




Bayesian Calibration of Inexact Computer Models

Matthew Plumlee


To cite this article: Matthew Plumlee (2017) Bayesian Calibration of Inexact Computer Models, Journal of the American Statistical Association, 112:519, 1274-1285, DOI: 10.1080/01621459.2016.1211016

To link to this article: <https://doi.org/10.1080/01621459.2016.1211016>

 View supplementary material [↗](#)

 Accepted author version posted online: 20 Jul 2016.
Published online: 13 Jun 2017.

 Submit your article to this journal [↗](#)

 Article views: 744

 View related articles [↗](#)

 View Crossmark data [↗](#)

 Citing articles: 2 View citing articles [↗](#)



Bayesian Calibration of Inexact Computer Models

Matthew Plumlee

Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI

ABSTRACT

Bayesian calibration is used to study computer models in the presence of both a calibration parameter and model bias. The parameter in the predominant methodology is left undefined. This results in an issue, where the posterior of the parameter is suboptimally broad. There has been no generally accepted alternatives to date. This article proposes using Bayesian calibration, where the prior distribution on the bias is orthogonal to the gradient of the computer model. Problems associated with Bayesian calibration are shown to be mitigated through analytic results in addition to examples. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received June 2015
Revised June 2016

KEYWORDS

Calibration; Computer experiments; Deterministic models; Gaussian processes; Identifiability; Kriging; Model inadequacy; Orthogonal processes; Uncertainty quantification

1. Background on Calibration

A computer model is the mathematical representation of a system evaluated via a single computer or a set of computers. Computer models are widely adopted to understand specific phenomena or the general behavior of systems. For example, computer models can assist an engineer in designing and testing new processes without costly prototype development.

The focus of this article is *calibration*. Calibration is the act of using observations from a real, physical experiment to learn about and adjust the computer model. This action is often required to ensure that the computer model's mathematical representation adequately emulates reality. For each observation, or potential observation, it is assumed the computer model has a corresponding response that deterministically depends on a set of specified conditions. The conditions are partitioned into two categories: (1) a set of general inputs that represent controllable or measurable quantities in a real system and (2) a set of calibration parameters which cannot be directly controlled or measured in a real system. This article will distinguish these by terming the groups the *input* and the *parameter* of the computer model, respectively.

From a broad statistical perspective, computer models can be considered nonlinear response functions. The key difference between computer models and typical regression models is the structure of the computer model is fixed by the physics/science of the system of interest. Therefore, the usual regression methods for addressing model inadequacy, such as including higher order or interaction terms, are not applicable. As a result, a computer model is often considered *inexact*. This means the computer model does not perfectly match the real system even if some *parameters* included with it are set optimally. In their landmark work, Kennedy and O'Hagan (2001a) took

a Bayesian approach to handling computer model *bias*, the difference between the model and the truth. In contrast, traditional nonlinear regression has operated under the assumption that the model accurately represents reality if the correct parameters are chosen (Bates and Watts 1988).

As an example, consider the following example of an inexact computer model. A ball is fired from a chosen vertical position y_0 with an unknown vertical velocity v_0 . The ball drops due to gravity as time increases. At select instances of time, both the time, denoted x , and the position of the ball, which is the response, can be recorded. From basic projectile mechanics, the second derivative of the vertical position is equal to $-g$, where g is an unknown rate of acceleration caused by gravity. The computer model is then given by

$$y_0 + v_0x - \frac{g}{2}x^2.$$

It follows that x is the input and (v_0, g) is the parameter of the computer model. In reality, air resistance will slow the ball via a quadratic drag force and thus the second derivative of the ball's vertical height is equal to $-g + cv^2(x)$, where $v^2(x)$ is the square of the vertical velocity of the ball at time x . Say that the initial position is 8, the initial velocity is -1 , the acceleration due to gravity is 10 and the coefficient of drag, c , is 0.2. Then, the ball's vertical position at time x is given by

$$\frac{5}{2} \log \left(\frac{50}{49} - \frac{50}{49} \tanh \left(\tanh^{-1}(\sqrt{0.02}) + \sqrt{2}x \right)^2 \right) + 8. \quad (1)$$

The computer model is biased; no matter what values are selected for v_0 and g , $8 + v_0x - \frac{g}{2}x^2$ will never align perfectly with the true response over all values of the input. Nonetheless, the computer model can be useful in emulating the system. Figure 1 illustrates this, showing the true function along with

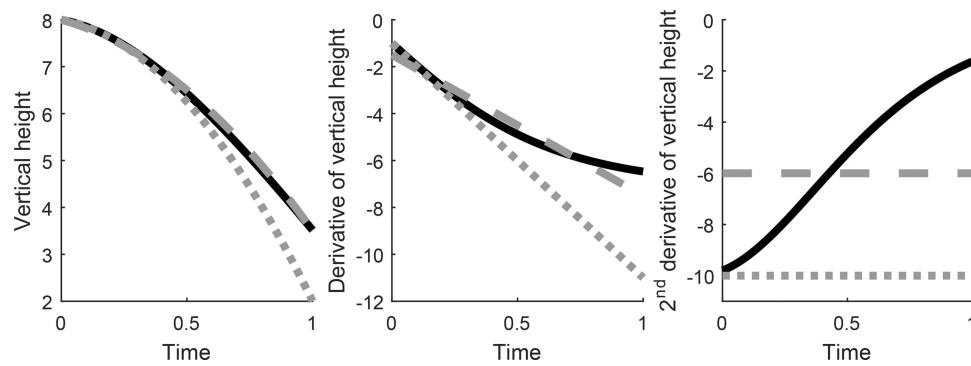


Figure 1. The response from the real system (solid, dark line) for the ball example in Section 1 versus the computer model with $(v_0, g) = (-1.0, 10.0)$ (short dashes) and $(-1.5, 6.0)$ (long dashes). The left panel is the vertical height, the middle panel is the derivative of the vertical height with respect to time, and the right panel is the second derivative of the vertical height with respect to time.

the computer model given two different values of the parameter. When $v_0 = -1.5$ and $g = 6$, the computer model represents the response well when time is between 0 and 1, despite the existence of a model bias. Compared to using $v_0 = -1$ and $g = 10$, $v_0 = -1.5$ and $g = 6$ yields a closer representation of the system in the domain $[0, 1]$. This is also true for the first and second derivatives of the vertical position, where using $v_0 = -1.5$ and $g = 6$ gives values close to the true function, even though it is not perfect.

Kennedy and O’Hagan (2001a) created a large impact among statisticians who use computer models. To name a select few, see Higdon et al. (2004), Goldstein and Rougier (2004), Bayarri et al. (2007), Qian and Wu (2008), Joseph and Melkote (2009), Wang, Chen, and Tsui (2009), Han, Santner, and Rawlinson (2009), Kleiber et al. (2013), Joseph and Yan (2015), and Storlie et al. (2015). Kennedy and O’Hagan’s statistical inference package includes, simultaneously, parameter and bias function estimation. This bias estimation allows for bias correction, where the estimate of the true response is formed by taking the sum of the computer model plus the bias. The analysis is done by looking at the posterior distribution of the parameter value and the bias function after placing prior distributions on both.

Despite the excitement over these tools, there have also been lingering concerns. As a recent example, Farah et al. (2014) stated some concern over using Kennedy and O’Hagan’s calibration tools. Similarly, Gramacy et al. (2015) discarded Kennedy and O’Hagan’s full posterior in part because of the “identification issues known to plague [Kennedy and O’Hagan]-style calibration.” Tuo and Wu (2015b) showed that, under a reasonable model of data generation, the posterior mode of the parameter will depend on the prior distribution of the bias even as the number of observations gets very large. Thus, the choice of the bias’s prior becomes a permanent fixture in the parameter’s posterior. This effect is undesirable because of the difficulty in conjecturing an exact prior for the bias. Inexperienced users are rightfully hesitant to adopt a Bayesian method with so much at stake in the prior specification stage. Even those with a great deal of experience, like Keith Beven and Peter Diggle, noted the challenge of this task in the written discussions of Kennedy and O’Hagan’s article.

This article will show that Bayesian calibration is possible without the presence of these issues by adopting the assumption that there is some value of that parameter that is optimal

under some loss function (defined generally). Under this framework, previously suggested priors for the bias are shown to be inadequate. We suggest general priors for the bias to be used instead. The core methods of Bayesian calibration are left intact, including the ability to account for the noise in the physical observations and the uncertainty induced when the computer model is not exactly known.

Section 2 introduces a framework for calibration and discusses the practical implications. Section 3 outlines the proposed Bayesian inference package that accounts for an orthogonality condition that the prior on the bias ought to have. Section 4 deals with the practical issues arising from this formulation, namely designing a computational approach to finding general bias functions priors that meet this orthogonality condition. Section 5 discusses three examples that illustrate the practical implications of the proposed methods. Section 6 offers some brief thoughts on future directions for calibration research.

2. The Parameter and Model Bias

The objective of this section is to delineate the assumptions built into, but often not explicitly stated, for calibration. Section 2.1 gives the three major assumptions needed to frame the calibration problem. Section 2.2 compares this framework to Kennedy and O’Hagan (2001a). Section 2.3 narrows the definition of the calibration problem to the loss functions that are considered in this article.

2.1. Definition of the Parameter and Model Bias

This subsection outlines the explicit mathematical definition and assumptions behind the true function, the computer model, and the parameter. Three assumptions are introduced labeled (i), (ii) and (iii). These form the basis for the proposed analysis method. This definition does not depend on the statistical model, nor does it depend on which observations are recorded or the inputs that generate the recorded observations.

Every possible input is located in a bounded, open X , $X \subset \mathbb{R}^d$. This space X is the physical experimental design region for the inputs. This does not imply that every point in X will be observed in a physical experiment, only that the points in X are potential experimental inputs.

The true, unknown function of the input is termed $y(\cdot)$, which is provided by nature and is thus considered fixed. This is commonly defined as the expected value of the observations given the input. Define $D^{(\alpha)}y(x)$ as the mixed partial derivative of $y(\cdot)$ with orders α at x . The first major assumption is thus:

- (i) $y(\cdot)$ is a deterministic map from X to \mathbb{R} . There is some integer $k \geq 0$ such that for all length d vectors of non-negative integers α with $\|\alpha\|_{L^1} \leq k$, $D^{(\alpha)}y(\cdot)$ exists and is bounded over X .

The computer model is termed $f(x, t)$, where x is a possible input and t is a possible parameter in some set $\Theta \subset \mathbb{R}^p$. For some value of $\theta \in \Theta$, we are hopeful that the computer model $f(\cdot, \theta)$ is very close to $y(\cdot)$. Analogous to the above statement of (i), the second key assumption is:

- (ii) $f(\cdot, \cdot)$ is a deterministic map from $X \times \Theta$ to \mathbb{R} such that for all length d vectors of nonnegative integers α with $\|\alpha\|_{L^1} \leq k$, $D^{(\alpha,0)}f(\cdot, \cdot)$ exists and is bounded over $X \times \Theta$.

The k -differentiability condition is carried over from nature's true function. $D^{(\alpha,0)}f(x, t)$ is the mixed partial derivative of $f(\cdot, t)$ with orders α at x .

With assumptions (i) and (ii), both nature's true function and the computer model are fixed mappings. It is thus reasonable that if an oracle knew these entire functions, this oracle could choose the "best" value of θ possible. To define the parameter in specific terms, we use an explicit decision rule for the oracle to make their decision based on the difference between $y(\cdot)$ and $f(\cdot, t)$ for all $t \in \Theta$. This leads to the third major assumption:

- (iii) There is a mapping L from the space of k differentiable functions on X to \mathbb{R} . There is some $\theta \in \Theta$ such that

$$L\{y(\cdot) - f(\cdot, \theta)\} < L\{y(\cdot) - f(\cdot, t)\},$$

for all $t \in \Theta$ such that $t \neq \theta$. (2)

That is, there is a function $L(\cdot)$ that can distinguish the best possible parameter from all other possible values of the parameter.

Borrowing from the notation of score functions (Gneiting and Raftery, 2007), the loss L is considered strictly proper if

$$L\{y(\cdot) - y(\cdot)\} < L\{y(\cdot) - g(\cdot)\},$$

for all k differentiable functions $g(\cdot) \neq y(\cdot)$.

Thus, if there is some unique $t^* \in \Theta$ such that $y(x) = f(x, t^*)$ pointwise for all $x \in X$, then if L is strictly proper, $\theta = t^*$. The computer model is biased if no such value exists.

The function $z_\theta(\cdot)$ is the bias function associated with the computer model. Using assumptions (i), (ii), and (iii), the bias is straightforwardly defined as

$$z_\theta(x) := y(x) - f(x, \theta),$$

and we have that

$$y(x) = f(x, \theta) + z_\theta(x).$$

The dependence of the bias on θ is often suppressed in the literature, but it will be included here for clarity.

The definition of θ in assumption (iii) is fixed in this article (up to the loss function), and this may not agree with other definitions of θ . As discussed in Higdon et al. (2004) and Han,

Santner, and Rawlinson (2009), there are cases when the parameter has an explicit physical meaning and other cases when the parameter has little or no physical meaning. We make no such distinction and always view the parameter as an artifact of nature's function, our computer model, and our choice of loss function for the oracle.

2.2. Comparison to Kennedy and O'Hagan (2001)

Kennedy and O'Hagan (2001a) stated that the parameter θ is the one that best explains the difference between the true model and the computer model based on the prior distribution. This statement appears to imply the parameter is the minimizer of the reproducing kernel Hilbert space norm based on the prior distribution (Bayesian framework) of the bias. This definition was not outlined in their original article but was attributed to them later (Tuo and Wu 2015b), as Kennedy and O'Hagan offered no formal definition of the parameter. When this interpretation is used, it agrees with the framework established by assumptions (i), (ii), and (iii). The use of the prior to define the parameter was critiqued by Tuo and Wu (2015a) and Tuo and Wu (2015b). Section 2.3 considers definitions of the parameter that are separate from our prior distribution to avoid conflating the prior and the definition of the parameter.

Even if this description of the parameter (as the minimizer of the reproducing kernel Hilbert space norm corresponding to the bias's prior) is acceptable, Kennedy and O'Hagan's approach to using the same prior distribution on the bias independent of θ is not. This is because by defining the parameter, as done in assumptions (i), (ii), and (iii), we implicitly require orthogonality of the bias function and an aspect of the computer model. Section 3 provides the full technical details.

2.3. Loss Functions

This subsection outlines the structure of the class of loss functions considered in this work. These are motivated by different applications, see Section 5 for more discussion of the practical implications of different losses. The structures of these losses are such that they can be easily incorporated into the proposed orthogonality approach (Section 3) for creating bias function prior distributions because of their quadratic form.

The recent article of Tuo and Wu (2015b) discussed a specific version of this framework. They considered the loss

$$\begin{aligned} L_{L^2}\{y(\cdot) - f(\cdot, t)\} &:= \int_X (y(\xi) - f(\xi, t))^2 d\xi \\ &= \|y(\cdot) - f(\cdot, t)\|_{L^2}^2. \end{aligned} \quad (3)$$

Han, Santner, and Rawlinson (2009) used a framework like this for tuning parameter calibration. This L^2 loss is named after the norm in the associated Hilbert space.

As another loss, consider the inner product of $y(\cdot) - f(\cdot, t)$ with itself in a Sobolev space (Wendland, 2004, p. 133). Looking at the Hilbert space W_k^2 , the Sobolev space of order k , the loss is thus

$$L_{W_k^2}\{y(\cdot) - f(\cdot, t)\} := \sum_{\|\alpha\|_{L^1} \leq k} \|D^{(\alpha)}y(\cdot) - D^{(\alpha,0)}f(\cdot, t)\|_{L^2}^2.$$

Another important extension of the loss is the incorporation of a weighting measure $\mu(\cdot)$ defined over a sigma-field on X . We presume that this measure is a probability measure for simplicity. This measure will define the relative importance of different sectors of the input space on the loss. Using μ ,

$$L_{L^2(\mu)}\{y(\cdot) - f(\cdot, t)\} := \int_X (y(\xi) - f(\xi, t))^2 d\mu(\xi) = \|y(\cdot) - f(\cdot, t)\|_{L^2(\mu)}^2.$$

The minimizer of this loss will attempt to make the computer model as close as possible to the true function, giving special weight to regions where the measure is large.

Placing these ideas together, the most general loss this article will consider is

$$L_{W_k^2(\mu)}\{y(\cdot) - f(\cdot, t)\} := \sum_{\|\alpha\|_{L^1} \leq k} \|D^{(\alpha)}y(\cdot) - D^{(\alpha,0)}f(\cdot, t)\|_{L^2(\mu)}^2. \tag{4}$$

This function is strictly proper in the sense that $L_{W_k^2(\mu)}\{y(\cdot) - g(\cdot)\} > 0$ if $g(\cdot) \neq y(\cdot)$ on a subset of X with positive measure and $L_{W_k^2(\mu)}\{y(\cdot) - y(\cdot)\} = 0$.

3. Bayesian Calibration, Orthogonality, and Bias Functions

This section motivates and outlines the Bayesian analysis of observational data using the principles described in Section 2. Section 3.1 gives a general overview of the approach to the proposed Bayesian analysis of data and describes the needed prior distributions and Sections 3.2–3.4 will give the details and illustrations that motivate this approach.

3.1. General Bayesian Framework

Say that there are a set of inputs, $\{x_1, \dots, x_n\}$, which correspond to a set of observations from the real system, labeled $\{Y_1, \dots, Y_n\}$. The observations are considered stochastic due to some source of noise in the real system and thus Y_i differs from $y(x_i)$ by an unknown value. The prior distributions on the differences between Y_i and $y(x_i)$ are assumed independent normal distributions with zero mean and variance v . This normality assumption may require some transformation or agglomeration of the raw data.

Let π generally represent a probability density and $\pi(a|b)$ implies the conditional probability density of a given b . The parameter θ will be assumed to have any prior distribution labeled $\pi(\theta)$ decided by a user. In a Bayesian inference setup, we thus would like to find

$$\pi(\theta|Y = (Y_1, \dots, Y_n)^T),$$

which would define the *posterior* of θ . The boldface notation will be used for matrices and vectors whose size depends on n and the superscript \top represents transpose. Using Bayes rule,

$$\pi(\theta|Y) \propto \pi(Y|\theta)\pi(\theta),$$

where \propto stands for equality up to a constant multiplier. Expanding out the conditional:

$$\pi(\theta|Y) \propto \int_{\mathbb{R}^n} \pi(Y|z_\theta(x_1), \dots, z_\theta(x_n), \theta)\pi(z_\theta(x_1), \dots, z_\theta(x_n)|\theta)\pi(\theta) d(z_\theta(x_1), \dots, z_\theta(x_n)). \tag{5}$$

Y given $(z_\theta(x_1), \dots, z_\theta(x_n), \theta)$ is thus a product of n independent normal densities with mean $y(x_i) = f(x_i, \theta) + z_\theta(x_i)$ and variance v .

Like Kennedy and O’Hagan (2001a), a Gaussian process model serves as the prior distribution on the function $z_\theta(\cdot)$, which is the bias conditional on θ . The Gaussian process is a valid probability distribution on continuous functions with easily computable conditional distributions. This prior says that for every $n = 1, 2, 3, \dots$ the joint distribution of $z_\theta(x_1), z_\theta(x_2), \dots, z_\theta(x_n)$ is multivariate normal. In this case, $z_\theta(x)$ has zero mean and the covariance between $z_\theta(x)$ and $z_\theta(x')$ is $\sigma^2 r_\theta(x, x')$, where σ^2 a scaling constant. This function $r_\theta(\cdot, \cdot)$ must be positive semidefinite on $X \times X$, which means that the covariance matrix of $z_\theta(x_1), z_\theta(x_2), \dots, z_\theta(x_n)$ is non-negative definite. Gaussian processes have been used for prior distributions on functions in a diverse set fields beyond computer model calibration such spatial statistics (Cressie 1993) and machine learning (Rasmussen and Williams 2006).

The key point of this article is the distribution of $\pi(z_\theta(x_1), \dots, z_\theta(x_n)|\theta)$ must account for the definition of the parameter from Section 2 to avoid identifiability problems. This article defines a new approach to creating the distribution of $\pi(z_\theta(x_1), \dots, z_\theta(x_n)|\theta)$ based on an orthogonality condition which will be illustrated in Section 3.2 and justified in Sections 3.3 and 3.4.

Let $r(\cdot, \cdot)$ represent any covariance function on $X \times X$. If one accepts assumptions (i), (ii), and (iii) with loss L_{L^2} , then the suggestion for the prior distribution on the bias is a Gaussian process with mean zero and a covariance function $\sigma^2 r_\theta(\cdot, \cdot)$, $\sigma^2 > 0$, with

$$r_\theta(x, x') = r(x, x') - h_\theta(x)^\top H_\theta^{-1} h_\theta(x'), \tag{6}$$

where $h_\theta(x)$ is the $p \times 1$ vector

$$h_\theta(x) = \int_X D^{(0,1)} f(\xi, \theta) r(x, \xi) d\xi$$

and H_θ is the $p \times p$ matrix

$$H_\theta = \int_X \int_X D^{(0,1)} f(\xi', \theta) \{D^{(0,1)} f(\xi, \theta)\}^\top r(\xi', \xi) d\xi' d\xi.$$

The notation $D^{(0,1)} f(\xi, \theta)$ represents the vector valued gradient of $f(\cdot, \cdot)$ with respect to the second argument. If the loss is $L_{W_k^2(\mu)}$ from Section 2.3, then

$$h_\theta(x) = \sum_{\|\alpha\|_{L^1} \leq k} \int_X D^{(\alpha,1)} f(\xi, \theta) D^{(0,\alpha)} r(x, \xi) d\mu(\xi)$$

and

$$H_\theta = \sum_{\|\alpha\|_{L^1} \leq k} \sum_{\|\alpha'\|_{L^1} \leq k} \int_X \int_X D^{(\alpha',1)} f(\xi', \theta) \{D^{(\alpha,1)} f(\xi, \theta)\}^\top \times D^{(\alpha',\alpha)} r(\xi', \xi) d\mu(\xi') d\mu(\xi).$$

The notation $D^{(\alpha,1)} f(\xi, \theta)$ represents the vector valued gradient of $D^{(\alpha,0)} f(\cdot, \cdot)$ with respect to the second argument. The notation $D^{(\alpha,\alpha')} r(\cdot, \cdot)$ represents the mixed partial derivative of $r(\cdot, \cdot)$ with α for the first d elements and α' for the last d elements.

Now consider the posterior distribution of the parameter and the bias function. Let \mathbf{f}_θ be an $n \times 1$ vector where the i th element is $f(x_i, \theta)$. Given θ , the vectorization of $\{y(x_1), \dots, y(x_n)\}$ is multivariate normal with mean \mathbf{f}_θ and covariance matrix $\sigma^2 \mathbf{R}_\theta$, where \mathbf{R}_θ is the $n \times n$ matrix with elements $r_\theta(x_i, x_j)$. Accounting for these facts and straightforwardly integrating (5),

$$\pi(\theta|\mathbf{Y}) \propto \frac{\pi(\theta)}{|\sigma^2 \mathbf{R}_\theta + v\mathbf{I}|^{1/2}} \times \exp \left\{ -\frac{1}{2} (\mathbf{Y} - \mathbf{f}_\theta)^\top (\sigma^2 \mathbf{R}_\theta + v\mathbf{I})^{-1} (\mathbf{Y} - \mathbf{f}_\theta) \right\},$$

where $|A|$ indicates the determinant of the matrix A . Now consider the posterior distribution of the bias function at some arbitrary point x_0 . Given θ , using z_θ 's prior,

$$\pi(z_\theta(x_0)|\mathbf{Y}, \theta) = \mathcal{N}\{a_z(x_0, \theta), b_z(x_0, \theta)\},$$

where $\mathcal{N}(a, b)$ stands for the normal distribution with mean a and variance b ,

$$a_z(x_0, \theta) = \mathbf{r}_\theta(x_0)^\top \left(\mathbf{R}_\theta + \frac{v}{\sigma^2} \mathbf{I} \right)^{-1} (\mathbf{Y} - \mathbf{f}_\theta)$$

and

$$b_z(x_0, \theta) = \sigma^2 r_\theta(x_0, x_0) - \sigma^2 \mathbf{r}_\theta(x_0)^\top \left(\mathbf{R}_\theta + \frac{v}{\sigma^2} \mathbf{I} \right)^{-1} \mathbf{r}_\theta(x_0).$$

Here, $\mathbf{r}_\theta(x_0)$ is understood to be the $n \times 1$ sized vector with i th element $r_\theta(x_0, x_i)$. One can then find the posterior distribution of the bias function by integrating out θ , that is

$$\begin{aligned} \pi(z(x_0)|\mathbf{Y}) &= \int_{\Theta} \pi(z(x_0)|\mathbf{Y}, \theta) \pi(\theta|\mathbf{Y}) d\theta, \\ &\propto \int_{\Theta} \frac{1}{\sqrt{b_z(x_0, \theta)}} \exp \left\{ -\frac{(z_\theta(x_0) - a_z(x_0, \theta))^2}{2b_z(x_0, \theta)} \right\} \\ &\quad \times \pi(\theta|\mathbf{Y}) d\theta. \end{aligned}$$

Typically, one can approximate this via Monte Carlo, where draws from the posterior distribution of θ are used to approximate the integral with a sum. Approximate samples of the posterior distribution of θ can be found with Markov chain Monte Carlo approaches such as the Metropolis-Hastings sampler (Gelman et al. 2014, Chap. 11).

A major obstacle in practice is the exact knowledge of σ^2 . If one cannot conjecture σ^2 a priori, the value of σ^2 can be handled through another level of a Bayesian model. Place a prior on σ^2 of $\pi(\sigma^2) \propto 1/\sigma^2$, which is the improper Jeffreys' prior. Given θ and $z_\theta(\cdot)$, the vectorization of $\{z_\theta(x_1), \dots, z_\theta(x_n)\}$, the posterior distribution of σ^2 is given by an inverse gamma distribution with shape $n/2$ and rate $\mathbf{z}_\theta^\top \mathbf{R}_\theta^{-1} \mathbf{z}_\theta / 2$. This conditional distribution then allows us to use the following Markov chain Monte Carlo sampler called a Gibbs sampler (Geman and Geman 1984):

- Draw θ from its posterior distribution given most recent draw of σ^2 . Draw z_θ from its posterior distribution given the most recent draws of θ and σ^2 .

- Draw σ^2 from its posterior distribution given most recent draws of θ and z_θ .

These steps can be iterated many times and eventually θ and σ^2 will be close to draws from their joint posterior.

Kennedy and O'Hagan (2001a) do not incorporate the definition of θ in (iii) into their posterior, which leads to the problems that will be illustrated in the next subsection. Thus, the prior on $z_\theta(\cdot)$ Kennedy and O'Hagan recommended as simply a Gaussian process with mean zero and covariance function $\sigma^2 r(\cdot, \cdot)$. Letting \mathbf{R} be defined the same as \mathbf{R}_θ , their posterior of θ is

$$\begin{aligned} \pi(\theta|\mathbf{Y}) &\propto \frac{\pi(\theta)}{|\sigma^2 \mathbf{R} + v\mathbf{I}|^{1/2}} \\ &\quad \times \exp \left\{ -\frac{1}{2} (\mathbf{Y} - \mathbf{f}_\theta)^\top (\sigma^2 \mathbf{R} + v\mathbf{I})^{-1} (\mathbf{Y} - \mathbf{f}_\theta) \right\}. \end{aligned}$$

Thus the only difference lies in our covariance matrix $\sigma^2 \mathbf{R}_\theta$ versus their $\sigma^2 \mathbf{R}$. As will be shown, this replacement is important because by defining the parameter as done in assumption (iii), $\sigma^2 \mathbf{R}_\theta$ is a better covariance for the bias conditioned on θ compared to $\sigma^2 \mathbf{R}$.

3.2. Illustration of the Orthogonality Condition

Consider an input space that consists of two points, $x_1 = 1$ and $x_2 = 2$. Say that the true output at these two points are $y(x_1) = 2.3$ and $y(x_2) = 3.9$, respectively. Consider the computer model given by

$$f(x, t) = t/4 + 2x + \sin(tx),$$

where $\Theta = [-\pi, \pi]$. There is no value of t for which $f(1, t) = 2.3$ and $f(2, t) = 3.9$, thus model bias exists.

A reasonable prior for the bias of these two possible outputs conditioned on θ , labeled $z_\theta(x_1)$ and $z_\theta(x_2)$, is a bivariate normal distribution with zero mean and a covariance matrix of

$$\frac{1}{25} \begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix}.$$

Thus, the minimizer of the associated reproducing kernel Hilbert space norm (Kennedy and O'Hagan-type definition of the parameter, see Section 2.2) is

$$\begin{aligned} \theta &= \arg \min_{t \in \Theta} (t/4 + 2x_1 + \sin(tx_1) - 2.3)^2 \\ &\quad - 1.5(t/4 + 2x_1 + \sin(tx_1) - 2.3) \\ &\quad \times (t/4 + 2x_2 + \sin(tx_2) - 3.9) \\ &\quad + (t/4 + 2x_2 + \sin(tx_2) - 3.9)^2 \approx -0.108. \end{aligned} \quad (7)$$

Kennedy and O'Hagan would then say that the joint distribution of $y(x_1)$ and $y(x_2)$ conditional on θ has a multivariate distribution normal with means $f(x_1, \theta)$ and $f(x_2, \theta)$ and a covariance matrix given previously.

Consider when the following loss (# 1) is used:

$$(t/4 + 2x_1 + \sin(tx_1) - 2.3)^2,$$

which only considers the discrepancy at x_1 . Since we can modify t for the optimal parameter θ in (iii), the minimizer of the loss

can be found by solving the following nonlinear equation:

$$\theta/4 + 2x_1 + \sin(\theta x_1) = 2.3.$$

This can be solved numerically to find $\theta \approx 0.241$. This naturally implies that we expect, given the true value of θ , that $z_\theta(x_1) = 0$. Thus by defining the parameter explicitly, only biases with $z_\theta(x_1) = 0$ should be considered. We can enforce this by using the covariance for the bias conditioned on θ of

$$\frac{1}{25} \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}.$$

Now, consider when the following loss (# 2) is used:

$$(t/4 + 2x_1 + \sin(tx_1) - 2.3)^2 + (t/4 + 2x_2 + \sin(tx_2) - 3.9)^2.$$

This puts an equal amount of weight toward the bias at each point and $\theta \approx 0.022$. We also find that, by standard optimality conditions,

$$\frac{\partial}{\partial t} (t/4 + 2x_1 + \sin(tx_1) - 2.3)^2 + (t/4 + 2x_2 + \sin(tx_2) - 3.9)^2|_{t=\theta} = 0,$$

and thus

$$1.249z_\theta(x_1) + 2.248z_\theta(x_2) \approx 0.$$

Here, there is some linear constraint on $z_\theta(x_1)$ and $z_\theta(x_2)$. We are implicitly reducing the degrees of freedom of $(z_\theta(x_1), z_\theta(x_2))$ by one because of standard optimization results coupled with assumptions (i), (ii), and (iii). In this case, the vector $(z_\theta(x_1), z_\theta(x_2))$ is orthogonal to the vector (1.249, 2.248). We can enforce this by using the covariance for the bias conditioned on θ of

$$\begin{pmatrix} 1.528 & -0.849 \\ -0.849 & 0.472 \end{pmatrix}.$$

One can then verify that

$$\begin{aligned} \text{var}(1.249z_\theta(x_1) + 2.248z_\theta(x_2)) \\ = 1.249^2 \cdot 1.528 - 2 \cdot 1.249 \cdot 2.248 \cdot 0.849 + 2.248^2 \cdot 0.472 \approx 0. \end{aligned}$$

Three different possible covariances for the bias at x_1 and x_2 given the true θ have been described: one provided by Kennedy and O’Hagan and two others which have a single degree of freedom because the parameter is defined as the minimizer of a loss (either #1 or #2). The latter two are referred to as orthogonal approaches.

Figure 2 illustrates the distribution of the Y_1 and Y_2 conditioned on the knowledge of the true parameter θ . It shows the ellipses of the 95% credible region corresponding to the joint distribution of the outputs using Kennedy and O’Hagan (2001a) and the orthogonal approach. Kennedy and O’Hagan’s suggestion leads a conditional distribution which is broader than needed considering the true θ is known. This is because they do not account for the definition of the parameter. The orthogonal approach leads to a thin conditional distribution that aligns closely with reality. This holds under both the losses # 1 and # 2 even though θ differs. Kennedy and O’Hagan (2001a)

have a noticeably large credible region relative to the conditioning on the true θ . The proposed orthogonal Gaussian process method yields a conditional distribution that agrees with the observations despite having significantly less area included in the credible region.

3.3. Orthogonality in General

In general, the bias function must be orthogonal some aspect of the computer model under assumptions (i), (ii), and (iii).

Let us expand out L_{L^2} , the popular loss studied in Tuo and Wu (2015b):

$$\begin{aligned} L_{L^2}\{y(\cdot) + f(\cdot, t)\} &= \int_X (f(\xi, \theta) + z_\theta(\xi) - f(\xi, t))^2 d\xi, \\ &= \int_X z_\theta(\xi)^2 d\xi + 2 \int_X z_\theta(\xi)(f(\xi, \theta) \\ &\quad - f(\xi, t)) d\xi + \int_X (f(\xi, \theta) - f(\xi, t))^2 d\xi. \end{aligned}$$

Let $f(x, t)$ be differentiable in t at θ for all $x \in X$. This is labeled $D^{(0,1)}f(x, \theta)$ and is a vector in \mathbb{R}^p . Evaluating the gradient of the loss at θ yields

$$-2 \int_X D^{(0,1)}f(\xi, \theta)z_\theta(\xi) d\xi.$$

If θ is in the interior of $\Theta \subset \mathbb{R}^p$, then the gradient of the loss is a vector of zeros at θ . Thus:

Theorem 1. Suppose (i), (ii), and (iii) hold with the loss L_{L^2} and θ located in the interior of $\Theta \subset \mathbb{R}^p$. Suppose that $D^{(0,1)}f(\cdot, \cdot)$ exists and is bounded on $X \times \Theta$. Then,

$$\int_X D^{(0,1)}f(\xi, \theta)\{y(\xi) - f(\xi, \theta)\} d\xi = 0. \tag{8}$$

The direct proof of this result is as follows. Since θ is in the interior of Θ , there exists a constant δ such that $\theta + ae_i$ is in Θ for all $-\delta \leq a \leq \delta$ and $i = 1, \dots, p$, where e_i is the column vector of zeros except for a one at the i th entry. Let the function γ_i be defined as

$$\gamma_i(a) = \frac{L_{L^2}\{y(\cdot) - f(\cdot, \theta + ae_i)\} - L_{L^2}\{y(\cdot) - f(\cdot, \theta)\}}{a}.$$

Under the assumptions of boundedness and the compactness of X , we can use the dominated convergence theorem to show this has a limit point at $a = 0$ of $e_i^T \int_X D^{(0,1)}f(\xi, \theta)\{y(\xi) - f(\xi, \theta)\} d\xi$. By (iii), for all a between 0 and δ the numerator of $\gamma_i(a)$ is nonnegative and thus $\gamma_i(a)$ is nonnegative. Taking the limit of $\gamma_i(\cdot)$ as $a \rightarrow 0$ from the right yields

$$e_i^T \int_X D^{(0,1)}f(\xi, \theta)\{y(\xi) - f(\xi, \theta)\} d\xi \geq 0.$$

Moreover, for all a between $-\delta$ and 0 the numerator of $\gamma_i(a)$ is nonnegative and thus $\gamma_i(a)$ is nonpositive. Thus

$$e_i^T \int_X D^{(0,1)}f(\xi, \theta)\{y(\xi) - f(\xi, \theta)\} d\xi \leq 0,$$

which finalizes the result since this is true for $i = 1, \dots, p$.

The condition (8) can be thought of as orthogonality in the L^2 Hilbert space between each element of $D^{(0,1)}f(\cdot, \theta)$ and $z_\theta(\cdot)$.

The major assumption, aside from the boundedness and differentiability conditions, is that θ is in the interior of Θ . If this is not true, the resulting bias function $z_\theta(\cdot)$ may not be orthogonal to the gradient of the computer model. Consider if Θ is the box $[0, 1]^2$ and $\theta = (0.5, 1)$. The proof can be modified to show that $z_\theta(\cdot)$ must be orthogonal to the first element in the vector $D^{(0,1)} f(\cdot, \theta)$, but not the second. The more general result would be that $z_\theta(\cdot)$ is orthogonal to $h^\top D^{(0,1)} f(\cdot, \theta)$ if there exists a $\delta > 0$ such that $\theta + \alpha h$ is in Θ for all $-\delta \leq \alpha \leq \delta$. Given that none of the considered examples discussed in Section 5 had a solution at a boundary point, this article will not further discuss the case when θ is on the boundary of Θ , though it may be a topic for future work.

This idea can easily be generalized to show that if the loss (4) is used, a similar orthogonality condition emerges on the bias:

Theorem 2. Suppose (i), (ii), and (iii) hold with the loss $L_{W_k^2(\mu)}$ and θ located in the interior of $\Theta \subset \mathbb{R}^p$. For all vectors of nonnegative integers α with $\|\alpha\|_{L^1} \leq k$, suppose that $D^{(\alpha,1)} f(x, \theta)$ exists and is bounded on $X \times \Theta$. Then $z_\theta(\cdot) = y(\cdot) - f(\cdot, \theta)$ must be such that

$$\sum_{\|\alpha\|_{L^1} \leq k} \int D^{(\alpha,1)} f(\xi, \theta) D^{(\alpha)} z_\theta(\xi) d\mu(\xi) = 0. \quad (9)$$

This shows that even if one adopts the idea that the minimizer should not necessarily be the minimizer of the squared error loss (L_{L^2}), there still should be consideration of orthogonality of the bias to the gradient of the computer model, albeit in a different Hilbert space.

If (8) or more generally (9) holds with probability one over the probability measure placed on $z_\theta(\cdot)$, then $z_\theta(\cdot)$ is said to be almost surely orthogonal to the gradient. The next subsection will discuss nontrivial prior distributions on $z_\theta(\cdot)$ that are almost surely orthogonal to the gradient (see (6)).

3.4. Orthogonal Prior Distributions for Bias Functions

Gaussian processes with almost sure orthogonality were not covered in the literature until recently. Plumlee and Joseph (2015) developed a model for Gaussian process regression that contained both a linear mean portion and a random field portion.

The purpose of that article was to reduce potential identifiability problems in universal kriging. In Plumlee and Joseph (2015), $f(x, t)$ was limited to the linear model $t^\top g(x)$, where $g(x)$ is a vector of known functions.

For calibration, $g(\cdot)$ can be exchanged for the gradient $D^{(0,1)} f(x, \theta)$, resulting in (6). The general form of the result of almost sure orthogonality, which follows from the proof of Theorem 1 in Plumlee and Joseph (2015), is:

Theorem 3. Suppose (ii) and for all pairs vectors of nonnegative integers α, α' with $\|\alpha\|_{L^1} \leq k$ and $\|\alpha'\|_{L^1} \leq k$,

$$\int_X \int_X D^{(\alpha',1)} f(\xi', \theta) \{D^{(\alpha,1)} f(\xi, \theta)\}^\top \times D^{(\alpha',\alpha)} r(\xi', \xi) d\mu(\xi') d\mu(\xi)$$

exists and this matrix is finite and positive definite, and $r(\cdot, \cdot)$ is bounded and continuous on $X \times X$. Then if $z_\theta(\cdot)$ is has covariance (6),

$$\sum_{\|\alpha\|_{L^1} \leq k} \int D^{(\alpha,1)} f(\xi, \theta) D^{(\alpha)} z_\theta(\xi) d\mu(\xi) = 0, \quad (10)$$

with probability one.

4. Practical Considerations of Orthogonal Bias Functions

Evaluating $r_\theta(\cdot, \cdot)$ has two major difficulties in practice. The first problem is that the integrals that define $r_\theta(\cdot, \cdot)$ are difficult to solve even in simple cases, see Plumlee and Joseph (2015). Second, $r_\theta(\cdot, \cdot)$ requires the exact specification of the gradient of the computer model. Often times this is not directly observable from a computer model, though there are exceptions (Morris, Mitchell, and Ylvisaker 1993). These issues will be addressed by slightly altering the definition of $r_\theta(\cdot, \cdot)$. The to-be-stated forms of $r_\theta(\cdot, \cdot)$ in (11) and (12) will likely prove to be the most useful expressions in practice.

4.1. Addressing Difficult Integration, $L^2(\mu)$ Case

Revisiting the definition of the loss, note that under assumptions (i), (ii), and (iii) the sample average approximation of the

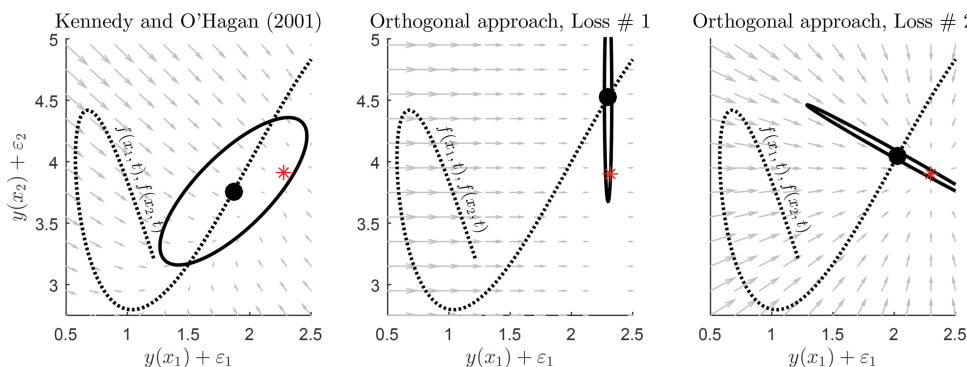


Figure 2. A comparison between conditional distribution of Y_1 and Y_2 given the true θ produced by Kennedy and O'Hagan (2001a) and the orthogonal approach. The background vector field is based on the loss function, where the vectors are oriented toward the optimum. The dashed line represents all possible paired values of $f(x_1, \cdot)$ and $f(x_2, \cdot)$. The dark circle is the point $(f(x_1, \theta), f(x_2, \theta))$ based on assumptions (i), (ii), and (iii) with the three different losses. The ellipsis in each panel represents the 95% creditable region corresponding to the joint distribution of Y_1 and Y_2 given θ , where ϵ_1 and ϵ_2 have prior standard deviation 0.02. The * near (2.3, 3.9) represents a random draw of $(y(x_1) + \epsilon_1, y(x_2) + \epsilon_2)$.

loss can be used (Shapiro, Dentcheva, and Ruszczyński 2014, Chap. 5). The measure μ can be approximated by μ_N , a uniform measure over a discrete set $\{\xi_1, \dots, \xi_N\}$ where each element is drawn independently from μ . The approximate loss is then given by

$$L_{L^2(\mu_N)}(t) := \frac{1}{N} \sum_{i=1}^N (y(\xi_i) - f(\xi_i, t))^2.$$

The minimizer of $L_{L^2(\mu_N)}$ is labeled θ_N . One might wonder how close this minimizer is to θ , the minimizer of $L_{L^2(\mu)}$. Under our conditions, the strong law of large numbers implies $L_{L^2(\mu_N)}$ almost surely converges to $L_{L^2(\mu)}$ as $N \rightarrow \infty$. From Shapiro, Dentcheva, and Ruszczyński's Theorem 5.3:

Proposition 1. Say that (i), (ii), and (iii) are true with the loss $L_{L^2(\mu)}$ and μ is a probability measure. Further suppose $L_{L^2(\mu)}\{y(\cdot) - f(\cdot, t)\}$ is continuous with respect to t and $L_{L^2(\mu_N)}$ has a unique minimizer in Θ labeled θ_N for all N . Then $\theta_N \rightarrow \theta$ with probability one as $N \rightarrow \infty$.

Thus, the minimizer of $L_{L^2(\mu_N)}$ can be made arbitrarily close to the minimizer of $L_{L^2(\mu)}$. Unlike the value of n , which requires physical experimentation, increasing the value of N requires only slightly more computational cost. A practitioner can thus increase N until $L_{L^2(\mu_N)}$ is a suitable replacement for $L_{L^2(\mu)}$. This motivates using a Gaussian process prior for $z_\theta(\cdot)$ with zero mean and covariance function $\sigma^2 r_\theta(\cdot, \cdot)$, $\sigma^2 > 0$ with

$$r_\theta(x, x') = r(x, x') - h_\theta(x)^T H_\theta^{-1} h_\theta(x'),$$

where

$$h_\theta(x) = \frac{1}{N} \sum_{i=1}^N D^{(0,1)} f(\xi_i, \theta) r(x, \xi_i),$$

and

$$H_\theta = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N D^{(0,1)} f(\xi_i, \theta) \{D^{(0,1)} f(\xi_j, \theta)\}^T r(\xi_j, \xi_i).$$

By Theorem 3, $z_\theta(\cdot)$ is almost surely orthogonal to gradient of the computer model on the set ξ_1, \dots, ξ_N .

The above relation is stated with summation notation but this covariance can also be represented using matrices and matrix multiplication. Let $w(x)$ be the $N \times 1$ vector with elements $r(x, \xi_i)$, F_θ be the $N \times p$ matrix with rows $D^{(0,1)} f(\xi_i, \theta)$ and W be an $N \times N$ matrix with elements $r(\xi_i, \xi_j)$. Then

$$r_\theta(x, x') = r(x, x') - w(x)^T F_\theta (F_\theta^T W F_\theta)^{-1} F_\theta^T w(x'). \quad (11)$$

Note also that $r_\theta(\cdot, \cdot)$ is a valid covariance for all values of N such that $F_\theta^T W F_\theta$ is positive definite:

Proposition 2. Suppose F_θ is finite, $r(\cdot, \cdot)$ is a continuous, bounded, positive semidefinite function on $X \times X$ and $F_\theta^T W F_\theta$ is positive definite. Then, $r_\theta(\cdot, \cdot)$ defined in (6) is a positive semidefinite function on $X \times X$.

This result follows Plumlee and Joseph's (2015) Lemma 1.

Practically, the formulation in (6) can be directly evaluated if $D^{(0,1)} f(\xi_1, \theta), \dots, D^{(0,1)} f(\xi_N, \theta)$ are available. The next section covers the case that the gradient is not available.

4.2. Addressing an Estimated $f(\cdot, \cdot)$, $L^2(\mu)$ Case

This section will discuss when the computer model response, $f(x, t)$, and its gradient, $D^{(0,1)} f(x, t)$, are not directly available for the entire region $X \times \Theta$. This commonly happens when the computer model takes a good deal of computational power to evaluate or it can only be run on specialized equipment. Assume that an experiment has been conducted on the computer model. The common approach for inference would say that a prior distribution of a Gaussian process is placed on $f(\cdot, \cdot)$ (Santner, Williams, and Notz 2003). Then, the posterior distribution of $f(\cdot, \cdot)$ given the results from the experiment still a Gaussian process, albeit with a different mean and covariance function.

This article takes the posterior distribution of $f(\cdot, \cdot)$ as a fixed probabilistic definition of the computer model: $f(\cdot, \cdot)$ follows a Gaussian process with mean at (x, t) of $m_f(x, t)$ and a covariance between (x, t) and (x', t') of $c_f((x, t), (x', t'))$. For emphasis, $m_f(\cdot)$ and $c_f(\cdot, \cdot)$ are the mean and covariance functions corresponding to the posterior of $f(\cdot, \cdot)$ after the computer experiment, thus they account for information gained during the experiment. This idea of first fixing the probabilistic definition of $f(\cdot, \cdot)$ is consistent with modularization as introduced by Bayarri et al. (2007). The core concept of this is separately analyzing the posterior of some components and data, here $f(\cdot, \cdot)$ and the observations from the computer experiment, independent of other components and data, here θ and Y . There are many motivations behind adopting modularization as discussed in Liu et al. (2009). Among the explanations in that work, by fixing $f(\cdot, \cdot)$ here, we avoid the corruption of $f(\cdot, \cdot)$ when inferring on θ and $z_\theta(\cdot)$.

Assumption (i), since it does not depend on $f(\cdot, \cdot)$, is unchanged. We must revisit the last two assumptions in this new context where $f(\cdot, \cdot)$ is a random function. We thus replace (ii) with:

- (ii*) $f(\cdot, \cdot)$ follows a Gaussian process with mean $m_f(\cdot, \cdot)$ and covariance function $c_f((\cdot, \cdot), (\cdot, \cdot))$. Suppose that $m_f(\cdot, \cdot)$ exists and is bounded on $X \times \Theta$ and $c_f((\cdot, \cdot), (\cdot, \cdot))$ exists and is bounded on $(X \times \Theta) \times (X \times \Theta)$. Further suppose that $D^{(0,1)} m_f(\cdot, \cdot)$ and $D^{(0,1), (0,1)} c_f((\cdot, \cdot), (\cdot, \cdot))$ exist and are bounded on $X \times \Theta$ and $(X \times \Theta) \times (X \times \Theta)$, respectively.

By Fubini's Theorem under (ii*)

$$E \left\{ \int_X (y(\xi) - f(\xi, t))^2 d\mu(\xi) \right\} = \int_X (y(\xi) - m_f(\xi, t))^2 + v_f(\xi, t) d\mu(\xi),$$

where

$$v_f(\xi, t) = c_f((\xi, t), (\xi, t)),$$

that is, $v_f(\xi, t)$ is the variance of $f(\cdot, \cdot)$ at (ξ, t) . This forms a good loss to use in these circumstances. This loss considers both the difference between the truth, $y(\cdot)$, and the emulated computer model, $m_f(\cdot, t)$, and the uncertainty in that emulation, $v_f(\cdot, t)$.

We now replace (iii) with

(iii*) Suppose that there is some $\theta \in \Theta$ such that for all $t \in \Theta$ such that $t \neq \theta$,

$$\int_X (y(\xi) - m_f(\xi, \theta))^2 + v_f(\xi, \theta) d\mu(\xi) < \int_X (y(\xi) - m_f(\xi, t))^2 + v_f(\xi, t) d\mu(\xi).$$

At a stationary point of the loss in (iii*), under (ii*),

$$\int_X D^{(0,1)} m(\xi, \theta) z_\theta(\xi) d\mu(\xi) = \int_X D^{(0,1)} v_f(\xi, \theta) d\mu(\xi).$$

Thus under this loss, we do not have orthogonality but still have a linear constraint that we can deal with in a similar way.

In these cases, the integral is also not available and thus we use the stochastic average approximation from the previous subsection. Let $w(x)$ and W be as defined in the previous subsection, M_θ be the $N \times p$ matrix with rows $D^{(0,1)} m_f(\xi_i, \theta)$ and Q_θ be the $p \times 1$ matrix

$$Q_\theta = \sum_{i=1}^N D^{(0,1)} v_f(\xi_i, \theta).$$

Pulling all of these ideas through, our suggested prior for $z_\theta(\cdot)$ is a Gaussian process with a mean function

$$\zeta_\theta(\cdot) = w(\cdot)^\top M_\theta (M_\theta^\top W M_\theta)^{-1} Q_\theta$$

and covariance $\sigma^2 r_\theta(\cdot, \cdot)$, $\sigma^2 > 0$, with

$$r_\theta(x, x') = r(x, x') - w(x)^\top M_\theta (M_\theta^\top W M_\theta)^{-1} M_\theta^\top w(x'). \quad (12)$$

Thus, the suggested covariance function for the bias function is the same as the previously stated covariance with a plug-in estimate in place of $D^{(0,1)} f(\cdot, \theta)$. The difference lies in the mean function for $z_\theta(\cdot)$, which was zero before but now incorporates in the uncertainty in $f(\cdot, \cdot)$.

With $r_\theta(\cdot, \cdot)$ is determined, we now closely follow the analysis outlined in Section 3.4 to derive the posterior. Let m_θ be the

vectorization of $\{m_f(x_1, \theta), \dots, m_f(x_n, \theta)\}$, ζ_θ be the vectorization of $\{\zeta_\theta(x_1), \dots, \zeta_\theta(x_n)\}$ and C_θ be the $n \times n$ covariance matrix with elements $c_f((x_i, \theta), (x_j, \theta))$. The analysis then proceeds as in Section 3 with the following substitutions:

$$\begin{aligned} \pi(\theta|Y) &\propto \frac{\pi(\theta)}{|\sigma^2 R_\theta + C_\theta + v I_n|^{1/2}} \exp \left\{ -\frac{1}{2} (Y - m_\theta - \zeta_\theta)^\top \right. \\ &\quad \times (\sigma^2 R_\theta + C_\theta + v I_n)^{-1} (Y - m_\theta - \zeta_\theta) \left. \right\}, \\ a_z(x_0, \theta) &= \zeta_\theta(x_0) + r_\theta(x_0)^\top \\ &\quad \times \left(R_\theta + \frac{1}{\sigma^2} C_\theta + \frac{v}{\sigma^2} I \right)^{-1} (Y - m_\theta - \zeta_\theta) \end{aligned}$$

and

$$\begin{aligned} b_z^2(x_0, \theta) &= \sigma^2 r_\theta(x_0, x_0) - \sigma^2 r_\theta(x_0)^\top \\ &\quad \times \left(R_\theta + \frac{1}{\sigma^2} C_\theta + \frac{v}{\sigma^2} I \right)^{-1} r_\theta(x_0). \end{aligned}$$

5. Illustrative Examples

5.1. Pedagogical Example

Say $X = [0, 1]$, $f(x, t) = tx$, and $y(x) = 4x + x \sin 5x$ and thus model bias exists. In the left panel of Figure 3, the loss functions L_{L^2} and $L_{W_1^2}$ are plotted as a function of the possible parameter values. For comparison sake, the loss that corresponds to the reproducing kernel Hilbert space associated with the bias's prior distribution, discussed in Section 2.2, is plotted as well. The losses L_{L^2} and $L_{W_1^2}$ behave differently, with the minimizer of the former being around 3.6 and the latter being around 3.2. The behavior of the loss $L_{W_1^2}$ is similar to the reproducing kernel Hilbert space loss.

Say we conduct an experiment where responses are observed at $\{0.00, 0.05, 0.10, \dots, 0.80\}$ and generate Y_i as $y(x_i)$ plus a normally distributed variable with zero mean and variance $v = 0.02^2$. The prior density on θ is normal with mean zero and unit variance.

If we use the proposed method with $\sigma^2 = 0$, this produces inference under the assumption of no bias, we can derive the posterior distribution of θ as normal. Consider the method of

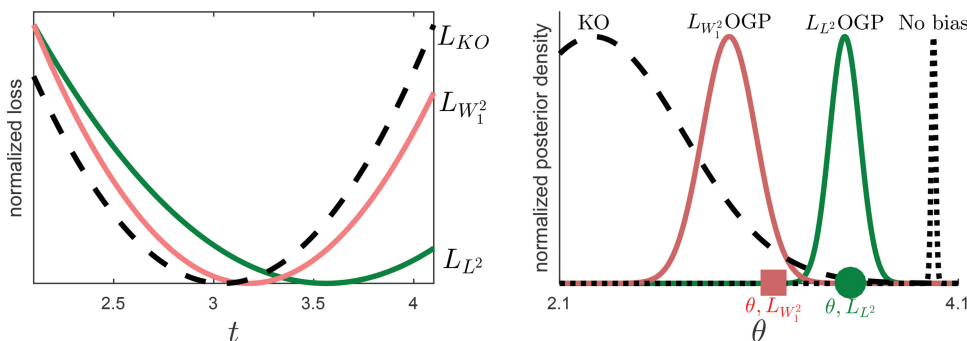


Figure 3. Graphs corresponding to Section 5.1. The left panel is $L_{L^2}(y(\cdot) - f(\cdot, t))$, $L_{W_1^2}(y(\cdot) - f(\cdot, t))$, and the loss attributed to Kennedy and O'Hagan (RKHS loss) and versus t . The right panel shows the posterior density using the four methods, where the solid lines are the orthogonal approach with the two losses, the long dashes represent Kennedy and O'Hagan's posterior and the short dashes represent the posterior if model bias is ignored. The square is the minimizer of $L_{W_1^2}$ and the circle is the minimizer of L_{L^2} .

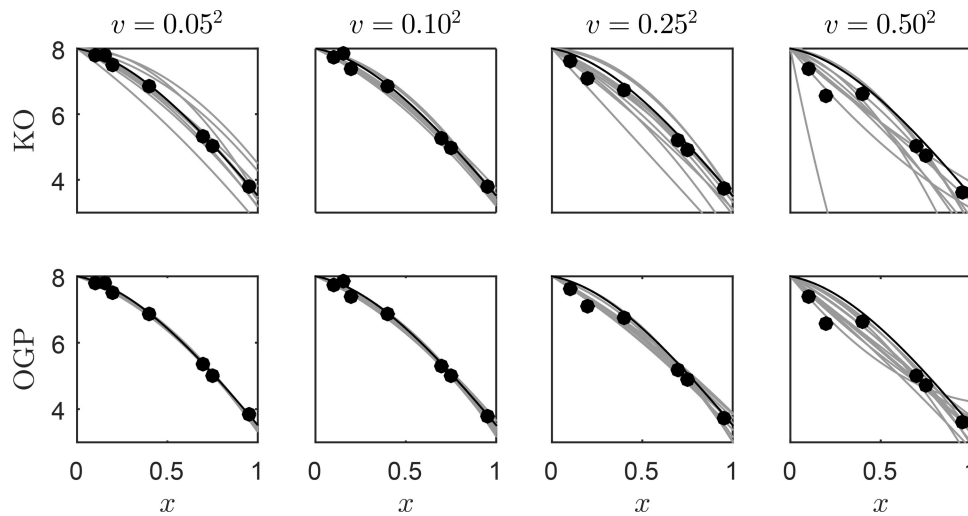


Figure 4. Posterior draws of $f(\cdot, \theta)$ along with the observations and true function for the ball example in Section 5.2. The true, unknown function is the solid, dark line. The dots are located at the observations. The lighter solid lines represent ten draws from the posterior distribution of $f(\cdot, \theta)$ using one of the four values of v with either Kennedy and O'Hagan's bias function priors (top) or the proposed orthogonal Gaussian process bias function priors (bottom) with L_{L^2} .

Kennedy and O'Hagan (2001a) with the covariance

$$r(x, x') = \sigma^2(1 + |x - x'|/\psi)(-|x - x'|/\psi). \quad (13)$$

This is a special case of the Matérn covariance, widely used in the analysis of computer experiments, where $\psi = 1/2$ is chosen as our lengthscale parameter. The posterior θ is also normal for the posterior of Kennedy and O'Hagan (2001a). To find r_θ for the orthogonal approach, we use symbolic integration packages to find $h_\theta(\cdot)$ and H_θ in both the case of L_{L^2} and $L_{W_1^2}$ and with a renormalization step so that the average prior variance over the input space is the same as a stationary correlation function. After completing these two steps, when the loss is L_{L^2} ,

$$r_\theta(x, x') = 5.19r(x, x') - 23.52h(x)h(x'),$$

$$h(x) = 2x + \frac{2x + 3}{4} \exp(-2x) + \frac{6x - 13}{4} \exp(2x - 2)$$

and when the loss is $L_{W_1^2}$,

$$r_\theta(x, x') = 1.27r(x, x') - 0.78h(x)h(x'),$$

$$h(x) = 2x - \frac{6x + 1}{4} \exp(-2x) - \frac{2x + 1}{4} \exp(2x - 2).$$

The right panel of Figure 3 shows the posterior distribution of the parameter under each framework. The posterior when bias is ignored is quite concentrated at a value that is beyond the actual minimizer of the L^2 norm of the bias. This is because we observed data only up to 0.8. There is a great deal of bias that occurs when the input is between 0.8 and 1 which is ignored if the prior on the bias is zero. The posterior distribution of θ using the orthogonal Gaussian process method is properly centered around the appropriate minimizer, with a little broader distribution to account for the uncertainty in the bias. The Kennedy and O'Hagan posterior is not centered properly, mostly existing to the left of the region of interest. This is despite the fact that the parameter they implicitly defined exists around 3.1 (see the left panel of Figure 3).

5.2. Falling Ball Example, revisited

Consider the falling ball example from the first section with the initial vertical height a known value of eight. The difference between Y_i and $y(x_i)$ is drawn independently from a normal distribution with mean zero and a known variance v .

The objective in this subsection is to compare the posterior distribution of $f(x, \theta)$ using Kennedy and O'Hagan's bias function priors and the proposed orthogonal Gaussian process bias function priors. Let the loss be L_{L^2} in this example and $r(\cdot, \cdot)$ be (13) with $\psi = 1$. In this example, σ^2 is handled using the Bayesian approach in Section 3.1. The observed responses are at $\{0.10, 0.15, 0.20, 0.40, 0.70, 0.75, 0.95, 1.00\}$.

Figure 4 shows draws from the posterior distribution of $f(\cdot, \theta)$ along with the true function with various values of v . Kennedy and O'Hagan's method results in a large posterior variance of $f(\cdot, \theta)$. Intuitively, one would expect a small value of the noise variance v would result in smaller variance in the posterior of $f(\cdot, \theta)$. For Kennedy and O'Hagan's method, this is not the case. Even when v is small at 0.05^2 , there is still a large posterior variance of $f(\cdot, \theta)$. In contrast, posterior variance of $f(\cdot, \theta)$ using the proposed method logically shrinks as the noise gets small. When v is 0.05^2 , the posterior $f(\cdot, \theta)$ is hard to distinguish from the true $y(x)$.

5.3. Ion Channel Example

This subsection will detail analysis of a small subset of real data analyzed in Plumlee, Joseph, and Yang (2016), which used Kennedy and O'Hagan's prior for the bias. That dataset contains the results from whole cell voltage clamp experiments on the sodium ion channels of cardiac cell membranes. These experiments return the current (response) needed to maintain a fixed membrane potential over time (input). We use a log scale for this input, that is, x is the log of time, for better presentation because most of the activity in the response is present in the beginning of the experiment. The observations will be the output of a single voltage clamp experiment on a cell. Figure 5 shows a series of observations over the course of an experiment at a fixed

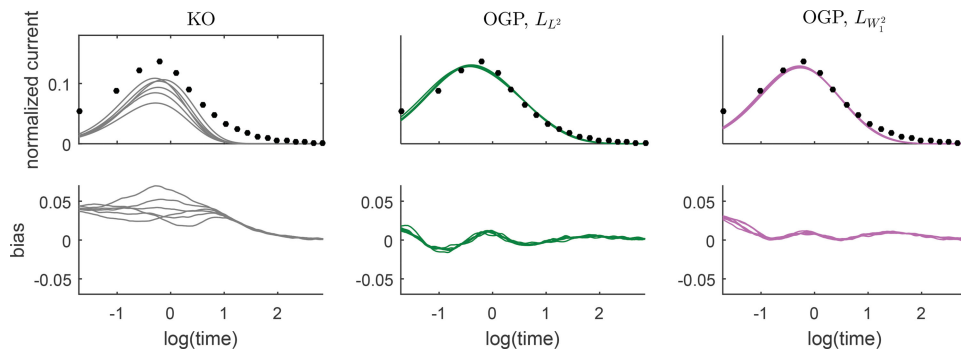


Figure 5. The top panels are posterior draws of $f(\cdot, \theta)$ for the ion channel example in Section 5.3 using Kennedy and O'Hagan's prior on the bias functions (left) and the proposed method (middle with L_{L^2} and right with $L_{W_1^2}$). The dots are located at the observations. The bottom panels are posterior draws of the bias using the same framework as the top panels.

membrane potential of -35mV . The response is normalized by a negative constant for clarity.

The computer model is the classic Markov model for sodium ion channels, see, for example, Clancy and Rudy (1999). This article will not detail the motivation behind the model, but the normalization mentioned earlier allows implies

$$f(x, t) = e_1^\top \exp(\exp(x)A(t))e_4,$$

where the first \exp implies the matrix exponential, e_i is the column vector of zeros with the i th element one and

$$A(t) = \begin{bmatrix} -t_2 - t_3 & t_2 & 0 & 0 \\ t_1 & -t_1 - t_2 & 0 & 0 \\ 0 & t_2 & -t_1 - t_2 & t_1 \\ 0 & 0 & t_2 & -t_1 \end{bmatrix}.$$

The parameter is three dimensional, $(\theta_1, \theta_2, \theta_3)$. The gradient, $D^{(0,1)}(x, \theta)$, can be evaluated directly in this case, as can $D^{(1,1)}(x, \theta)$. Take $r(\cdot, \cdot)$ as given in (13) with $\psi = 1$. The value of v is fixed at 0.001^2 and σ^2 at 0.02^2 . For our method, $v = 0.001^2$ and the average prior variance of $z_\theta(\cdot)$ over X is fixed at 0.02^2 for comparison. Since $f(x, \theta)$ is a nonlinear and θ is three dimensional, the Metropolis-Hastings algorithm is used to sample from the posterior of θ for both the Kennedy and O'Hagan's method and the proposed method.

Figure 5 graphs six draws from the posterior of $f(\cdot, \theta)$ using both the traditional and proposed approach to the bias function (with two losses). The posterior distribution of $f(\cdot, \theta)$ does not cover the data in either approach. This is expected because there is some model bias. But it is apparent that Kennedy and O'Hagan's posterior of $f(\cdot, \theta)$ is further from the data compared to the proposed posterior. Kennedy and O'Hagan's posterior variance of $f(\cdot, \theta)$ is also much larger than one would expect given the information available. These two issues are resolved when the proposed bias function priors are used.

This also illustrates the different conclusions one can get by using different loss functions. When $L_{W_1^2}$ is used, there is a penalty placed on the discrepancy between the derivative of the response and the derivative of the actual response. The loss L_{L^2} has no such penalty. The posterior draws from the bias when using L_{L^2} have smaller magnitude than the bias when using $L_{W_1^2}$, but have more oscillations across the input. This illustrates the effect the loss plays in estimation. It is up to a user to decide

which parameter/bias pair they are interested in when deciding a loss.

6. Conclusions and Discussion

This article introduces a technique for Bayesian calibration when the computer model does not align perfectly with reality. This article assumes that the parameter is defined as the minimizer of a loss. Under this assumption, it was shown that the bias function should be orthogonal to the gradient of the computer model. Previously suggested bias function priors do not have this property. This work details the construction of bias function priors that do have this property.

The parameter in this work is defined as an artifact of the goal we would like to achieve, for example, minimizing the squared error difference between the computer model and nature's function. Other options include minimizing the squared error difference between the computer model's derivative and the derivative of nature's function. Defining the parameter based on these type of criteria instead of from a probability distribution is not uncommon to other fields, for example see Bouchard and Triggs (2004) and Liang and Jordan (2008). From the perspective of a statistical researcher, this definition allows for the evaluation and comparison of different calibration methods.

There are some additional computational costs associated with using the proposed bias functions as opposed to the bias structure suggested by Kennedy and O'Hagan. The major cost to compute the posterior of θ and $z_\theta(\cdot)$ will be the inversion of an $n \times n$ matrix, where n is the number of physical observations. This exists in both Kennedy and O'Hagan and the proposed approach. In Kennedy and O'Hagan, only $n^2/2$ evaluations are of $r(\cdot, \cdot)$ required during likelihood computation. When the ideas from Section 4.1 and Section 4.2 are used, $r(\cdot, \cdot)$ is evaluated $(n + N)^2/2$ times. Recall that N is the number of samples drawn uniformly from the input space to approximate the integrals. Thus, if N is much larger than n , there could be a significant amount of additional computation. A repeatable strategy for having N be large enough such that θ_N is close to θ without computational issues was not found by the author. The other additional cost is the inversion of a $p \times p$ matrix $M_\theta^\top W M_\theta$, which is relatively cheap if p is small, but may be a consideration if p is very large. With the impact of these values delineated, consider the practical ranges of each value. Typically there are fewer parameters than observations,

thus $p \ll n$, and there is typically not enough observations to fill the space, thus $n \ll N$. The major additional cost will then be evaluating $r(\cdot, \cdot)$ and additional $N^2/2$ times, which is a new cost to the proposed method. In Section 4.2, the computation of the predictive mean and variance for the computer model also requires inverting a square matrix with size equal to the number of computer model observations, which is also implicitly required for Kennedy and O'Hagan's posterior. If specialized computer experiment designs, such as grids (Kennedy and O'Hagan 2001b) or sparse grids (Plumlee 2014), are used, the cost of the computer model inference can be significantly reduced.

Supplementary Materials

The supplementary materials contain the MATLAB code used to generate the figures.

Acknowledgment

The authors would like to acknowledge the support from the National Science Foundation (CMMI-1266025); thank Andrew R. Ednie and Eric S. Bennett for sharing the heart cell dataset; and thank Daniel W. Apley, Robert B. Gramacy, David M. Higdon, V. Roshan Joseph, Barry L. Nelson, Rui Tuo, C. F. Jeff Wu, two anonymous reviewers and an anonymous associate editor for their comments toward improving this work.

References

- Bates, D. M., and Watts, D. G. (1988), *Nonlinear Regression: Iterative Estimation and Linear Approximations*, Wiley Online Library, available at onlinelibrary.wiley.com/doi/10.1002/9780470316757.fmatter/pdf. [1274]
- Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-H., and Tu, J. (2007), "A Framework for Validation of Computer Models," *Technometrics*, 49, 138–154. [1275,1281]
- Bouchard, G., and Triggs, B. (2004), "The Tradeoff Between Generative and Discriminative Classifiers," in *16th IASC International Symposium on Computational Statistics (COMPSTAT'04)*, pp. 721–728. [1284]
- Clancy, C. E., and Rudy, Y. (1999), "Linking a Genetic Defect to its Cellular Phenotype in a Cardiac Arrhythmia," *Nature*, 400, 566–569. [1284]
- Cressie, N. A. C. (1993), *Statistics for Spatial Data*, New York: Wiley. [1277]
- Farah, M., Birrell, P., Conti, S., and Angelis, D. D. (2014), "Bayesian Emulation and Calibration of a Dynamic Epidemic Model for A/H1N1 Influenza," *Journal of the American Statistical Association*, 109, 1398–1411. [1275]
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014), *Bayesian Data Analysis* (3rd ed.), Boca Raton, FL: CRC Press. [1278]
- Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741. [1278]
- Gneiting, T., and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378. [1276]
- Goldstein, M., and Rougier, J. (2004), "Probabilistic Formulations for Transferring Inferences from Mathematical Models to Physical Systems," *SIAM Journal on Scientific Computing*, 26, 467–487. [1275]
- Gramacy, R. B., Bingham, D., Holloway, J. P., Grosskopf, M. J., Kuran, C. C., Rutter, E., Trantham, M., Drake, P. R., et al. (2015), "Calibrating a Large Computer Experiment Simulating Radiative Shock Hydrodynamics," *The Annals of Applied Statistics*, 9, 1141–1168. [1275]
- Han, G., Santner, T. J., and Rawlinson, J. J. (2009), "Simultaneous Determination of Tuning and Calibration Parameters for Computer Experiments," *Technometrics*, 51, 464–474. [1275,1276]
- Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D. (2004), "Combining Field Data and Computer Simulations for Calibration and Prediction," *SIAM Journal on Scientific Computing*, 26, 448–466. [1275,1276]
- Joseph, V. R., and Melkote, S. N. (2009), "Statistical Adjustments to Engineering Models," *Journal of Quality Technology*, 41, 362–375. [1275]
- Joseph, V. R., and Yan, H. (2015), "Engineering-Driven Statistical Adjustment and Calibration," *Technometrics*, 57, 257–267. [1275]
- Kennedy, M. C., and O'Hagan, A. (2001a), "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society, Series B*, 63, 425–464. [1274,1275,1276,1277,1278,1283]
- (2001b), "Supplementary Details on Bayesian Calibration of Computer Models," Tech. Rep., Internal Report. Available at <http://www.shef.ac.uk/~st1ao/ps/calsup.ps>. [1285]
- Kleiber, W., Sain, S. R., Heaton, M. J., Wiltberger, M., Reese, C. S., and Bingham, D. (2013), "Parameter Tuning for a Multi-Fidelity Dynamical Model of the Magnetosphere," *The Annals of Applied Statistics*, 7, 1286–1310. [1275]
- Liang, P., and Jordan, M. I. (2008), "An Asymptotic Analysis of Generative, Discriminative, and Pseudolikelihood Estimators," in *Proceedings of the 25th International Conference on Machine Learning*, ACM, pp. 584–591. [1284]
- Liu, F., Bayarri, M., Berger, J., et al. (2009), "Modularization in Bayesian Analysis, With Emphasis on Analysis of Computer Models," *Bayesian Analysis*, 4, 119–150. [1281]
- Morris, M. D., Mitchell, T. J., and Ylvisaker, D. (1993), "Bayesian Design and Analysis of Computer Experiments: Use of Derivatives in Surface Prediction," *Technometrics*, 35, 243–255. [1280]
- Plumlee, M. (2014), "Fast Prediction of Deterministic Functions Using Sparse Grid Experimental Designs," *Journal of the American Statistical Association*, 109, 1581–1591. [1285]
- Plumlee, M., and Joseph, V. R. (2015), "Orthogonal Gaussian Process Models," *Under review*. [1280,1281]
- Plumlee, M., Joseph, V. R., and Yang, H. (2016), "Calibrating Functional Parameters in the Ion Channel Models of Cardiac Cells," *Journal of the American Statistical Association*, 111, 500–509. [1283]
- Qian, P. Z., and Wu, C. J. (2008), "Bayesian Hierarchical Modeling for Integrating Low-Accuracy and High-Accuracy Experiments," *Technometrics*, 50, 192–204. [1275]
- Rasmussen, C. E., and Williams, C. (2006), *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press. [1277]
- Santner, T. J., Williams, B. J., and Notz, W. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer Science & Business Media. [1281]
- Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014), *Lectures on Stochastic Programming: Modeling and Theory*, Philadelphia, PA: SIAM. [1281]
- Storlie, C. B., Lane, W. A., Ryan, E. M., Gattiker, J. R., and Higdon, D. M. (2015), "Calibration of Computational Models With Categorical Parameters and Correlated Outputs via Bayesian Smoothing Spline ANOVA," *Journal of the American Statistical Association*, 110, 68–82. [1275]
- Tuo, R., and Wu, C. F. J. (2015a), "Efficient Calibration for Imperfect Computer Models," *The Annals of Statistics*, 43, 2331–2352. [1276]
- Tuo, R., and Wu, C. F. J. (2015b), "A Theoretical Framework for Calibration in Computer Models: Parametrization, Estimation and Convergence Properties," *arXiv:1508.07155*. [1275,1276,1279]
- Wang, S., Chen, W., and Tsui, K.-L. (2009), "Bayesian Validation of Computer Models," *Technometrics*, 51, 439–451. [1275]
- Wendland, H. (2004), *Scattered Data Approximation*, Cambridge, UK: Cambridge University Press. [1276]