of the remaining 9. The 8 input parameters and their initial ranges are

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

The inputs were linearly transformed to vary over $[-1, 1]$.

Outputs: GV 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. GV compare the average over 40 of the 512 sub-volumes of the logarithms of 11 representative outputs (6 from the bj band and 5 from the K band) to the corresponding observational data z from the 2dFGRS galaxy survey, where

$$z = -(1.70, 1.94, 2.43, 3.31, 4.56, 5.42, 2.06, 2.22, 2.38, 3.46, 5.04) \tag{25}$$

Design: 1000 evaluations of Galform were made in a Latin hypercube design across the 8-dimensional input space. Seven evaluations were "burnt" (the model failed to compute), so that X is 993×8 and F is 993×11.

Model Discrepancy: A leading cosmologist's opinion is that there is no overall bias of the model, so that $\mathrm{E}[\varepsilon] = 0$. 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 11 luminosity outputs all being too high (or too low), suggesting positive correlation between all 11 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, $\mathrm{Var}[\varepsilon]$ was parameterised as

$$\Sigma_\varepsilon(a, b, c) = a \begin{bmatrix} 1 & b & .. & c & .. & c \\ b & 1 & .. & c & . & c \\ : & : & : & : & : & : \\ c & .. & c & 1 & b & .. \\ c & .. & c & b & 1 & .. \\ : & : & : & : & : & : \end{bmatrix} \tag{26}$$

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 $\mathrm{E}[e] = 0$ and known $\mathrm{Var}[e] = \Sigma_e$.

6.3.1 Diagnostics for the example

The REML estimates of the correlation lengths are $(.87, 1.2, .81, 1.3, .93, .86, 1.7, 2.0)$. The regression terms for the emulator were chosen to be $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$.

The maximum likelihood estimate of x^* and Hessian based variance are $\hat{x}^* = (0.928, 0.241, -0.214, 0.100, 0.100, -0.004, 0.501, -0.506)$ and

```
 0.028  0.023  0.000 -0.011  0.000  0.000 -0.021 -0.026
 0.023  0.116  0.010  0.014  0.028 -0.005  0.016 -0.140
 0.000  0.010  0.026 -0.043 -0.007  0.002  0.010  0.019
-0.011  0.014 -0.043  0.171  0.044 -0.054  0.001 -0.053
 0.000  0.028 -0.007  0.044  0.063 -0.034  0.017 -0.045
 0.000 -0.005  0.002 -0.054 -0.034  0.059  0.000  0.005
-0.021  0.016  0.010  0.001  0.017  0.000  0.074  0.009
-0.026 -0.140  0.019 -0.053 -0.045  0.005  0.009  0.318
```

Recall that $\mathrm{Var}[\varepsilon]$ is specified as

$$\Sigma_\varepsilon = a \begin{bmatrix} 1 & b & .. & c & .. & c \\ b & 1 & .. & c & . & c \\ : & : & : & : & : & : \\ c & .. & c & 1 & b & .. \\ c & .. & c & b & 1 & .. \\ : & : & : & : & : & : \end{bmatrix} \tag{27}$$

where a is a variance and b and c are correlations. The cosmologist's initial specifications were $a = 5.66 \times 10^{-3}$, $b = 0.8$ and $c = 0.6$.

To investigate this initial specification, 5000 simulations from the joint posterior distribution for ε and x^* in (13), using the Gaussian approximation to the calibration distribution, based on $\hat{x}^*$ and its Hessian based variance, gave a non-significant Mahalanobis P-value (0.2876), suggesting that the specification $\mathrm{E}[\varepsilon] = 0$ is not in doubt. This suggestion is supported further by the following small ratios of the posterior expectations of ε to their posterior standard deviations.

```
0.12  0.32 -0.20  0.21 -0.01 -0.06 -0.23 -0.28  0.18  -0.20 -0.12
```

The following ratios of posterior standard deviations of ε to their prior standard deviations are an indication that almost nothing has been learnt about the specification of Σ_ε.

```
0.89 0.85 0.91 0.91 0.93 0.93 0.96 0.95 0.97 0.97 0.97
```

The following eigen-values of $\mathrm{Var}[\varepsilon]^{-1}\mathrm{Var}[\varepsilon \mid z]$ are a further indication that very little has been learnt about the specification of Σ_ε.

```
1.03 1.00 1.00 0.98 0.95 0.94 0.90 0.80 0.78 0.71 0.45
```

For comparison, 5000 simulations from the joint posterior distribution for ε and x^*, using a multivariate t-dstribution with 3 degrees-of-freedom as a thicker tailed approximation to the calibration distribution, gave a similar Mahalanobis P-value (0.3133); and the ratios of posterior expectations of ε to their posterior standard deviations, ratios of posterior standard deviations of ε to their prior standard deviations, and the eigen-values of $\mathrm{Var}[\varepsilon]^{-1}\mathrm{Var}[\varepsilon \mid z]$ all agree with the indications using the Gaussian approximation to $p(x^* \mid z)$.

```
0.08 0.26 -0.16 0.17 -0.01 -0.06 -0.19 -0.26 0.17 -0.17 -0.09

0.961 0.938 0.981 0.994 0.989 0.977 0.981 0.970 1.003 0.985 0.996

1.053 1.010 0.990 0.986 0.982 0.958 0.903 0.824 0.812 0.761 0.609
```

6.3.2 Inference for the example

We re-parameterised Σ_ε by setting $\varphi = (0.5\log(a), \mathrm{arctanh}(b), \mathrm{arctanh}(c))$. Maximum likelihood estimates of (a, b, c) are $(0.02, 1, 0.9)$, but these and their standard errors are unreliable. However, a more detailed inspection of the likelihood surface does suggest larger values for all three parameters than the upper limits suggested by the cosmologist. Hence, restricting the likelihood to the cosmologist's limits, gives estimates $(0.00566, 0.8, 0.8)$ for (a, b, c). Further simulations with these values had little effect on the diagnostics.

If we fix the correlations b and c at their initial specified values of $b = 0.8$ and $c = 0.6$, the maximum likelihood estimate of the variance a is 0.04, which is order of magnitude larger than its initial specified value of $a = 0.00566$; in fact, $a = 0.00566$ is outside a likelihood based 95% interval for a: see Fig. 6.3.2.

7 Discussion

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

The model error, also referred to as model discrepancy, must 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. The distribution of ε is sometimes completely specified or $\mathrm{Var}[\varepsilon]$ is specified partially to be a function of parameters φ; and in both cases, $\mathrm{E}[\varepsilon]$ is usually specified to be zero.

In the completely specified case, observed data z on the real system, in combination with runs on the simulator of the system and the initial specification for model error are used to learn further about ε. In this report, the learning process is based on approximations to the posterior distribution $p(\varepsilon \mid z)$ based on fully specified distributions for all random quantities. We examine the posterior distribution of ε to see to what extent our initial beliefs are changed by z.

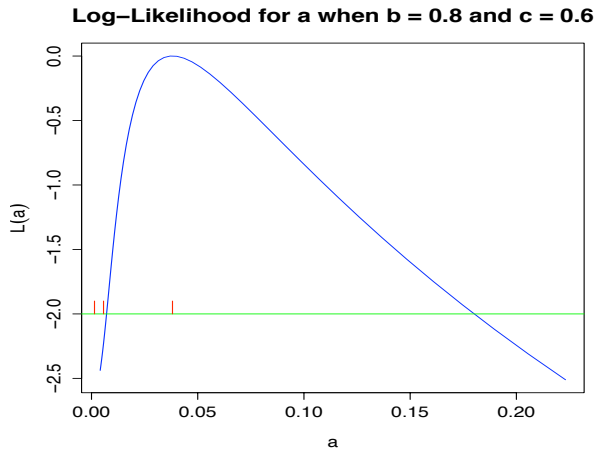

Figure 2: Likelihood for the variance a when the correlations b and c are set at their initial specifications 0.8 and 0.6.

In the case where $\mathrm{Var}[\varepsilon]$ is specified to be a function of parameters φ, we use a likelihood analysis to learn about φ.

The methods are illustrated on three reasonably substantial examples.

An important issue concerns subsequent analysis, such as calibration and prediction. Should we replace the prior specification of ε by its posterior specification based on the posterior distribution $p(\varepsilon \mid z)$ in the completely specified case, or based on the posterior distribution of φ (possibly approximated by its likelihood) in the partially specified case? If we do this, then we double-count the system observations z. The "correct" form of analysis analysis should proceed as follows.

In the parameterised case, it is straightforward to show that the predictive distribution for y given z can be calculated as:

$$p(y \mid z) = \int p(y \mid z, \varphi)\, p(\varphi \mid z)\, d\varphi \tag{28}$$

where, in the case that a prior distribution for φ is not specified, the posterior distribution $p(\varphi \mid z)$ may be approximated by its likelihood when the number of observations z is large. However, the first term in the integrand is the usual predictive distribution, but a separate one for each φ, so that the integral is essentially intractable. Similarly for calibration, where the calibration distribution for x^* given z can be calculated as

$$p(x^* \mid z) = \int p(x^* \mid z, \varphi)\, p(\varphi \mid z)\, d\varphi \tag{29}$$

but again the integral is practically intractable.

When the prior distribution of φ is completely specified, the predictive distribution can be calculated as

$$p(y \mid z) = \int p(y \mid z, \varepsilon)\, p(\varepsilon \mid z)\, d\varepsilon \tag{30}$$

Both terms in the integrand are too complicated for the integral to be calculated. However, the integral can be evaluated if we were to approximate each term in the integrand by a Gaussian distributions. The calibration distribution is similarly complicated.

Fortunately, we can side-step these integrations by simulation similar to that used to simulate (approximately) from the joint posterior distribution of x^* and ε. These ideas will be explored and implemented in a subsequent account.

Acknowledgements

I would like to thank Peter Craig for early discussions on REML, Jonathan Cumming for helpful discussions on many aspects of computer models and patiently answering my questions about `R`, Michael Goldstein for invaluable discussions and ideas on model error and not discouraging exploration of fully Bayesian methods, Jonathan Rougier for helpful discussions and access to his "hat run" R code, and job-share partner Ian Vernon for many insightful, illuminating discussions, and for providing Galform data on which to explore my ideas.

This report was produced with the support of the Basic Technology initiative as part of the "Managing Uncertainty for Complex Models" project.

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," To be published.

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.

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," *University of British Colombia Technical Report 229*.

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

Rougier, J. (2009), "Formal Bayes methods for model calibration with uncertainty," in *Applied Uncertainty Analysis for Flood Risk Management*, eds. Beven, K. and Hall, J., Imperial College Press / World Scientific.

Zickfeld, K., Slawig, T., and Rahmstorf, S. (2004), "A low-order model for the response of the Atlantic thermohaline circulation to climate change," *Ocean Dynamics*, 54, 8–26.