Computer Model Calibration or Tuning in Practice

Jason L. Loeppky
Department of Statistics
University of British Columbia
Vancouver, BC, V6T 1Z2, CANADA
(jason@stat.ubc.ca)

Derek Bingham
Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC, V5A 1S6, CANADA
(dbingham@stat.sfu.cc)

William J. Welch
Department of Statistics
University of British Columbia
Vancouver, BC, V6T 1Z2, CANADA
(will@stat.ubc.ca)

May 5, 2006

Abstract

Computer models to simulate physical phenomena are now widely available in engineering and science. Before relying on a computer model, a natural first step is often to compare its output with physical or field data, to assess whether the computer model reliably represents the real world. Field data, when available, can also be used to calibrate or tune unknown parameters in the computer model.

Calibration is particularly problematic in the presence of systematic discrepancies between the computer model and field observations. We introduce a likelihood alternative to previous Bayesian methodology for estimation of calibration or tuning parameters. In an important special case, we show that maximum likelihood estimation will asymptotically find values of the calibration parameter that give an unbiased computer model, if such a model exists. However, the calibration parameters are not necessarily estimable. We also show in some settings that calibration or tuning need to take into account the end-use prediction strategy. Depending on the strategy, the choice of the parameter values may be critical or unimportant.

KEYWORDS: Computer experiment, Gaussian stochastic process, Maximum likelihood, Random function.

1. INTRODUCTION

In many engineering and science disciplines, deterministic computer models or codes are used to simulate complex physical processes. For example, Sacks, Schiller and Welch (1989), Sacks, Welch, Mitchell and Wynn (1989), and Currin, Mitchell, Morris and Ylivisaker (1991) described applications in chemical kinetics, electrical engineering, and the physics of heat storage.

Ideally, a deterministic computer model is a perfect mathematical representation of the physical system. If so, at every setting of the input (explanatory) variables used in the field, the mean of the field response and the computer model output agree exactly. In addition, one or more of the input variables to the computer code may be unknown or unmeasurable in the physical system. Thus, the output values from limited computer runs are compared to the field response measurements to adjust these input variables. The objective of adjustment is either (i) to estimate the true values of the corresponding physical parameters, the *calibration* problem or (ii) to *tune* adjustment parameters of the computer code without physical meaning to be a better surrogate for the physical process. An underlying objective of the analysis of such data is to build an accurate predictor of the field mean, so avoiding the need for field response measurements in new settings.

Kennedy and O'Hagan (2001) introduced a Bayesian model incorporating both the computer-model and field data. Their formulation also recognized that the computer model will inevitably be biased to some extent. As pointed out, by discussants of Kennedy and O'Hagan (2001), in the presence of bias it is questionable whether calibration is possible in the sense of estimating physically meaningful constants. This practical issue is a major thrust of our paper.

To illustrate the practical complexities, an application described in greater detail in Section 4 concerns a resistance spot welding process (Bayarri et al. 2002, 2005 and Higdon et al. 2004). There are three input variables to the computer code that may also be set for field measurements. In addition, the computer code requires an input value for a parameter, τ, which is not a variable in the field. Following the statistical model and maximum likelihood approach of Section 2 leads to a profile likelihood for τ, as shown in Figure 1. The profile likelihood is seen to be highly multi-modal, which leaves the practitioner with many questions regarding the statistical model and the appropriate choice of τ. This finding is consistent with the Bayesian marginal posterior distribution for τ given by Higdon et al. (2004) for the same application. We argue in Section 6, however, that in the most common setting for end-use prediction based on such a statistical model,

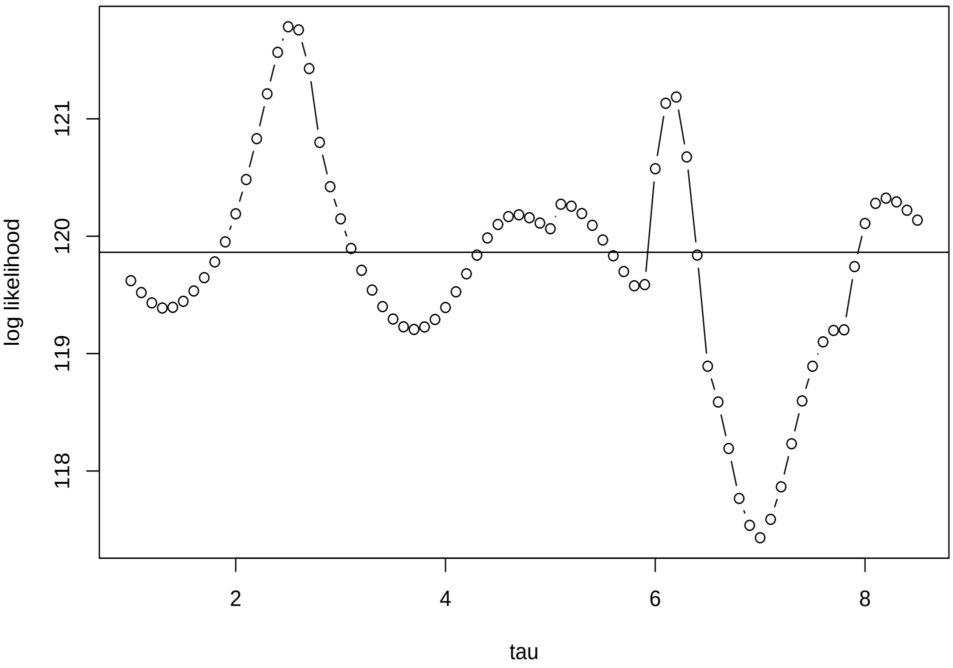

Figure 1: Profile Likelihood for τ in the Spot Weld Example. An approximate 95% confidence interval for τ is given by values with a likelihood above the horizontal line.

agonizing over the choice of τ is largely irrelevant to prediction accuracy.

This paper is outlined as follows. In Section 2 we introduce the standard model for combining data from the field and the computer model (see Bayarri et al. 2002, 2005; Higdon et al. 2004; Kennedy and O'Hagan, 2001) and develop a likelihood approach to estimate the parameters in the model. In Section 3, we discuss the identifiability of the model in Section 2 and show via a theoretical result that maximum likelihood estimation will asymptotically find values of the calibration parameter that give an unbiased computer model, if such a model exists. Section 4 presents detailed analysis of two examples. Sections 5 and 6 focus on practical issues related to the implementation of the model and formally discuss three end-use prediction strategies. Finally we will end with a brief discussion in Section 7.

2. STATISTICAL MODEL

To build a joint model for both sources of data, the formulation given by Kennedy and O'Hagan (2001) is adopted. The computer code has two types of input variables: the

d variables $\boldsymbol{x} = (x_1, \ldots, x_d)$ can also be manipulated and measured in the field, whereas the q further variables $\boldsymbol{t} = (t_1, \ldots, t_q)$ relate to calibration or tuning and cannot be varied or observed in the field. The function coded in the computer model to generate output is denoted by $\kappa(\boldsymbol{x}, \boldsymbol{t})$.

For the model generating the field observations in the real world, it is helpful conceptually to distinguish calibration from tuning. For calibration, the variables $\boldsymbol{t} = (t_1, \ldots, t_q)$ in the computer model correspond to q parameters, $\boldsymbol{\tau}$, in the real world. These parameters denote the unknown values of physical constants, for example rate constants in a chemical kinetics system. The parameters $\boldsymbol{\tau}$ define the physical process; they cannot be adjusted in practice. Field observations on a scalar response variable are assumed to be generated by

$$Y_F(\boldsymbol{x}) = \phi(\boldsymbol{x}|\boldsymbol{\tau}) + \varepsilon, \tag{1}$$

where $\phi(\boldsymbol{x}|\boldsymbol{\tau})$ describes how the field response mean changes with $\boldsymbol{x}$, conditional on $\boldsymbol{\tau}$, and ε is white-noise measurement error with distribution $N(0, \sigma_\varepsilon^2)$. In contrast, when tuning, $\boldsymbol{t}$ is a set of variables to adjust the computer model to more closely represent the physical system in some way, but with no analogous $\boldsymbol{\tau}$ in the real world. Thus, $\boldsymbol{\tau}$ can be removed from (1). Alternatively, we might think of $\boldsymbol{\tau}$ as the best set of values of $\boldsymbol{t}$, (i.e., those that make $\kappa(\boldsymbol{x}, \boldsymbol{\tau})$ and $\phi(\boldsymbol{x})$ close in some sense). Even in the case of calibration, $\boldsymbol{\tau}$ is fixed, and hence $Y_F(\boldsymbol{x})$ is not written as a function of $\boldsymbol{\tau}$. Mathematically, then, all these situations are indistinguishable. In Section 3, when we discuss the identifiability of $\boldsymbol{\tau}$ in calibration, it turns out that including $\boldsymbol{\tau}$ in (1) is helpful. This we write $\phi(\boldsymbol{x}|\boldsymbol{\tau})$ throughout as a convention.

Kennedy and O'Hagan (2001) formulated a model combining the computer-model and field data:

$$\begin{aligned} y_C(\boldsymbol{x}, \boldsymbol{t}) &= \kappa(\boldsymbol{x}, \boldsymbol{t}) & (2) \\ Y_F(\boldsymbol{x}) &= \kappa(\boldsymbol{x}, \boldsymbol{\tau}) + \delta(\boldsymbol{x}) + \varepsilon. & (3) \end{aligned}$$

Comparing (1) and (3), we see that the mean of the field response, $\phi(\boldsymbol{x}|\boldsymbol{\tau})$, is modeled as the sum of $\kappa(\boldsymbol{x}, \boldsymbol{\tau})$ and a discrepancy term, $\delta(\boldsymbol{x})$. Typically, $\delta(\boldsymbol{x})$ is attributed to deficiencies in the computer model, though bias in the field measurement system cannot be distinguished. Note that the field-data model (3) is in terms of the computer code with $\boldsymbol{t}$ hypothetically set to $\boldsymbol{\tau}$. As $\boldsymbol{t} = \boldsymbol{\tau}$ is fixed in $\kappa(\boldsymbol{x}, \boldsymbol{\tau})$, the discrepancy, $\delta(\boldsymbol{x})$, is a function of $\boldsymbol{x}$ only. Kennedy and O'Hagan (2001) also included a scale parameter multiplying $\kappa(\boldsymbol{x}, \boldsymbol{\tau})$ in (3), but this appears to be unimportant for the applications we have encountered. The statistical model in (2) for the computer runs allows arbitrary values of $\boldsymbol{t}$ according to some experimental plan.

We prefer a variation on this formulation where the mean field response is expressed in terms of the computer model with $\boldsymbol{t}$ set to a "chosen" value, $\boldsymbol{t}_0$, not necessarily equal to $\boldsymbol{\tau}$. First, as we see from Figure 1, $\boldsymbol{\tau}$ might not be estimated well. Secondly, as we shall illustrate in Section 6, different choices for $\boldsymbol{t}_0$ may be appropriate for prediction, depending on exactly how the statistical model and future computer model runs are used. For these reasons, we will use the combined model,

$$\begin{aligned} y_C(\boldsymbol{x}, \boldsymbol{t}) &= \kappa(\boldsymbol{x}, \boldsymbol{t}) && (4) \\ Y_F(\boldsymbol{x}) &= \kappa(\boldsymbol{x}, \boldsymbol{t}_0) + \delta_{\boldsymbol{t}_0}(\boldsymbol{x}) + \varepsilon. && (5) \end{aligned}$$

The field mean in (5) is now in terms of the computer model with $\boldsymbol{t} = \boldsymbol{t}_0$. Different choices for $\boldsymbol{t}_0$ will lead to different discrepancy functions, $\delta_{\boldsymbol{t}_0}(\boldsymbol{x})$. We can expand $\delta_{\boldsymbol{t}_0}(\boldsymbol{x})$ in (5) as

$$\delta_{\boldsymbol{t}_0}(\boldsymbol{x}) = (\phi(\boldsymbol{x}|\boldsymbol{\tau}) - \kappa(\boldsymbol{x}, \boldsymbol{\tau})) + (\kappa(\boldsymbol{x}, \boldsymbol{\tau}) - \kappa(\boldsymbol{x}, \boldsymbol{t}_0)).$$

The first term on the right hand side might be called the "intrinsic" discrepancy if $\boldsymbol{\tau}$ were known and used for $\boldsymbol{t}$, whereas the second term is the "miscalibration" discrepancy. The model (4) for the computer model runs is the same as (2).

Again following Bayarri et al. (2002, 2005), Higdon et al. (2004) and Kennedy and O'Hagan (2001), $\kappa(\boldsymbol{x}, \boldsymbol{t})$ and $\delta_{\boldsymbol{t}_0}(\boldsymbol{x})$ are modeled as realizations of two independent Gaussian stochastic processes. This approach has been widely used for modeling deterministic computer codes (see, for example, Currin et al. 1991; O'Hagan 1992; Sacks et al. 1989; and Welch et al. 1992).

Thus, we will treat $\kappa(\boldsymbol{x}, \boldsymbol{t})$ as if it is the realization of a random function, $\mu + Z_\kappa(\boldsymbol{x}, \boldsymbol{t})$.

Here, μ is a mean parameter, and $Z_\kappa(\boldsymbol{x}, \boldsymbol{t})$ is a Gaussian stochastic process indexed by $(\boldsymbol{x}, \boldsymbol{t})$ with mean zero and variance σ_κ^2. The covariance between $Z_\kappa(\boldsymbol{x}, \boldsymbol{t})$ and $Z_\kappa(\boldsymbol{x}', \boldsymbol{t}')$ at the two input vectors $(\boldsymbol{x}, \boldsymbol{t})$ and $(\boldsymbol{x}', \boldsymbol{t}')$ is denoted by $\sigma_\kappa^2 R_\kappa((\boldsymbol{x}, \boldsymbol{t}), (\boldsymbol{x}', \boldsymbol{t}'))$. Also following common practice, the correlation function $R_\kappa(\cdot, \cdot)$ will be a product of one-dimensional, Gaussian correlation functions,

$$R_\kappa((\boldsymbol{x}, \boldsymbol{t}), (\boldsymbol{x}', \boldsymbol{t}')) = \exp\left\{ -\sum_{j=1}^{d} \theta_{\kappa,j}(x_j - x_j')^2 - \sum_{j=1}^{q} \theta_{\kappa,d+j}(t_j - t_j')^2 \right\}, \qquad (6)$$

which depends on the unknown parameters $\boldsymbol{\theta}_\kappa = (\theta_{\kappa,1}, \ldots, \theta_{\kappa,d+q})$, where $\theta_{\kappa,j} \geq 0$.

Similarly, we assume that $\delta_{\boldsymbol{t}_0}(\boldsymbol{x})$ is a realization from a second mean-zero Gaussian stochastic process, $Z_\delta(\boldsymbol{x})$, which is independent of $Z_\kappa(\boldsymbol{x}, \boldsymbol{t})$. It is set up analogously to (6), with its own variance, σ_δ^2, and its own correlations parameters, $\boldsymbol{\theta}_\delta = (\theta_{\delta,1}, \ldots, \theta_{\delta,d})$, in the correlation function, $R_\delta(\boldsymbol{x}, \boldsymbol{x}')$. Note that there is no need for parameters corresponding to $\boldsymbol{t}$ in $R_\delta(\boldsymbol{x}, \boldsymbol{x}')$, because whatever value $\boldsymbol{t}_0$ is chosen for the field data model (5), it is constant for all field observations.

Data are collected by running the computer model and from the field. There are m computer-code evaluations at $(\boldsymbol{x}_1^*, \boldsymbol{t}_1), \ldots, (\boldsymbol{x}_m^*, \boldsymbol{t}_m)$ and n field observations at $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Note that the values of $\boldsymbol{x}$ need not be the same in the two data sources, and we use the notation $\boldsymbol{x}^*$ to identify input values for the computer-model runs. The observed values of the output variable of interest are $\mathbf{y}_C = (y_C(\boldsymbol{x}_1^*, \boldsymbol{t}_1), \ldots, y_C(\boldsymbol{x}_m^*, \boldsymbol{t}_m))^T$ from the computer model and $\mathbf{y}_F = (y_F(\boldsymbol{x}_1), \ldots, y_F(\boldsymbol{x}_n))^T$ from the field.

The likelihood function for the joint data, $\mathbf{y} = (\mathbf{y}_C^T, \mathbf{y}_F^T)^T$, is

$$L(\mathbf{y}; \mu, \sigma_\kappa^2, \sigma_\delta^2, \sigma_\varepsilon^2, \boldsymbol{\theta}_\kappa, \boldsymbol{\theta}_\delta, \boldsymbol{t}_0) \propto \quad |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2}(\mathbf{y} - \mu\mathbf{1})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mu\mathbf{1}) \right\}. \qquad (7)$$

Here, $\boldsymbol{\Sigma}$ is the $(m+n) \times (m+n)$ covariance matrix,

$$\boldsymbol{\Sigma} = \sigma_\kappa^2 \boldsymbol{R}_\kappa + \begin{pmatrix} 0 & 0 \\ 0 & \sigma_\delta^2 \boldsymbol{R}_\delta + \sigma_\varepsilon^2 \boldsymbol{I}_n \end{pmatrix},$$

where $\boldsymbol{R}_\kappa$ is the $(m+n) \times (m+n)$ correlation matrix for the combined data from the correlation function $R_\kappa(\cdot, \cdot)$, $\boldsymbol{R}_\delta$ is the $n \times n$ correlation matrix for the field data from the correlation function $R_\delta(\cdot,\cdot)$, and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Element $(i, m+j)$ of $\boldsymbol{R}_\kappa$ is the correlation between $Z_\kappa(\boldsymbol{x}_i^*, \boldsymbol{t}_i)$ and $Z_\kappa(\boldsymbol{x}_j, \boldsymbol{t}_0)$, corresponding to run i of the computer model and observation j in the field $(i = 1, \ldots, m; j = 1, \ldots, n)$. The likelihood depends on $\boldsymbol{t}_0$ through these elements of $\boldsymbol{R}_\kappa$.

Bayarri et al. (2002, 2005), Higdon et al. (2004) and Kennedy and O'Hagan, (2001) considered Bayesian approaches to estimate the unknown parameters $(\mu, \sigma_\kappa^2, \sigma_\delta^2, \sigma_\varepsilon^2, \boldsymbol{\theta}_\kappa, \boldsymbol{\theta}_\delta, \boldsymbol{\tau})$. In this article, we focus instead on the method of maximum likelihood. When interest centers on choosing values for $\boldsymbol{t}_0$, we can examine the profile likelihood. This approach is conceptually and numerically simple. At least for low-dimensional $\boldsymbol{t}_0$, the profile likelihood is easily visualized, and gives the experimenter an indication of uncertainty.

3. IDENTIFIABILITY

Wynn (2001) raised the issue of identifiability of the terms $\kappa(\boldsymbol{x}, \boldsymbol{\tau})$ and $\delta(\boldsymbol{x})$ in model (3), asking whether the two terms are separately estimable. The same concerns would apply to the revised model (5) and we attempt to more formally discuss these issues in this section.

If the computer model does not represent the field mean for any value of $\boldsymbol{t}_0$, (i.e., there does not exist any value for $\boldsymbol{t}_0$ such that $\delta_{\boldsymbol{t}_0}(\boldsymbol{x}) = \phi(\boldsymbol{x}|\boldsymbol{\tau}) - \kappa(\boldsymbol{x}, \boldsymbol{t}_0) = 0$ in (5) for all $\boldsymbol{x}$ of interest), then the meaning of $\boldsymbol{t}_0$ is unclear. In particular, the feasibility of calibration of $\boldsymbol{\tau}$ is questionable. In much simpler contexts, (e.g., linear models), the parameters in an inadequate model will not necessarily be estimated consistently. Thus, we investigate the more promising case where the computer model does match the field mean for at least one value of $\boldsymbol{t}_0$.

Formally, we define the issue of identifiability in terms of two questions. (i) Suppose the computer model is a perfect representation of reality up to specifying an appropriate value for $\boldsymbol{t}_0$, (i.e., $\delta_{\boldsymbol{t}_0}(\boldsymbol{x}) = 0$ in (5) for all $\boldsymbol{x}$ of interest). Whatever value of $\boldsymbol{t}_0$ is used in $\kappa(\boldsymbol{x}, \boldsymbol{t}_0)$, however, the bias term can model the slack, $\phi(\boldsymbol{x}|\boldsymbol{\tau}) - \kappa(\boldsymbol{x}, \boldsymbol{t}_0)$. Does maximum likelihood asymptotically choose a value of $\boldsymbol{t}_0$ such that $\delta_{\boldsymbol{t}_0}(\boldsymbol{x}) = 0$? (We show the answer to this question is "yes", at least in one important special case.) (ii) If a value of $\boldsymbol{t}_0$ is found such that $\delta_{\boldsymbol{t}_0}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$, does this imply $\boldsymbol{t}_0 = \boldsymbol{\tau}$? (We argue the answer to this question is "maybe".)

We give a formal answer to the first question in the important special case of the computer code being computationally efficient. Following Higdon et al. (2004) if the computer code is computationally efficient than an essentially unlimited number of runs from the code can be performed and one can model the difference between the field data and the computer code at a fixed value of $\boldsymbol{t}_0$. From (4) and (5), for any value of $\boldsymbol{t}_0$ the differences between the field and computer model responses are

$$Y(\boldsymbol{x}_i) = Y_F(\boldsymbol{x}_i) - y_\kappa(\boldsymbol{x}_i, \boldsymbol{t}_0) = \delta_{\boldsymbol{t}_0}(\boldsymbol{x}_i) + \varepsilon_i \qquad (i = 1, \ldots, n). \tag{8}$$

The following lemma relates to modelling such data using a random function model for $\delta_{\boldsymbol{t}_0}(\boldsymbol{x})$ as in Section 2. It shows that if $\delta_{\boldsymbol{t}_0}(\boldsymbol{x}_i) = 0$ for $i = 1, \ldots, n$, then asymptotically the method of maximum likelihood will lead to an estimated random-function model that is always 0.

Lemma 1. Suppose response data are generated from

$$Y_i = \delta_i + \epsilon_i \qquad (i = 1, \ldots, n), \tag{9}$$

where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ are unknown constants, and $\epsilon_1, \ldots, \epsilon_n$ are independent Gaussian random variables with mean zero and unknown variance σ^2_ε. Consider representing $\boldsymbol{\delta}$ as a realization of a random function, that is,

$$Y_i = Z_i + \epsilon_i, \qquad (i = 1, \ldots, n), \tag{10}$$

where $\epsilon_1, \ldots, \epsilon_n$ are defined as in (9) and are independent of $Z_1, \ldots, Z_n$, which are Gaussian random variables with mean zero, unknown variance σ^2_δ, and any given positive definite correlation matrix $\boldsymbol{R} \neq \boldsymbol{I}_n$. As $n \to \infty$, the maximum likelihood estimator of σ^2_δ is 0 if $\boldsymbol{\delta} = 0$.

Proof: see appendix.

Turning to the second question, suppose a value of $\boldsymbol{t}_0$ is chosen by maximum likelihood such that $\delta_{\boldsymbol{t}_0}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$. All this says is that $\kappa(\boldsymbol{x}, \boldsymbol{t}_0) = \phi(\boldsymbol{x}|\boldsymbol{\tau})$ for all $\boldsymbol{x}$. As we discussed in Section 2 the discrepancy is the sum of the intrinsic discrepancy and the miscalibration discrepancy. If the sum is zero, this does not imply that the miscalibration discrepancy is zero, i.e. that $\boldsymbol{t}_0 = \boldsymbol{\tau}$.

4. EXAMPLES

We consider an example from chemical kinetics and than revisit the spot welding process introduced in Section 1.

Chemical Kinetics model

The following example is a demonstration of the above methodology, using a model from chemical kinetics with simulated field data. The model under consideration predicts the concentration $y(x)$ of the chemical as a function of time, x, with mean response governed by the equation

$$\phi(x|\tau) = c + y_0 \exp(-\tau x),$$

where c the residual concentration of the chemical at the end of the reaction process, y_0 is the initial concentration of the chemical and τ is an unknown decay rate that is specific to the chemical reaction under consideration (see Fogler, 1999 for more details of the chemical kinetics model).

Assume that three replicates were collected at 11 equi-spaced time points in [0,3] from the following model

$$\phi(x|\tau = 1.7) = 1.5 + 3.5 \exp(-1.7x) + \varepsilon; \quad x \in [0,3] \text{ and } \varepsilon \sim N(0, 0.36).$$

Let the computer model be $\kappa(x,t) = 5 \exp(-tx)$ for $x \in [0,3]$. For this example, we evaluate the computer model at (x,t) values given by a 21-point random Latin Hypercube design (McKay, Conover and Beckman, 1979).

Applying the methodology from Section 2, Figure 2 shows the profile likelihood as a function of the value of t_0 used in the computer model in (5). A grid of 31 values of t_0 on $[0,3]$ is used. Notice that the likelihood surface is multi-modal, suggesting that various values of t_0 will lead to a good model.

For each value of t_0, Figure 3 plots the corresponding predicted computer model, bias, and field mean as functions of x. The predicted computer model and bias functions

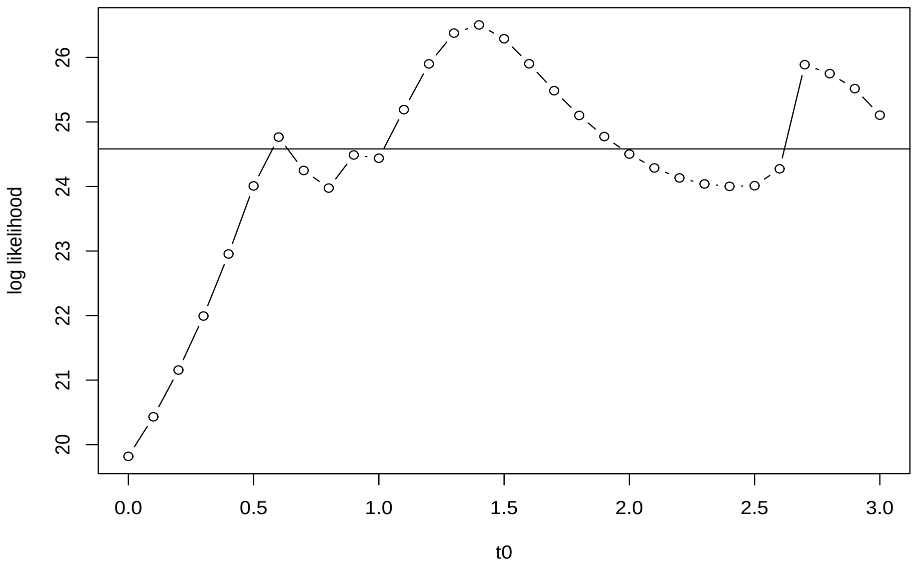

Figure 2: Profile Log Likelihood for The Chemical Kinetics Example. Values of t_0 with a log likelihood above the horizontal line are an approximate 95% confidence region.

change considerably with t_0, as would be expected. Surprisingly, the predicted field mean functions show little variation with respect to t_0.

If the experimenter is concerned only with predicting the field mean, Figure 3 sshows that careful estimation of t_0 is not important in this case. In order to further investigate this we now consider the spot welding process that was introduced in Section 1.

Resistance Spot Welding

In resistance spot welding, two metal sheets of gauge, G, are welded together by compressing them on the top and bottom by copper electrodes with an applied load (L). A current (C) is then applied to the sheets which produces a local heating at the interface (faying surface) where the two sheets have been pressed together. The heat supplied causes the two sheets of metal to melt and after cooling produces a weld nugget. See Figure 4 for a diagram of the process.

In the automotive industry it is often necessary to know the size of the weld nugget under a variety of conditions as this feature is an indication of the weld quality. Since

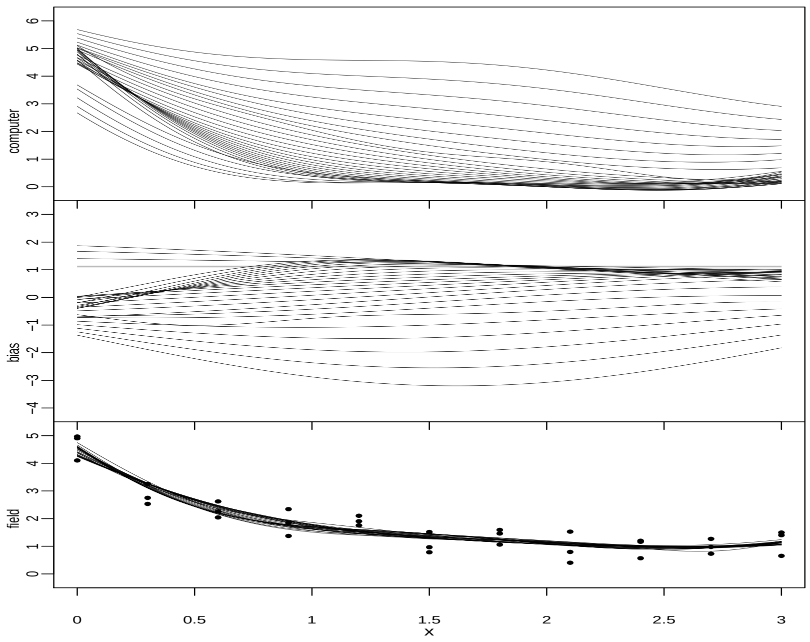

Figure 3: Predicted Curves for the Chemical Kinetics Example for Various Values of t_0. Top panel: computer model, $\kappa(x, t_0)$. Middle panel: discrepancy, $\delta_{t_0}(x)$. Bottom panel: field mean, $\phi(x|\tau) = \kappa(x, t_0) + \delta_{t_0}(x)$, with the field observations superimposed.

the physical process is costly and time consuming to perform, a computer model is constructed to simulate the spot welding process. The computer code uses a coupling of partial differential equations (see Wang and Hayden, 1999) and requires inputs (L, C, G, t). However, only the factors (L, C, G) can be adjusted in the physical process. One of the important aspects of the spot welding process is the amount of heat generated by electrical resistance at the faying surface. The input variable t affects the amount of heat that is generated at the faying surface. Additional details regarding the process can be found in Bayarri et al. (2002, 2005) and Higdon et al. (2004).

There are at least two different data sets relating to this example. The analysis presented here is based on the data reported in Higdon et al. (2004), with $m = 47$ computer runs and $n = 120$ field observations. The latter are 10 replicates of a $2 \times 3 \times 2$ design for (L, C, G). Bayarri et al. (2002, 2005) used another data set with $m = 35$

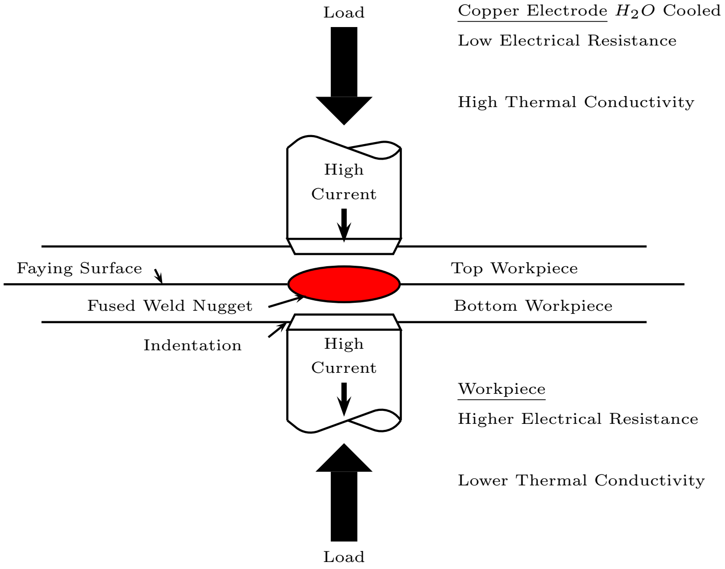
Load
Copper Electrode H_2O Cooled
Low Electrical Resistance
High Thermal Conductivity
High
Current
Faying Surface
Top Workpiece
Fused Weld Nugget
Bottom Workpiece
Indentation
High
Current
Workpiece
Higher Electrical Resistance
Lower Thermal Conductivity
Load

Figure 4: Simplified Diagram of the Resistance Spot Welding Process (see Bayarri et al. 2002).

computer model runs; it gives similar results, and for brevity is not reported here.

Following the notation in Section 2 and the identifiability argument in Section 3, the profile log likelihood in Figure 1 should be a function of t_0 in (5), not a function of τ. The multi-modal plot casts further doubt on whether to relate t_0 to a scalar-valued parameter τ. For comparison, Figure 5 shows the Posterior histogram for τ based on the Higdon et al. (2004) MCMC approach with 50,000 runs.

Comparing the plots in Figure 1 and Figure 5 notice that the posterior histogram is similarly multi-modal, suggesting that the Bayesian approach could be replaced with the simpler likelihood approach here.

The rows of the plot in Figure 6 shows the predicted surfaces for the computer code, the bias and the field mean for each value of t_0 from the profile plot. The columns show the four unique combination of Load and Gauge that were observed in the field. We see the same features that we saw in Figure 3; the predicted surfaces for the field mean, $\kappa(\boldsymbol{x}, t_0) + \delta_{t_0}(\boldsymbol{x})$, for various choices of t_0 are very similar, and again the predicted

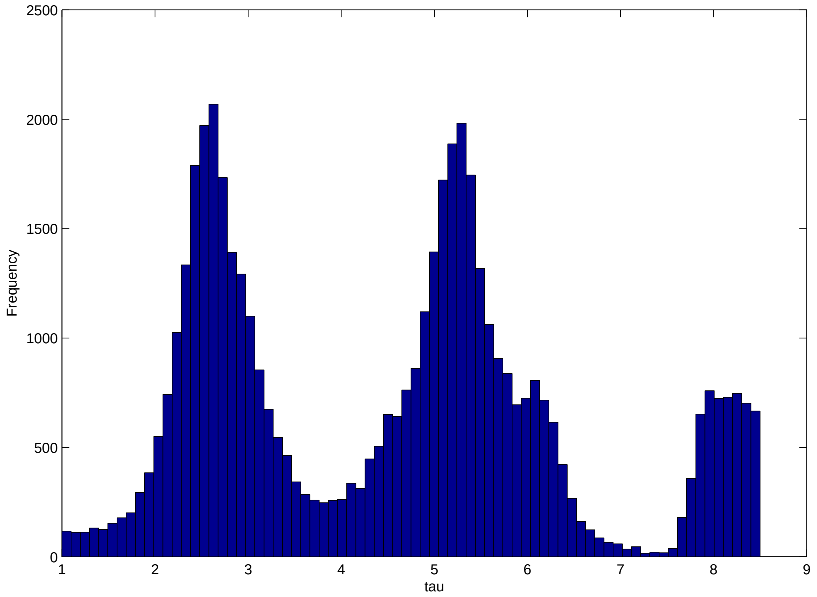

Figure 5: Posterior Distribution for τ for the Spot Weld Data

computer model and the predicted bias functions are quite different. It is also interesting to note that some of the predicted bias functions are flat and very close to zero. This suggests that it may be possible to run the computer code at a carefully chosen value of t_0 to make reliable predictions of the physical process. We will return to this in Sections 5 and 6.

5. TESTING FOR DISCREPANCY

In this section we propose a likelihood ratio test to check if the bias term in (5) is significant. There are two reasons one may wish to do this; first, if the discrepancy is not significant then adequate predictions of the field data may be found by simply using the computer code; secondly, the model below is simpler than the model in (5).

Consider the following model,

$$\begin{aligned} y_C(\boldsymbol{x}, \boldsymbol{t}) &= \kappa(\boldsymbol{x}, \boldsymbol{t}), \\ Y_F(\boldsymbol{x}) &= \kappa(\boldsymbol{x}, \boldsymbol{t}_0) + \varepsilon. \end{aligned} \tag{11}$$

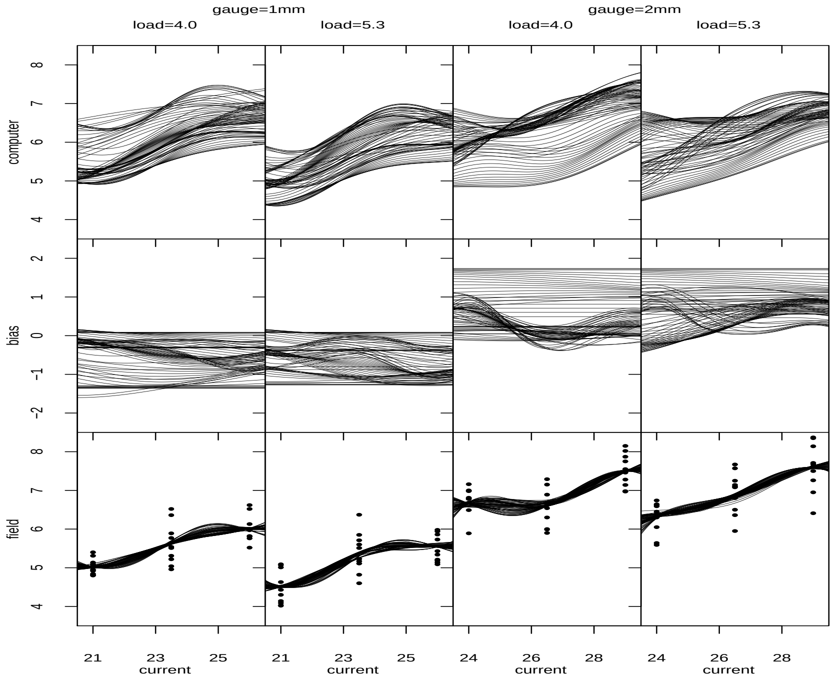

Figure 6: Top Row: Predicted surface for the computer model $\kappa(x, t_0)$. Middle Row: Predicted surface for the discrepancy function $\delta_{t_0}(x)$. Bottom Row: Predicted surface for the field mean, with the actual field observations superimposed.

The likelihood for model (11) is the same as the likelihood in (7) except

$$\boldsymbol{\Sigma} = \sigma_\kappa^2 \boldsymbol{R}_\kappa + \begin{pmatrix} 0 & 0 \\ 0 & \sigma_\varepsilon^2 \boldsymbol{I}_n \end{pmatrix}.$$

The estimation of $\boldsymbol{t}_0$ can again be done using the profile likelihood, or by treating $\boldsymbol{t}_0$ as further parameters to be numerically optimized.

Since (11) is a sub-model of (5) standard likelihood results can be used to test if there is a discrepancy between the two sources of data. The appropriate asymptotic test would be to compare twice the difference in the two maximum log-likelihood values with a χ^2_ν distribution where $\nu = d + 1$ is the difference in the number of parameters.

In general, the maximum likelihood estimates of $\boldsymbol{t}_0$ will not be the same in both models, thus the above test is not of practical interest. That is, the experimenter is interested in knowing if there exists a value of $\boldsymbol{t}_0$ that can be used to make the bias term essentially zero for all values of $\boldsymbol{x}$. Thus it would be more appropriate to check if the there is a significant bias by using the same value of $\boldsymbol{t}_0$ in both models.

Recall in Section 4 we mentioned that some of the bias functions for the spot weld data were flat and relatively close to zero. Applying the methodology above leads to $\hat{\boldsymbol{t}}_0 = 7.7$ for the model without bias, whereas the model with bias led to an estimate of $\hat{\boldsymbol{t}}_0 = 2.5$. Using the test above with different values of $\boldsymbol{t}_0$ and degrees of freedom equal to four the test indicates that the bias is not significant. However, as mentioned above it is more appropriate to compare the likelihoods at the same value of $\boldsymbol{t}_0$, i.e. $t_0 = 7.7$. In this case the degrees of freedom are three and the test still indicates that the bias is not significant.

6. STRATEGIES FOR PREDICTION

In Section 1 we outlined three possible strategies for predicting the mean of the physical system, $\phi(\boldsymbol{x}|\tau)$ at $\boldsymbol{x}$:

$$\hat{\phi}(\boldsymbol{x}|\tau) = \kappa(\boldsymbol{x}, \boldsymbol{t}_0), \tag{12}$$

where $\boldsymbol{t}_0$ is chosen to be appropriate for model (11);

$$\hat{\phi}(\boldsymbol{x}|\tau) = \kappa(\boldsymbol{x}, \boldsymbol{t}_0) + \hat{\delta}_{\boldsymbol{t}_0}(\boldsymbol{x}), \tag{13}$$

where $\hat{\delta}_{\boldsymbol{t}_0}(\boldsymbol{x})$ is the best linear unbiased predictor (BLUP) of $\delta_{\boldsymbol{t}_0}(\boldsymbol{x})$ in the full model (5), and $\boldsymbol{t}_0$ is chosen for this model; and

$$\hat{\phi}(\boldsymbol{x}|\tau) = \kappa(\boldsymbol{x}, \widehat{\boldsymbol{t}_0) + \delta}_{\boldsymbol{t}_0}(\boldsymbol{x}), \tag{14}$$

which is the prediction of the sum of the computer model and bias functions from the full model (5). The strategies in (12) and (13) involve running the computer model again at $\boldsymbol{x}$. For models with multiple sources of data, the forms of the BLUPs in (13) and (14) and their associated mean squared errors of prediction are given by Santner, Williams and Notz (2003). (See also Journel and Huijbregts, 1978, where similar methodology applied to geostatistics is called cokriging).

To explore the merits of the three strategies and various choices of value for t_0, we again consider the chemical kinetics example in Section 4. As $\phi(x|\tau)$ is known, it is possible to calculate the root mean squared error (RMSE) of prediction to assess accuracy. Figure 7 gives the RMSE values averaged over the 10 sites $x = \{0.15, 0.45, \ldots, 2.85\}$, which are the midpoints of the 11 values of x for the field data. Each strategy is evaluated over a grid of t_0 values.

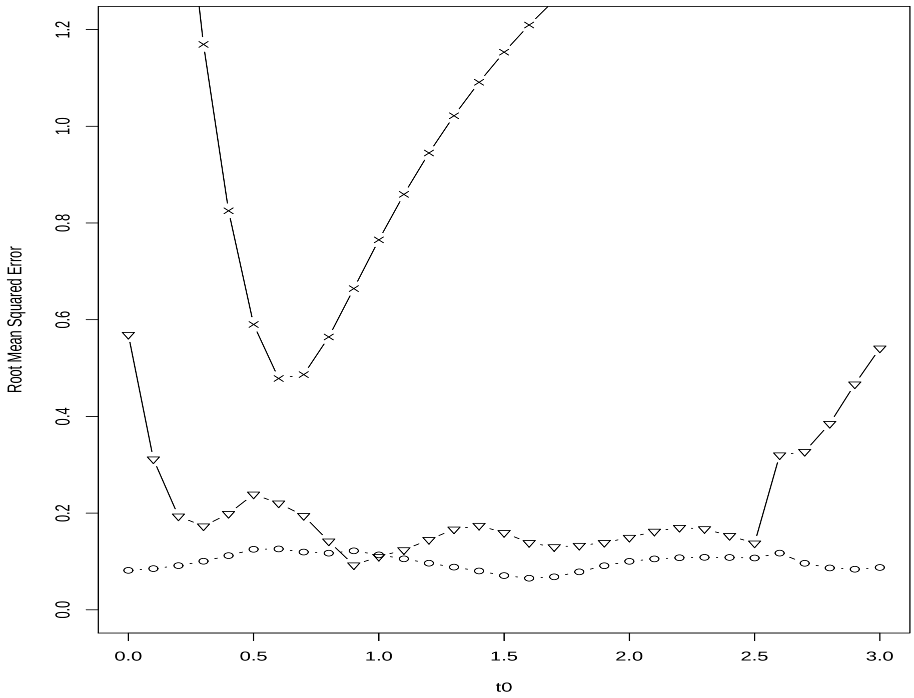

Figure 7: RMSE of prediction for various values of t_0. Strategies (12), (13), and (14) are denoted by "$\times$", "$\triangledown$", and "$\circ$", respectively.

We see immediately in Figure 7 that carefully choosing a value of t_0 is important when predictions are from (12) or (13) but not important for strategy (14). Secondly, strategy (14) typically performs much better.

We next give more detailed discussion of strategies (12) and (14), and conclude with the more difficult (13).

Using the Computer Code Directly As In (12)

We assume that the computer code is computationally demanding, but it is possible to run the code at a small number of future sites. From Figure 7 we can see that $t_0 = 0.6$ minimizes the RMSE. This is consistent with the MLE, $\hat{t}_0 = 0.6$ found using model (11).

The RMSE is large, however, relative to the other two strategies. This is not surprising since the likelihood ratio test in Section 5 indicates that there is a significant bias.

Using the Full Model for Prediction As In (14)

In Figure 7 we see that the RMSE is relatively small for various choices of t_0. The minimum RMSE occurs when $t_0 = 1.6$, whereas the MLE is $\hat{t}_0 = 1.4$. It is interesting to note that the likelihood for this example appears to provide very little information for selecting an appropriate value of t_0. However, the root mean squared error at $t_0 = 1.4$ is about 1.2 times larger than the minimum value at $t_0 = 1.6$. Overall, however, this method performs well and is robust to choice of t_0.

Using the Computer Code Directly With an Adjustment for the Bias As In (13)

The RMSE values in Figure 7 show a few surprises. First, the minimum RMSE occurs at $t_0 = 0.9$, and secondly the RMSE for the full model is smaller at all other values of t_0. At first glance this appears to be counter-intuitive: One would expect that knowing the true value of the code would result in a better prediction. If this strategy is to be effective, however, the model in (5) has to provide a good estimate of the bias function. Consider Figure 8, which compares the estimated and true functions for each component of the model, with t_0 set to 0, for example. It is clear that the estimates of both the computer model and the bias functions are relatively poor, whereas the field mean function is well predicted.

Alternatively, to utilize a new computer model run, in general one could use the following procedure: i) run the code at $\kappa(\boldsymbol{x}, \hat{\boldsymbol{t}}_0)$; ii) add the new observation to the existing data and reconstruct the full model from (5); and iii) use the new full model and predict based on (14).

Using this strategy for the kinetics data and the MLE, $\hat{t}_0 = 1.4$, the RMSE is about one third of that from (13) and about two thirds that from (14) without a new run. Thus, if one is prepared to run the code again, it seems more advantageous to recompute the full model and use it for prediction.

7. DISCUSSION

In this paper we introduced a likelihood based approach for model calibration in

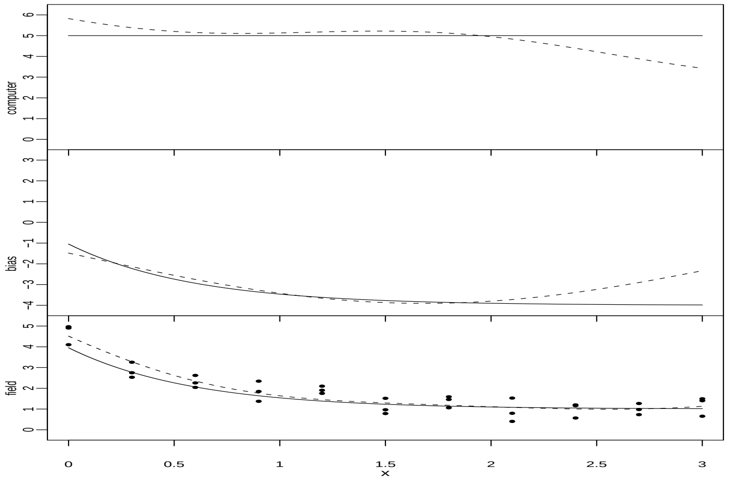

Figure 8: Top Panel: Dashed line shows the predicted surface for the computer model $\kappa(x, 0)$, solid line shows the true function $\kappa(x, 0)$. Middle Panel: Dashed line shows the predicted surface for the discrepancy function $\delta_0(x)$, the solid line shows the true discrepancy. Bottom Panel: Dashed line shows the predicted surface for the field mean function, solid line shows the true function, with the actual field observations superimposed.

computer experiments. By using the likelihood approach we were able to make three simple points. First, it maybe impossible to find an interpretable estimate of a calibration parameter. However, even if one could estimate the calibration parameter exactly, it may be the case that this value leads to a worse prediction of the process mean than a value of the parameter that was used to adjust the computer code (see Kennedy and O'Hagan, 2001). Secondly, the choice of the tuning parameter is relatively unimportant when a model with discrepancy is considered. Lastly, the choice of the tuning parameter varies depending on how one intends to use the computer model. Finally, we concluded the paper with some practical considerations for the experimenter when applying the methodology for model calibration.

APPENDIX

Proof of Lemma 1: The log likelihood for the model in (10) is

$$l(\sigma_\delta^2, \boldsymbol{R}, \sigma_\varepsilon^2; \mathbf{Y}) = -\frac{1}{2}(n \ln(2\pi) + \ln|\mathbf{\Sigma}| + \mathbf{Y}^T\mathbf{\Sigma}^{-1}\mathbf{Y}),$$

where $\mathbf{\Sigma} = \sigma_\delta^2 \boldsymbol{R} + \sigma_\varepsilon^2 \boldsymbol{I}_n$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$. Applying the results for differentiating a log determinant or a matrix inverse given by, for example, Searle, (1982, pp. 335, 337), a second-order Taylor series expansion of the log likelihood around $\sigma_\delta^2 = 0$ is

$$l(\sigma_\delta^2, \boldsymbol{R}, \sigma_\varepsilon^2; \mathbf{Y}) \simeq l(0, \boldsymbol{R}, \sigma_\varepsilon^2; \mathbf{Y}) + \frac{\sigma_\delta^2}{2\sigma_\varepsilon^2}\left(\frac{\mathbf{Y}^T \boldsymbol{R} \mathbf{Y}}{\sigma_\varepsilon^2} - n\right) + \frac{(\sigma_\delta^2)^2}{4\sigma_\varepsilon^4}\left(\mathrm{tr}(\boldsymbol{R}^2) - \frac{2\mathbf{Y}^T \boldsymbol{R}^2 \mathbf{Y}}{\sigma_\varepsilon^2}\right). \tag{15}$$

Because the data, $\mathbf{Y}$, are assumed to be generated by (9), $\mathbf{Y}^T \boldsymbol{R} \mathbf{Y}/\sigma_\varepsilon^2$ is a quadratic form in Gaussian random variables. Thus, the coefficient of σ_δ^2 in (15) has expectation $\boldsymbol{\delta}^T \boldsymbol{R} \boldsymbol{\delta}/(2\sigma_\varepsilon^4)$ (e.g., Seber, 1977, p.13), which is zero if $\boldsymbol{\delta} = 0$ and positive otherwise. The coefficient's standard deviation is $\sqrt{n/2}/\sigma_\varepsilon^2$ (e.g., Seber, 1977, p.16). Similarly, the coefficient of $(\sigma_\delta^2)^2$ in (15) has expectation $-(\mathrm{tr}(\boldsymbol{R}^2) + \boldsymbol{\delta}^T \boldsymbol{R}^2 \boldsymbol{\delta}/\sigma_\varepsilon^2)/(4\sigma_\varepsilon^2)$ and standard deviation $\sqrt{\mathrm{tr}(\boldsymbol{R}^2)}/(4\sigma_\varepsilon^2)$. We also note that $n < \mathrm{tr}(\boldsymbol{R}^2) \leq n^2$. Thus, the coefficient of $(\sigma_\delta^2)^2$is negative with probability 1 as $n \to \infty$. Comparing the distributions of the two coefficients, we see that $\sigma_\delta^2 = 0$ is a maximum likelihood solution as $n \to \infty$ if $\boldsymbol{\delta} = 0$. □

ACKNOWLEDGEMENTS

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the National Program on Complex Data Structures, Canada. This work was motivated by discussions with J. Sacks.

REFERENCES

Bayarri,M., Berger,J.O., Higdon, D., Kennedy, M., Kottas,A., Paulo,R., Sacks,J., Cafeo,J.A., Cavendish,J.C., Lin,C-H., Tu, J. (2002), "A Framework for Validation of Computer Models," *National Institute of Statistical Sciences Technical Report*, 128, www.niss.org/downloadabletechreports.html.

Bayarri, M., Berger, J.O., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J. C.H., Lin and Tu, J. (2005), "A Framework for Validation of Computer Models," *National Institute of Statistical Sciences Technical Report.*

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991), "Bayesian Prediction of Deterministic Functions With Applications to the Design and Analysis of Computer Experiments," *Journal of the American Statistical Association*, 86, 953-963.

Fogler, H. S., (1999), *Elements of Chemical Reaction Engineering,* Prentice Hall.

Higdon, D., Kennedy, M., Cavendish, J.C., Cafeo, J.A. and Ryne, R.D. (2004), "Combining Field Data and Computer Simulation for Calibration and Prediction," *SIAM Journal on Scientific Computing*, 26, 448-466.

Journel, A.G. and Huijbregts, C.J. (1978), *Mining geostatistics,* Academic Press.

Kennedy, M.C. and O'Hagan, A. (2001), "Bayesian Calibration of Computer Models," (with discussion) *Journal of the Royal Statistical Society, Series B*, 63, 424-462.

McKay, M.D. Conover, W.J. and Beckman, R.J. (1979), " A Comparison of Three Methods for Selecting the Inputs Variables in the Analysis of the Output From a Computer Code," *Technometrics*, 21, 239-245.

O'Hagan, A. (1992), "Some Bayesian Numerical Analysis" (with discussion), *Bayesian Statistics 4* (eds. J.M. Bernardo,J.O. Berger, A.P. Dawid and A.F.M. Smith), 345-363. Oxford.

Sacks, J., Schiller, S.B. and Welch, W.J. (1989), "Designs for Computer Experiments," *Technometrics*, 31, 41-47.

Sacks, J., Welch, W.J., Mitchell, T. and Wynn, H.P. (1989), "Designs and Analysis of Computer Experiments"(with discussion), *Statistical Science*, 4, 409-435.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Designs and Analysis of Computer Experiments,* Springer-Verlag.

Seber, G.A.F., (1977), *Linear Regression Analysis,* Wiley.

Searle, S.R. (1982), *Matrix Algebra Useful for Statistics,* Wiley.

Wang, P.C., and Hayden, D.B. (1999), "Computational Modeling of Resistance Spot Welding of Aluminum," *GM Research Report* R& D-9152, General Motors Research and Development Center, Warren MI.

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J. and Morris, M.D. (1992), "Screening Predicting and Computer Experiments," *Technometrics*, 34, 15-25.

Wynn, H.P. (2001), "Discussion of the paper by Kennedy and O'Hagan," *Journal of the Royal Statistical Society, Series B*, 63, 450-451.