ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2019)
81, *Part* 3, *pp.* 519–545

Computer model calibration with confidence and consistency

Matthew Plumlee

Northwestern University, Evanston, USA

[Received November 2017. Revised January 2019]

Summary. The paper proposes and examines a calibration method for inexact models. The method produces a confidence set on the parameters that includes the best parameter with a desired probability under any sample size. Additionally, this confidence set is shown to be consistent in that it excludes suboptimal parameters in large sample environments. The method works and the results hold with few assumptions; the ideas are maintained even with discrete input spaces or parameter spaces. Computation of the confidence sets and approximate confidence sets is discussed. The performance is illustrated in a simulation example as well as two real data examples.

Keywords: Frequentist; Inverse problem; Model discrepancy; Model inadequacy; Model misspecification; Non-linear regression; Uncertainty quantification

1. Introduction

Computer models are complex representations of systems that are implemented in computer code: a term dating back to at least Sacks *et al*. (1989). Calibration leverages observations of input–output pairs from the real system to make statistical adjustments that align the model with reality via a parameter. The goal is to find the parameter that would best represent nature's output across all possible inputs. Statistical adjustments are primarily used to combat two issues. First, the observations are noisy perturbations of nature's response. Second, the inputs in the data set are only a subset of all potential inputs and the generation or choice of the inputs may not be under the control of the statistician.

The wrinkle in the calibration problem is that most computer models are inexact representations of nature, i.e. no value of the parameter will exactly align the computer model with nature's system. This is the primary distinction between statistical models and computer models; classical statistical inference assumptions for non-linear functions assume model exactness (Box and Coutie, 1956; Beale, 1960). Statistical models guard against an incorrect model by augmenting it with extra terms if it is found to be inexact. Computer models are used because they exhibit reasonable behaviour and trends, even if the model is slightly inexact. When there are very little data, such as the example in Section 9.2 with nine data points in a four-dimensional space, it can become difficult to trust a data-driven statistical model in comparison with a well-engineered computer model. The user would like to leverage the coherent behaviour of a model with a parameter that best represents the system. Because we have collected stochastic data at limited inputs, we

Address for correspondence: Matthew Plumlee, McCormick School of Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208-3109, USA.
E-mail: mplumlee@northwestern.edu

 1369–7412/19/81519

cannot expect to identify this parameter exactly; so in this paper we aim to find some good region or set of parameters to understand both the location of the best parameter and the uncertainty.

Modern calibration methods estimate both the parameter and the discrepancy between nature's response and the computer model. The two primary thrusts of research operate along the major statistical justifications (Efron, 2005): Bayesian and large sample frequentist. Kennedy and O'Hagan (2001) proposed to infer the parameter via a posterior distribution by placing a prior distribution on both the parameter and the discrepancy. This framework has been widely used (Goldstein and Rougier, 2006; Bayarri *et al*., 2007b; Higdon *et al*., 2008; Vernon *et al*., 2014; Plumlee *et al*., 2016). The statistical guarantees are from a Bayesian perspective. This brand of Bayesian calibration is criticized because of the relationship between the discrepancy's prior and the posterior of the parameter because of confounding between the discrepancy and the parameter. Tuo and Wu (2016) formally described this, demonstrating that the posterior mode of the parameter goes to some value that is dependent on the discrepancy's prior in a frequentist setting. Plumlee (2017) included a potential fix but offered only Bayesian support and did not confirm that the problems that were uncovered by Tuo and Wu (2016) are alleviated. This fix also requires the use of the derivative of the computer model, which can be burdensome in some cases and impossible in others, such as the discrete parameter case of Section 9.2.

The second thrust of research on this topic came a little later in the form of more traditional mechanisms. Joseph and Melkote (2009) outlined some principals of a frequentist framework, but the creation of replacement mechanisms with large sample statistical guarantees is a more recent development. Tuo and Wu (2015) and Wong *et al*. (2017) both advocated methods that perform well in large samples in a frequentist setting. These methods both produce a confidence set for the parameter with large sample frequentist justification. Specifically, although estimating both the discrepancy and the parameter, these methods leverage arguments of semiparametric efficiency to conclude that we can ignore the uncertainty in the discrepancy when calculating confidence sets for the parameter. More details will be provided later in this paper, but intuitively ignoring estimation of the discrepancy seems to create overly small confidence sets for the parameter.

This paper proposes a framework for creating a confidence set based on a confidence level α with two key properties:

(a) *confidence*; the probability that the best parameter is excluded is at most α for all sample sizes and for any generation of inputs;
(b) *consistency*; the probability that a suboptimal parameter is excluded goes to 1 as the number of samples goes to ∞ if the collected inputs eventually cover the input space.

For reasons that will become clear, the confidence set that produces these properties is termed the conservative and consistent set. The coincidence of these two properties is not present in existing Bayesian or large sample frequentist methods. The *caveat* is that, to build the conservative and consistent set, the norm of the discrepancy at the best parameter must be bounded by a value that is either user specified or estimated from the data. Aside from this assumption, the confidence and consistency results follow from mild regularity conditions. The conclusions hold for discrete or continuous input spaces as well as discrete and continuous parameter spaces. This paper also discusses the efficient learning of this bound from the data if we do not feel comfortable bounding it *a priori*.

This paper will outline the calibration setting in Section 2. Three sets are introduced in Section 3; the last two are shown to have the desired confidence property. Section 4 details the numerical procedure for building these sets, with some theoretical explanation. The consistency arguments are in Section 5. Section 6 compares this approach with other existing methods in a

general setting. Section 7 explores estimating the norm bound by using data. Section 8 contains a numerical simulation that illustrates some problems with competing methods and compares with the proposed set. Sections 9 and 10 close the paper with a pair of real data examples and a brief discussion.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

```
https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-
b-datasets
```

2. Setting and notation

This section describes standard assumptions for the calibration problem and the definition of the best parameter. Calibration, as taken in this paper, is the act of using data collected from the real system to conjecture on the optimal parameter of an inexact model. The only significant departure from standard assumptions is the generalization to a set of optimal parameters *versus* a singleton optimal parameter.

2.1. Data and stochastic assumptions

The data consist of n sets of inputs and scalar outputs represented by

$$(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n).$$

The input space, which is labelled $\mathbb{X}$, contains all the possible inputs of interest; some inputs are included in the data and some are not. The space $\mathbb{X}$ is left purposely vague at this point in the paper; it can be continuous or discrete, bounded or unbounded, closed or open. The points $x_1, \ldots, x_n$ are selected from somewhere in the input space via either a deterministic or a random mechanism. Nature's response $y(x)$ maps an input $x \in \mathbb{X}$ to a scalar. The user-defined computer model $f_\theta(x)$ maps an input $x \in \mathbb{X}$ and a parameter $\theta \in \Theta$ to a scalar. The objective is to find the parameter value that best aligns f_θ with y by leveraging the data.

Each observation Y_i differs from $y(x_i)$ by some random variable $e_i = Y_i - y(x_i)$ and these differences are independent across i and independent from the generation of $x_1, \ldots, x_n$. Consider a fixed confidence level labelled $1 - \alpha$, which is greater than 0 and smaller than 1. Suppose that the distribution of e_i is well understood in the sense that there is some known $q_n(\alpha)$ that is the $(1 - \alpha)$-quantile of the sum of squared values of differences between observed responses and the true function, i.e.

$$\mathbb{P}\left\{ \sum_{i=1}^{n} e_i^2 \leqslant q_n(\alpha) \right\} = 1 - \alpha.$$

The most common example of this is e_is being normal random variables with mean 0 with variance σ^2, where $q_n(\alpha)$ can then be chosen as σ^2 times the $(1 - \alpha)$-quantile of the χ^2-distribution with n degrees of freedom.

2.2. Definition of the parameter

To begin studying the calibration problem, we first need to address a slightly philosophical question. If the model is incorrect for all potential parameters, what does it mean to have a good parameter? Borrowing from the notation of scoring rules (Gneiting and Raftery, 2007), let $l(\cdot, \cdot)$ be a scoring metric that operates on functions that map $\mathbb{X}$ to a scalar. A scoring metric is considered proper if

$$l(y, y) \leqslant l(y, g) \qquad \text{for all } g.$$

It is strictly proper if this inequality holds with equality only if $y(x) = g(x)$ everywhere of interest. Consider the strictly proper scoring metric that was used in Han *et al.* (2009), Tuo and Wu (2015), Plumlee (2017) and Wong *et al.* (2017) of

$$l(g, h) = \int_{\mathbb{X}} \{g(x) - h(x)\}^2 \mathrm{d}\mu(x),$$

where μ is some measure over $\mathbb{X}$. This is a strictly proper scoring metric that weights the relative inputs by some measure μ. This measure ought to come from the modeller's intended use of the model. The common assumption is a uniform measure over the input space, but this is not required. The measure μ is assumed fixed and independent of the data.

Given the scoring metric, the definition of the best parameter θ is such that

$$l(y, f_\theta) \leqslant l(y, f_t) \qquad \text{for all } t \in \Theta. \tag{1}$$

This value cannot directly be evaluated without knowledge of the true function y even if the scoring metric is given. But this construction ensures that the calibration problem is well defined in terms of an oracle. If a calibration scheme is good, it will yield values that meet condition (1).

Definition 1. The optimal parameter set Θ^* is all θ in Θ that meet criterion (1).

This set could be a single point θ^*, as is typically assumed, but this paper considers the general case.

Assumption 1. For the remainder of this paper, assume that $y(x)$ and $f_\theta(x)$ are bounded over $\mathbb{X}$ and Θ and $\mu(\mathbb{X})$ is finite and non-zero.

3. Conservative and consistent sets

This paper discusses confidence sets for the optimal parameters. A confidence set for the calibration problem should be designed such that it provides sufficient coverage.

Definition 2. A confidence set S which is based on the data $(x_i, Y_i)_{i=1}^n$ has sufficient coverage if, for all $\theta \in \Theta^*$,

$$\mathbb{P}(\theta \in S) \geqslant 1 - \alpha,$$

regardless of the method by which $x_1, \ldots, x_n$ were generated, so long as this generation does not depend on $e_1, \ldots, e_n$.

A key point is that coverage applies to all geneses of $x_1, \ldots, x_n$. Importantly, a set with sufficient coverage can be used in conjunction with limited sample sizes n. This concept also applies when an experimental design is used to decide $x_1, \ldots, x_n$ deterministically. Sufficient coverage additionally includes cases where $x_1, \ldots, x_n$ are generated outside the control of the user, i.e. when the generating mechanism for $x_1, \ldots, x_n$ does not match μ, the weighting measure that defines our best parameter.

This section describes three sets and their possession, or lack thereof, of the sufficient confidence property:

(a) *naive set*, a classical set without sufficient coverage but motivates the new confidence sets;
(b) *conservative set*, a new set with sufficient coverage but is, in general, larger than needed;
(c) *conservative and consistent set*, a new set with sufficient coverage that is smaller than the conservative set.

3.1. Naive set NS

We do not have access to $l(y, f_\theta)$ without an oracle to provide nature's function y. An empirical approximation is

$$\hat{l}(\text{data}, f_\theta) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - f_\theta(x_i)\}^2.$$

Thus, in the absence of an oracle, a set of reasonable parameters could be defined as

$$\text{NS} = \left\{\theta \in \Theta \text{ such that } \hat{l}(\text{data}, f_\theta) \leqslant \frac{q_n(\alpha)}{n}\right\}.$$

To provide some historical context, this type of naive set was mentioned in Beale's (1960) work on non-linear confidence sets. Beale briefly criticized the set, noting that it could be empty. This concern is reduced by choosing a sufficiently small α, provided that the model is exact.

If $y(x) = f_\theta(x)$ everywhere when $\theta \in \Theta^*$, then it is obvious that the naive set has sufficient coverage. However, the naive set often has insufficient coverage if the model is inexact. This is because

$$\hat{l}(\text{data}, f_\theta) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 + \frac{1}{n}\sum_{i=1}^{n}\{y(x_i) - f_\theta(x_i)\}^2 + \frac{2}{n}\sum_{i=1}^{n} e_i\{y(x_i) - f_\theta(x_i)\},$$

and the second term tends to be larger than the last term.

Proposition 1. Suppose that, for all $\theta \in \Theta^*$, there is some input x_i in the data such that $y(x_i) \neq f_\theta(x_i)$. Also assume that $e_1, \ldots, e_n$ are independent and identically normally distributed with zero mean and variance σ^2. Then NS does not have sufficient coverage.

We therefore conclude that new approaches are needed to achieve the sufficient coverage property.

3.2. Building a set of reasonable discrepancies

Building the next two sets requires creating a space of functions D that contains $y(\cdot) - f_\theta(\cdot)$ for all $\theta \in \Theta^*$. If Θ^* consists of only a singleton labelled θ^*, then $y - f_{\theta^*}$ is termed the discrepancy in the literature. Because there can be many elements in Θ^* as defined in this work, these can be considered discrepancies. Creating D is spiritually similar to creating a prior distribution on the discrepancy as proposed by Kennedy and O'Hagan (2001).

This paper constructs D by confining the norm of the discrepancy to be less than some value η specified by the user. This norm ought to communicate more than simply the size of the function. One selection of norms that have this property are norms that are associated with a reproducing kernel Hilbert space. Sometimes known as roughness penalties (Wahba, 1978), we refer to Wendland (2004) and Scholkopf and Smola (2001) for a detailed background on reproducing kernel Hilbert spaces.

Given a continuous, positive definite kernel function $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, consider the following space which is dense in the reproducing kernel Hilbert space:

$$G = \left\{g(\cdot) = \sum_{i=1}^{N} \beta_i \kappa(\cdot, x_i) \text{ such that } x_i \in \mathbb{X}, N = 1, 2, \ldots\right\}.$$

The reproducing kernel Hilbert space is equipped with an inner product labelled '$\langle \cdot, \cdot \rangle$' such that, for any two elements in G,

$$g(\cdot) = \sum_{i=1}^{N} \beta_i \kappa(\cdot, x_i)$$

and

$$h(\cdot) = \sum_{i=1}^{N'} \gamma_i \kappa(\cdot, x_i);$$

then

$$\langle g, h \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N'} \beta_i \gamma_j \kappa(x_i, x_j). \tag{2}$$

This inner product has the reproducing property where

$$\langle g, \kappa(\cdot, x) \rangle = g(x). \tag{3}$$

The norm that is induced by this space is $\|g\| = \sqrt{\langle g, g \rangle}$. This norm reflects the size of the discrepancy relative to its chosen representation. Given this description, D can be defined by

$$D = \{d \in G \text{ such that } \|d\| \leqslant \eta\}, \tag{4}$$

where $\eta > 0$ is some constant that controls the size of the set.

When the input space is Euclidean, a common kernel is the power exponential. Other options are available from Santner *et al.* (2003), although kernels are commonly referred to as covariance functions in the literature on computer experiments. The choice of kernel reflects the overall structure and smoothness of the discrepancies that are under consideration.

Assumption 2. For the remainder of this paper assume that $\sup_{x,x'} \kappa(x, x')$ and $\sup_{\theta} \|y - f_\theta\|$ are finite. Also assume that κ is continuous to enable us to employ Mercer's theorem in some of the proofs. Also assume that the norm is such that $\|g\| = 0$ implies that $g(x) = 0$ almost everywhere with respect to μ.

3.3. Conservative set

To construct a set with sufficient coverage requires the acknowledgement and estimation of model discrepancy. This was the key idea behind the influential paper of Kennedy and O'Hagan (2001) and the subsequent references that were mentioned in Section 1.

Consider the set

$$\text{CS} = \left\{ \theta \in \Theta \text{ such that } \min_{d \in D} \hat{l}(\text{data}, f_\theta + d) \leqslant \frac{q_n(\alpha)}{n} \right\},$$

where the empirical approximation to the score function matches the naive set and D is given by definition (4). Because this set will be conservative with respect to the discrepancy, it is termed the conservative set.

Theorem 1. Suppose that $y - f_\theta \in D$ for all $\theta \in \Theta^*$. Then CS has sufficient coverage.

Proof. The condition yields

$$\min_{d \in D} \sum_{i=1}^{n} \{Y_i - f_\theta(x_i) - d(x_i)\}^2 \leqslant \sum_{i=1}^{n} \{Y_i - y(x_i)\}^2 = \sum_{i=1}^{n} e_i^2.$$

The result then follows from the definition of $q_n(\alpha)$. This holds for all $x_1, \ldots, x_n$. □

The choice of D represents the most difficult part of constructing a conservative set. The conservative set has sufficient coverage if D is sufficiently large to contain the discrepancy function $y - f_\theta$ for the best parameter. But a D that is too large creates an empirical score that is

uniformly 0; thus all possible parameters will be in the conservative set. Balancing the problems with a small and large set should be considered equivalent to the difficulty of constructing a prior on the discrepancy in the Bayesian approach (Kennedy and O'Hagan, 2001) or defining a reproducing kernel Hilbert space which contains the discrepancy in the large sample frequentist approaches (Tuo and Wu, 2015). The selection of η based on data is discussed in more detail in Section 7. Until then, imagine it fixed by a user's understanding of the potential discrepancy.

3.4. Conservative and consistent set

Though the conservative set has sufficient coverage, one naturally wonders whether the size of the conservative set is as small as possible. Certainly, if we make D sufficiently small to contain only $y - f_\theta$ for all $\theta \in \Theta^*$, then the size concern is resolved. But a user does not have access to y or Θ^* and typically D should be made fairly large to ensure sufficient coverage. The problem becomes that a large D, in general, results in the inclusion of suboptimal parameters in the conservative set. This is because the discrepancy for suboptimal parameters can be absorbed by D. So the conservative set is not the smallest set. This subsection demonstrates a set that is always contained in the conservative set and is sometimes much smaller than the conservative set.

The key novelty that is expressed in this paper is to use a new constraint on the discrepancy that removes a potential problem with the conservative set without harming the coverage of the set. Specifically, consider the space of discrepancies

$$I(\theta) = \{d \in D \text{ such that } l(f_\theta + d, f_\theta) \leqslant l(f_\theta + d, f_t) \text{ for all } t \in \Theta\}.$$

The motivation is that we should consider only discrepancies that do not corrupt the optimality of θ. This can be thought of as partitioning the response space into regions which correspond to a single θ. Also, if Θ is a Euclidean subspace and the model is differentiable with respect to θ, there is a descriptor of this space in terms of orthogonality that is discussed in Section 4.

By using the intersection of the reasonable discrepancies and this new constraint, the proposed conservative and consistent set is described by

$$\text{CCS} = \left\{ \theta \in \Theta \text{ such that } \min_{d \in I(\theta)} \hat{l}(\text{data}, f_\theta + d) \leqslant \frac{q_n(\alpha)}{n} \right\}.$$

This provides an intuitive resolution to potential problems with the conservative set. Formally, we still need to show that this set has sufficient coverage.

Theorem 2. Suppose that $y - f_\theta \in D$ for all $\theta \in \Theta^*$. Then CCS has sufficient coverage.

Proof. Using the same concept as theorem 1, we need only to show that $\theta \in \Theta^*$ implies that $y - f_\theta \in I(\theta)$. The optimality of Θ implies that $l(y, f_\theta) \leqslant l(y, f_t)$ for all $t \in \Theta$ and thus

$$l\{f_\theta + (y - f_\theta), f_\theta\} \leqslant l\{f_\theta + (y - f_\theta), f_t\} \qquad \text{for all } t \in \Theta.$$

3.5. Comparing the sets

The ordering of these sets by inclusion follows from their respective constructions.

Proposition 2. $\text{NS} \subset \text{CCS} \subset \text{CS}$.

The ideas behind these sets apply without using the structure of the parameter space or the input space. Fig. 1 illustrates the construction of the sets with a simple two-dimensional example. The true response is located at (0.3, 0.9) and the four parameters generate responses (0.1, 0.1), (0.5, 0.5) and (0.9, 0.9). The observations are located at each input and have values 0.2 and 0.8 and the confidence region has radius 0.25. The set D is a ball centred at zero with radius

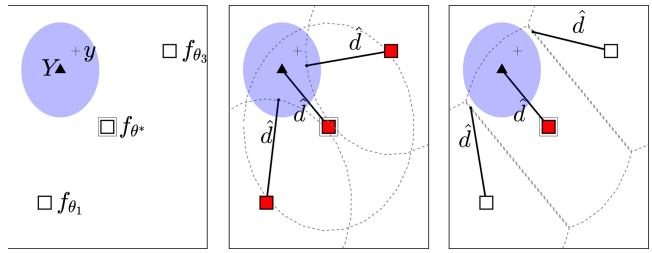

Fig. 1. Diagram illustrating the construction of the three sets described in Section 3 (+, nature's response; ▲, the observation; ●, confidence region for nature's response based on the observations; each of the three squares is a model's response; ⧈, ■, model with the optimal parameter; ■, in the confidence set; □, not in the confidence set): (a) naive set, where a model could be chosen only if it is in the confidence region around the observation; (b) conservative set, where a parameter is chosen if there is a point in the big, broken circle around each model that lies in the confidence region; (c) conservative and consistent set, where a parameter is chosen similarly to the conservative set but the broken circles are replaced by partitioned areas

0.25. After doing the respective optimizations, one can find that no parameters are chosen in the naive set, the conservative set includes all parameters, and the conservative and consistent set includes only the optimal parameter.

4. Solving the optimization problems

Deciding whether a value of a parameter θ is in CS or CCS hinges on the ability to calculate

$$\min_{d\in D} \hat{l}(\text{data}, f_\theta + d)$$

and

$$\min_{d\in I(\theta)} \hat{l}(\text{data}, f_\theta + d).$$

This section discusses the computation by casting these optimizations as finite convex programmes. This means that there is a finite number of decision variables and a finite number of constraints. Once placed as a finite, convex programme, many numerical methods can be leveraged to find the respective minima (Wright and Nocedal, 1999).

The optimization that is associated with CS can be recast as a finite dimensional programme via the kernel trick (Schölkopf *et al.*, 2001). This is sometimes known as the representer theorem. All proofs for the results in this section are in Appendix A.

Proposition 3. $\min_{d\in D} \hat{l}(\text{data}, f_\theta + d)$ is equal to

$$\min_{\delta_1,\ldots,\delta_n} \frac{1}{n}\sum_{i=1}^{n}\{f_\theta(x_i)+\delta_i - Y_i\}^2 \qquad \text{such that } \sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\delta_j k_{ij} \leqslant \eta^2$$

and k_{ij} is the ijth element of K_n^{-1} where

$$K_n = \begin{pmatrix} \kappa(x_1,x_1) & \cdots & \kappa(x_n,x_1) \\ \vdots & \ddots & \vdots \\ \kappa(x_1,x_n) & \cdots & \kappa(x_n,x_n) \end{pmatrix}.$$

Similarly, if the parameter space is finite, the same trick can be employed to cast the optimization that is associated with CCS as a finite convex problem.

Proposition 4. If $\Theta = \{1, \ldots, m\}$, $\min_{d \in I(\theta)} \hat{l}(\text{data}, f_\theta + d)$ is equal to

$$\min_{\delta_1,\ldots,\delta_n,\delta_{n+1},\ldots,\delta_{n+m}} \frac{1}{n} \sum_{i=1}^{n} \{f_\theta(x_i) + \delta_i - Y_i\}^2$$
$$\text{such that } \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i \delta_j k_{ij} + 2 \sum_{i=1}^{n} \sum_{j=n+1}^{n+m} \delta_i \delta_j k_{ij} + \sum_{i=n+1}^{n+m} \sum_{j=n+1}^{n+m} \delta_i \delta_j k_{ij} \leqslant \eta^2 \tag{5}$$
$$\delta_{n+j} \leqslant \frac{1}{2} \int \{f_j(x) - f_\theta(x)\}^2 \mathrm{d}\mu(x) \qquad \text{for } j = 1, \ldots, m$$

where k_{ij} is the ijth element of the $(n+m \times n+m)$-sized matrix

$$\begin{pmatrix} K_n & K_\theta \\ K_\theta^{\mathrm{T}} & K_{\theta\theta} \end{pmatrix}^{-1}$$

with K_θ an $n \times m$ matrix with ijth element $\int \kappa(x_i, x)\{f_j(x) - f_\theta(x)\}\mathrm{d}\mu(x)$ and $K_{\theta\theta}$ an $m \times m$ matrix with ijth element $\int \int \kappa(x, x')\{f_i(x) - f_\theta(x)\}\{f_j(x') - f_\theta(x')\}\, \mathrm{d}\mu(x)\, \mathrm{d}\mu(x')$.

When there are an infinite number of elements in Θ, evaluating whether θ is in CCS requires solving an optimization problem with an infinite number of constraints and an infinite number of decision variables. We shall reduce our goal to finding good programmes that offer reasonable approximations to the optimization problem. These approximations will produce sets that are still endowed with coverage and, under more assumptions, will be endowed with consistency in Section 5.

One approximation involves using a finite number of constraints, which is often termed a discretization approach in semifinite programming (Hettich and Kortanek, 1993).

Theorem 3. Choose m elements from Θ to use with problem (5). Define

$$\text{CCSD} = \left\{ \theta \in \Theta \text{ such that the solution to problem (5)} \leqslant \frac{q_n(\alpha)}{n} \right\}.$$

If $y - f_\theta \in D$ for all $\theta \in \Theta^*$, then for all $\theta \in \Theta^*$

$$\mathbb{P}(\theta \in \text{CCSD}) \geqslant 1 - \alpha.$$

Another approximation is a local approach that exploits differentiability. Say that the parameter space $\Theta \subset \mathbb{R}^p$. Assume that the computer model is differentiable with respect to the parameter almost everywhere with respect to μ. Call this gradient at θ evaluated at x $\nabla f_\theta(x)$. If Θ is open, then, for all $\theta \in \Theta^*$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} l(y, f_\theta) = \mathbf{0},$$

where the left-hand term is the gradient and the $\mathbf{0}$ represents a vector of 0s. This can be rewritten

$$\int \{y(x) - f_\theta(x)\} \nabla f_\theta(x) \mathrm{d}\mu(x) = 0.$$

Leveraging this constraint, the optimization problem can be written as

$$\min_{\delta_1,\ldots,\delta_n} \frac{1}{n} \sum_{i=1}^{n} \{f_\theta(x_i) - \delta_i - Y_i\}^2 \qquad \text{such that } \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i \delta_j \tilde{k}_{ij} \leqslant \eta^2 \tag{6}$$

where $\tilde{k}_{ij}$ is the ijth element of the $(n \times n)$-sized matrix $(K_n - q_\theta^{\mathrm{T}} Q_\theta^{-1} q_\theta)^{-1}$ with q_θ a $p \times n$ matrix with ith row $\int \kappa(x_i, x) \nabla f_\theta(x) \mathrm{d}\mu(x)$ and

$$Q_\theta = \int \int \kappa(x, x') \nabla f_\theta(x) \nabla f_\theta(x')^{\mathrm{T}} \mathrm{d}\mu(x) \mathrm{d}\mu(x').$$

Although this is a convex problem, the question of sufficient coverage again must be addressed.

Theorem 4. Suppose that Θ is an open subset of $\mathbb{R}^p$. Suppose also that $\nabla f_\theta(x)$ exists and is bounded across θ and $\mathbb{X}$. Define

$$\mathrm{CCSL} = \left\{ \theta \in \Theta \text{ such that the solution to problem (6)} \leqslant \frac{q_n(\alpha)}{n} \right\}.$$

If $y - f_\theta \in D$ for all $\theta \in \Theta^*$, then for all $\theta \in \Theta^*$

$$\mathbb{P}(\theta \in \mathrm{CCSL}) \geqslant 1 - \alpha.$$

The intersection $\mathrm{CCSD} \cap \mathrm{CCSL}$ of these two approximations still has confidence and is thus better than each individual set. This borrows the global guarantees of the former with the local guarantees of the latter. This method was used for all examples whenever CCS is used in a numerical setting. Focusing purely on the complexity of the resulting optimization problem, the programme that is associated with CCSL has m fewer decision variables and constraints compared with the programme that is associated with CCSD. Fewer constraints and decision variables are important to reducing the computational cost of a convex programme. Another important computational concern is the inversion of a matrix, sized $n+m$ for the finite approach and n for the local approach. However, the local approach has the potentially computationally taxing step of differentiating the computer model. Different problems will probably have different relative computational cost for optimization and gradient evaluation; thus there is probably no universal best computational option between these two approaches.

5. Consistency

This section demonstrates that the conservative and consistent set proposed does indeed have the desired consistency property. In this paper a set is consistent if, as the number of observations becomes large, the probability that a suboptimal parameter, $\theta \notin \Theta^*$, is in the set tends to 0. We shall show that this property holds for CCS as well as the discretized and local approximations CCSD and CCSL. This result does not extend to CS. Thus, of NS, CS and CCS, only CCS has both the coverage and the consistency property.

This result will be shown under two assumptions.

Assumption 3. The random variables $e_1, e_2, \ldots$ are identically distributed, independent random variables with mean 0 and variance σ^2.

Assumption 4. The data $x_1, x_2, \ldots$ are independent draws from a distribution ν. Also, there is some M such that $\mu(A)/\nu(A) < M$ for all $A \subset \mathbb{X}$.

The sampling plan assumption ensures that everywhere is eventually sampled. The sampling plan assumption can be replaced with some more complicated constraints on deterministic sampling plans. These conditions are considerably looser than those considered in Tuo and Wu (2015) and Wong *et al.* (2017), but consistency is a weaker claim than the efficiency claims in those works.

The consistency arguments are based on the following lower bound shown in Appendix B.

Lemma 1. Under assumptions 3 and 4, let θ^* be some element in Θ^*. Then

$$\min_{d \in I(\theta)} \hat{I}(\text{data}, f_\theta + d) \geqslant \sigma^2 + \frac{\{l(y, f_\theta) - l(y, f_{\theta^*})\}^2}{M \int \{f_\theta(x) - f_{\theta^*}(x)\}^2 \mathrm{d}\mu(x)}$$

occurs with probability tending to 1 as $n \to \infty$ for all $\theta \in \Theta$.

The penalty for suboptimal parameters in terms of the scoring metric is thus reflected in the evaluation criteria. Lemma 1, combined with

$$\frac{q_n(\alpha)}{n} \to \sigma^2,$$

yields the consistency result.

Theorem 5. Under assumptions 3 and 4, let CCS_n be the conservative and consistent set built with n data points. Suppose further that Θ^* is not empty. If $\theta \notin \Theta^*$, then

$$\mathbb{P}(\theta \in \mathrm{CCS}_n) \to 0.$$

This consistency argument directly carries over to the finite sample approximation that was introduced in Section 4 under an additional condition.

Corollary 1. Under the same assumptions as for theorem 5, let CCSD_n be the discrete parameter approximation to the conservative and consistent set built with n data points. Further suppose that some element from Θ^* is included in the m parameters that are used for the finite approximation. If $\theta \notin \Theta^*$, then

$$\mathbb{P}(\theta \in \mathrm{CCSD}_n) \to 0.$$

To extend these arguments to the local approximation to CCS, more assumptions are needed on the structure of the problem. Considering that the constraint that generates CCSL depends only on the first-order condition for optimality of θ, we require an assumption that translates first-order optimality to global optimality. This condition is convexity.

Theorem 6. Under the same assumptions as for theorem 5, suppose that Θ is an open subset of $\mathbb{R}^p$ and $\nabla f_\theta(x)$ exists and is bounded across Θ and $\mathbb{X}$. Suppose also that the partial derivatives $\nabla_i f_\theta$ are such that $\int \{\nabla_i f_\theta(x)\}^2 \mathrm{d}\mu(x) > 0$ for $i = 1, \ldots, p$. Let CCSL_n be the local approximation to the conservative and consistent set built with n data points. Further say that for all that $0 \leqslant w \leqslant 1$, if $\theta = w\theta_1 + (1-w)\theta_2$ is in Θ, then

$$l(y, f_\theta) \leqslant w l(y, f_{\theta_1}) + (1-w) l(y, f_{\theta_2}).$$

If $\theta \notin \Theta^*$, then

$$\mathbb{P}(\theta \in \mathrm{CCSL}_n) \to 0.$$

The proof for theorem 6 is also in Appendix B.

Comparing these results, if we can directly compute the conservative and consistent set, then this set has the most relaxed conditions of the three to produce consistency. If there are an infinite number of elements in Θ, then we either need to use a finite approximation to CCS with enough elements to capture an optimal parameter or to use a local approximation to CCS and to assume a convexity property on the scoring metric.

6. Connections and contrasts with other methods

We now compare the proposed conservative and consistent set with existing alternatives in a general setting. For this section, we shall presume that our e_is are independent and identically normally distributed with variance σ^2. To compare with other methods, we recast the proposed set in terms of a parameter evaluation function R. For the conservative and consistent set,

$$R(\theta) = \min_{d \in I(\theta)} \hat{l}(\text{data}, f_\theta + d),$$

and then the resulting confidence set is defined as

$$\text{CCS} = \left\{ \theta \in \Theta \text{ such that } R(\theta) \leqslant \frac{q_n(\alpha)}{n} \right\}.$$

The remainder of this section will define alternative approaches to build confidence sets in our context by altering $R(\cdot)$ to one of four alternatives. For all comparable methods, a parameter will be in the set if the parameter evaluation function at θ is less than $q_n(\alpha)/n$.

6.1. Non-linear least squares

It is uncommon for anyone actually to employ the naive set in practice as it is often empty for even slightly inaccurate models (see proposition 1). A more popular alternative uses the minimizing average squared error instead of σ^2, i.e.

$$\hat{\sigma}^2 = \min_{\theta \in \Theta} \hat{l}(\text{data}, f_\theta).$$

Clearly, some discrepancy becomes rolled into the calculation of $\hat{\sigma}^2$. This can be considered similar to the classical recommendations in Box and Coutie (1956) and Beale (1960), at least asymptotically. The set could then be built by deflating the parameter evaluation function with this new $\hat{\sigma}^2$, i.e.

$$R_1(\theta) = \frac{\sigma^2}{\hat{\sigma}^2} \hat{l}(\text{data}, f_\theta). \tag{7}$$

However, this set still does not maintain sufficient coverage under mild data generation assumptions.

Proposition 5. Suppose that $e_1, \ldots, e_n$ are independent and identically normally distributed with zero mean and variance σ^2. Let $x_1, \ldots, x_n$ be generated independently from a distribution ν. If $\theta^* \in \Theta^*$ but there is some $t \in \Theta$ such that

$$\int_{\mathbb{X}} \{y(x) - f_t(x)\}^2 \mathrm{d}\nu(x) < \int_{\mathbb{X}} \{y(x) - f_{\theta^*}(x)\}^2 \mathrm{d}\nu(x),$$

then $\mathbb{P}\{R_1(\theta^*) \leqslant q_n(\alpha)/n\} \to 0$ as $n \to \infty$.

Thus, although it is an obvious alternative with a storied history, this does not resolve the issues that are discussed in this paper because it will not provide sufficient coverage under our assumptions.

6.2. Stochastic models of the discrepancy

Studies such as Craig *et al.* (1997) and Kennedy and O'Hagan (2001) introduce the idea of treating the discrepancy like a random object. If the discrepancy is a random object, the differences $Y_i - f_\theta(x_i)$ come from two separate sources: the random e_i and the random model discrep-

ancy term. Craig *et al.* (1997) and Kennedy and O'Hagan (2001) noted that the discrepancy should not be modelled as independent and identically distributed errors. Consider using the conservative set with η set to $\sqrt{\{q_n(\alpha)/\sigma^2\}}$, i.e.

$$R_2(\theta) = \min_{\|d\|^2 \leqslant q_n(\alpha)/\sigma^2, d \in G} \hat{l}(\text{data}, f_\theta + d).$$

There are some probabilistic guarantees that come with this parameter evaluation function.

Proposition 6. Say that $y(\cdot) - f_\theta(\cdot)$ is a Gaussian process with mean 0 and the covariance function between $y(x) - f_\theta(x)$ and $y(x') - f_\theta(x')$ is $\kappa(x, x')$; then $R_2(\theta) \leqslant q_n(\alpha)/n$ with probability at least $1 - 2\alpha$.

It would appear that the statistical approach of modelling the discrepancy is closest to the conservative set. However, the history matching and posterior analyses of Craig *et al.* (1997) and Kennedy and O'Hagan (2001) mean that this analogy may not always directly apply. But, if we accept the argument that the conservative set is too large, we might extend this argument to these other approaches.

There are also some connections to non-linear generalized least squares; take for example Tarantola and Valette (1982). However, that stream of literature typically does not operate under the premise that the model is wrong; thus this paper will not directly compare with those ideas and the subsequent references. But the arguments against the conservative set can also extend to these types of approach.

6.3. Semiparametric efficiency arguments

Tuo and Wu (2015) introduced an estimation method termed L_2-calibration and demonstrated that the estimate is semiparametric efficient (see Bickel *et al.* (1998)). Semiparametric efficiency implies that the confidence sets that we construct, in large samples, can ignore all uncertainty in the discrepancy. We now give a confidence set that employs this concept, even if it is not exactly the same as in Tuo and Wu (2015), to draw connections. Say that $\hat{y}(\cdot)$ is a functional estimate of the true function and $\hat{\theta}$ is our estimate of a single point in Θ^*. We can then imagine that $\hat{d}(\cdot) = \hat{y}(\cdot) - f_{\hat{\theta}}(\cdot)$ is a good estimate of $y(\cdot) - f_{\theta^*}(\cdot)$, where θ^* is the single element in Θ^*. Then a semiparametric efficiency argument would imply that we can use a parameter evaluation function of

$$R_3(\theta) = \hat{l}(\text{data}, f_\theta + \hat{d}).$$

It becomes clear that this set is a more aggressive confidence set compared with the proposed CCS. The confidence set that is implied by L_2 ignores the uncertainty that is caused by learning the discrepancy. Since learning a function with few samples can be difficult, this set is often much more aggressive compared with CCS and as a result does not give sufficient coverage in most calibration problems.

6.4. Weighted least squares

Say that $x_1, \ldots, x_n$ are independently sampled according to some known measure $\nu(\cdot)$. Similarly in spirit to finding unbiased estimators with importance sampling (see, for example, Owen and Zhou (2000) and references within), we can use the parameter evaluation function of

$$R_4(\theta) = \frac{\sigma^2}{\min_\theta (1/n) \sum_{i=1}^n \partial\mu(x_i)/\partial\nu \{Y_i - f_\theta(x_i)\}^2} \frac{1}{n} \sum_{i=1}^n \frac{\partial\mu}{\partial\nu}(x_i) \{Y_i - f_\theta(x_i)\}^2,$$

where $\partial\mu/\partial\nu$ is the Radon–Nikodym derivative of μ with respect to ν. The leading term is merely the adjustment for the adjusted variance, like in equation (7). Wong *et al*. (2017) used a criterion that was similar in spirit to this with a semiparametric bootstrap. Their theoretical arguments operate in the regime of semiparametric efficiency; thus similar problems to L_2-calibration can appear in small samples. However, there is a bigger distinction between that method and our approach. The method of Wong *et al*. (2017) requires all x_is to be independent and randomly generated via some known distribution. In contrast, we are attempting to provide inference even if we are running a designed experiment where x_is are not random or maybe the generation of x_is is outside the control of the user.

7. Specifying η

Careful choices of $\kappa(\cdot,\cdot)$ and η are critical to successful implementation of the confidence set proposed.

In contrast with a Bayesian paradigm, the results that are implied in this paper do not require the use of some best κ; any κ should work so long as $\|y - f_\theta\|$ is finite for all $\theta \in \Theta^*$. Mild tuning of the function can occur without harming the conceptual underpinning of the method. There is some risk in drastically adjusting the κ-function on the basis of observed data because it determines the relationship between the discrepancy at the observed and unobserved locations.

The selection of η, being a single dimension and generally unknown at the start of calibration, represents a reasonable choice to tune or estimate. Before looking at specific estimates, we first examine the properties that an estimate should have. Say that we use a value $\hat{\eta}_n$ that depends on the first n observations. Any value of $\hat{\eta}_n$ that does not wander off to ∞ will maintain the consistency property. If $\hat{\eta}_n \geqslant \|y - f_\theta\|$ for all $\theta \in \Theta^*$, then confidence properties are maintained and there should not be much concern. If it is underestimated, the next result explains how far underestimated it can be and maintain confidence. This result shows that it is sufficient to find some estimate of $\hat{\eta}_n$ that behaves like a good statistical estimate. The proof is in Appendix C.

Theorem 7. Assume that the random variables $e_1, e_2, \ldots$ are identically distributed, independent random variables with mean 0 and $\mathbb{E}(e_i^4)$ is finite. Let CCS_n be the conservative and consistent set built with n data points using $\hat{\eta}_n$. For all $\theta \in \Theta^*$, if, for all constants $M > 0$,

$$\mathbb{P}(\|y - f_\theta\| - \hat{\eta}_n \leqslant Mn^{-1/2}) \to 0,$$

then, for all $\epsilon > 0$, there is some sufficiently large n such that $\mathbb{P}(\theta \in \text{CCS}_n) \geqslant 1 - \alpha - \epsilon$.

We now construct an estimate for η based on data in a general setting. Consider that there is a function $\delta(\cdot) = y(\cdot) - f_\theta(\cdot)$ in our reproducing kernel Hilbert space G with a norm $\|\delta\|$. We would like to use observations $\Delta_1 = \delta(x_1) + e_1, \ldots, \Delta_n = \delta(x_n) + e_n$ to estimate the norm of the underlying function when $e_1, \ldots, e_n$ are independent random variables with mean 0 and variance σ^2. Estimating η is closely related to the idea of smoothing data, which has a long and celebrated history (Wahba, 1990). Surprisingly, we have found no existing statistical tools to estimate the norm in this general setting. The oracle estimate

$$\eta_n = \min_{d \in G} \|d\| \qquad \text{such that } d(x_i) = \delta(x_i) \text{ for } i = 1, \ldots, n$$

gives the smallest norm possible given knowledge of $\delta(x_1), \ldots, \delta(x_n)$. The obvious estimate of

$$\min_{d \in G} \|d\| \qquad \text{such that } d(x_i) = \Delta_i \text{ for } i = 1, \ldots, n$$

has an extremely large variance. In general, we suggest decreasing the variance of this estimator by employing estimators of $\delta(x_i)$, $\hat{\Delta}_i$, yielding

$$\hat{\eta}_n = \min_{d \in G} \|d\| \qquad \text{such that } d(x_i) = \hat{\Delta}_i \text{ for } i = 1, \ldots, n. \tag{8}$$

Specifically, we consider using an $n \times n$ approximation matrix $A_n(\lambda)$ that depends on a smoothing parameter λ,

$$\begin{pmatrix} \hat{\Delta}_1 \\ \vdots \\ \hat{\Delta}_n \end{pmatrix} = A_n(\lambda) \begin{pmatrix} \Delta_1 \\ \vdots \\ \Delta_n \end{pmatrix}.$$

$A_n(\lambda)$ is typically designed to reduce variance at the cost of bias. The mean-square error of our estimator is bounded by using the reverse triangle inequality and carrying through the expectation.

Proposition 7. Let I_n be the $n \times n$ identity matrix and define

$$S_n(\lambda) = \begin{pmatrix} \delta(x_1) \\ \vdots \\ \delta(x_n) \end{pmatrix}^{\mathrm{T}} \{I_n - A_n(\lambda)\} K_n^{-1} \{I_n - A_n(\lambda)\} \begin{pmatrix} \delta(x_1) \\ \vdots \\ \delta(x_n) \end{pmatrix} + \sigma^2 \mathrm{tr}\{A_n(\lambda) K_n^{-1} A_n(\lambda)\}.$$

Then $\mathbb{E}\{(\hat{\eta}_n - \eta_n)^2\} \leqslant S_n(\lambda)$.

Although we would like to set λ to the minimizer of $S_n(\lambda)$, this value is unknown as it depends on $\delta(x_1), \ldots, \delta(x_n)$. Our suggestion is to use the following estimate of $S(\lambda)$:

$$\hat{S}_n(\lambda) = \begin{pmatrix} \Delta_1 - \hat{\Delta}_1 \\ \vdots \\ \Delta_n - \hat{\Delta}_n \end{pmatrix}^{\mathrm{T}} K_n^{-1} \begin{pmatrix} \Delta_1 - \hat{\Delta}_1 \\ \vdots \\ \Delta_n - \hat{\Delta}_n \end{pmatrix} + 2\sigma^2 \mathrm{tr}\{A_n(\lambda) K_n^{-1}\} - \sigma^2 \mathrm{tr}(K_n^{-1}).$$

This is an unbiased estimate, $\mathbb{E}\{\hat{S}_n(\lambda)\} = S_n(\lambda)$. The process of finding this type of unbiased estimate of an upper bound on the mean-squared error has connections with general shrinkage research, but this digression is outside the scope of this paper (for example, see Xie *et al.* (2012)).

In summary, our recommendation is to select a value λ_n by minimizing $\hat{S}_n(\lambda)$, and then to use this to find the estimator listed in result (8). For each θ being investigated, this process can be repeated. All that is left is to specify the exact form of $A_n(\lambda)$. We found that $(I_n + \lambda K_n^{-2})^{-1}$ performed better compared with the traditional smoother $(I_n + \lambda K_n^{-1})^{-1}$. If we want to offer some insurance to the user, doubling this estimate is a reasonable way to ensure that we comply with theorem 7. This action was taken for all examples.

We now address the theoretical argument that is needed to have guarantees of large sample coverage.

Proposition 8. If, as n becomes large, $\eta_n \to \eta$ in probability and there is some sequence $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n$ such that $S_n(\tilde{\lambda}_n) \to 0$, then $\hat{\eta}_n$ from result (8) using $\tilde{\lambda}_n$ converges to η in probability.

Under the generality that is considered in this paper, we could not provide a formal proof for the conditions of this result. In practice, we found that the effect of altering estimation methods for η was relatively minor, but the above approach worked well.

8. Simulation comparison

We shall now use a simulation study to illustrate some advantages of the proposed method over some of the existing and modern methods. This is not intended to be an exhaustive study, but instead to highlight the potential drawbacks of other approaches when the true model discrepancy is known. The test bed will be two model–system pairs from the literature with the input domain [0,1]. Courtesy of Tuo and Wu (2015), the first model–system pair is

$$f_\theta(x) = y(x) - \sqrt{(\theta^2 - \theta + 1)}\{\sin(2\pi x\theta) + \cos(2\pi x\theta)\} \qquad y(x) = \exp\left(\frac{2\pi x}{10}\right)\sin(2\pi x). \quad (9)$$

The second, from Plumlee (2017), is

$$f_\theta(x) = \theta x, \qquad y(x) = 4x + x\sin(5x). \quad (10)$$

Seven approaches for 90% confidence sets are chosen for comparison: $\widehat{\text{NS}}$, the set that was described in Section 6.1; KO, the 90% credible set from Kennedy and O'Hagan (2001) with correlation structure $\kappa(\cdot,\cdot)$; the 90% confidence set for L_2-calibration from Tuo and Wu (2015); WSL, the 90% bootstrapped confidence set of Wong *et al.* (2017); OGP, the orthgonal Gaussian process modification of Kennedy and O'Hagan (2001) that was described in Plumlee (2017); oracle CCS, the optimal η gifted to us; CCS, η chosen as described in Section 7.

Two observation schemes are considered. The first can be described as data poor: six observations occur at $\{0, 0.1, 0.2, 0.4, 0.6, 0.7\}$. The second can be described as data rich: 12 observations occur at equally spaced points between 0 and 1. We shall use the kernel function

$$\kappa(x, x') = (1 + |x - x'|)\exp(-|x - x'|).$$

Each Y_i is an independent, normally distributed random variable with mean $y(x_i)$ and variance 0.2^2. We measure two properties of the resulting predictive sets on θ. The first is the coverage probability, with the goal of its being at least 90%. The second is the width of the interval, with the goal of its being as small as possible while still meeting the coverage probability objective.

The results after 1000 replicates are described in Table 1. It becomes clear that the numerous other methods, which are not designed for sufficient coverage in small sample sizes, produce inadequate coverage. The only approach that nearly gives coverage for all cases is CCS, being at worst 5% below ideal. Given that sufficient coverage is not the goal of any of the previous approaches, this should not be surprising. Although we expected the KO set to mimic properties

Table 1. Frequency of coverage and average width of interval for the study in Section 8

Method	*Frequency of coverage*				*Width of interval*			
	y, f_θ *in equation (9)*		*y*, f_θ *in equation (10)*		*y*, f_θ *in equation (9)*		*y*, f_θ *in equation (10)*	
	Poor	*Rich*	*Poor*	*Rich*	*Poor*	*Rich*	*Poor*	*Rich*
$\widehat{\text{NS}}$	1.000	1.000	0.561	1.000	1.090	0.169	1.038	0.997
KO	0.109	0.000	0.452	0.387	0.401	0.140	0.769	0.594
L_2	0.063	0.779	0.728	0.878	0.092	0.021	0.632	0.282
WSL	0.153	0.673	0.043	0.589	0.031	0.023	0.303	0.198
OGP	0.998	0.967	0.387	0.951	1.067	0.741	0.723	0.310
Oracle CCS	0.998	0.992	0.893	0.989	0.136	0.058	2.073	0.672
CCS	0.997	0.992	0.849	0.997	0.283	0.080	2.285	0.740

of the conservative set, i.e. to be too large but to have coverage, it turns out to have surprisingly poor coverage properties. This might be due to the complicated nature of posteriors with unknown (and integrated-out) variance terms. The other interesting comparison lies between the OGP set and the proposed CCS. Although coverage appears good in the model–system pair from expression (9), the width of the OGP set is significantly wider. This is because orthogonality is a necessary but not sufficient condition for parameter optimality in the case of non-convex $l(y, f_\theta)$. Because expression (9) implies that $l(y, f_\theta)$ is non-convex, the OGP approach yields an interval that is too large compared with CCS. $\widehat{\text{NS}}$ is also large relative to CCS for this example, especially in the data poor case. In general, there does not appear to be much loss between CCS and the oracle version of CCS.

9. Examples

This section will discuss two real data examples where the true parameter is unknown, but the sets that are introduced by this paper can still be constructed and contrasted. These examples are intended to demonstrate the existence of practical differences.

9.1. Box and Coutie's example

Owing to the long history of the calibration problem, consider the work of Box and Coutie (1956). They introduced one of the first methods for, and implementations of, statistical calibration using a numerical algorithm conducted by a computer. Their calibration problem consists of learning the behaviour of a consecutive reaction which follows the first-order differential equation

$$\frac{\mathrm{d}M_1}{\mathrm{d}x} = -0.001^{\theta_1} M_1$$

and

$$\frac{\mathrm{d}M_2}{\mathrm{d}x} = 0.001^{\theta_2} M_1 - 0.001^{\theta_2} M_2$$

with initial condition $M_1(0) = 100$ and $M_2(0) = 0$.

The goal is calibration of the outcome M_2 based on observations at time points 10, 20, 40, 80, 160 and 320, with two replications. We consider a region of possible time points [0, 400]. The observations are assumed to be normally distributed. The replications in the experiment justify σ^2 to be near 13, which agrees with Box and Coutie's conclusion. We let

$$\kappa(x, x') = (1 + |x - x'|/400) \exp(-|x - x'|/400)$$

and consider θ-values in the box $[0.8, 1.3] \times [0.6, 1.1]$. The value of η was chosen via the automated mechanism that was described in Section 7.

Fig. 2 shows three sets that are discussed in this paper alongside that proposed by Box and Coutie (1956), which can be considered a large sample approximation to $\widehat{\text{NS}}$. The algorithm was reimplemented on a modern computer, giving slightly different results from those reported in Box and Coutie (1956). The Kennedy and O'Hagan (2001) set seems to imply that the best model has a large discrepancy near 120, which disagrees with all other sets. The L_2-set seems overly optimistic on the location of the best parameter considering that only a single time point is investigated after 200. The key comparison is between the traditional set that was proposed by Box and Coutie and the set that is proposed in this paper. The CCS is more elliptical in a

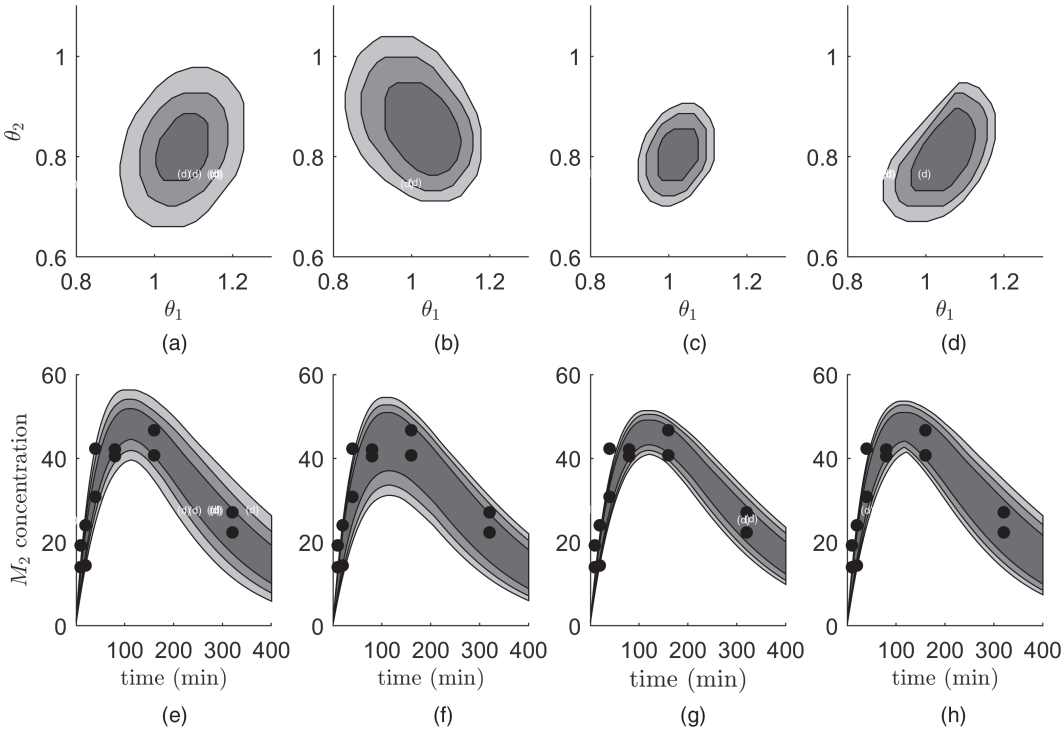

Fig. 2. Example from Section 9.1: (a)–(d) two-dimensional plots of the sets for α chosen as 0.001, 0.01 or 0.1, with the darkest area representing the largest α (some details of the methods to build confidence intervals are dicussed in Section 8); (e)–(h) data collected from the real system (●) alongside pointwise intervals for $f_\theta(x)$ when θ is in the respective confidence set; (a), (e) Box and Coutie; (b), (f) Kennedy and O'Hagan; (c), (g) L_2; (d), (h) CCS

specific way. The shape difference can be explained by propagating the parameter through a computer model. In our observational data, there is an absence of points in the right-hand half of the input space, implying that we are less sure of the optimal behaviour in that region. This is seen in CCS, where the uncertainty is large in the right-hand half of the input space. Since Box and Coutie's sets do not account for potential discrepancy, their set is unnaturally small in the right-hand part of the input space.

Although computational speed heavily depends on implementation, CCS and the Kennedy and O'Hagan set took roughly the same amount of time to construct, whereas the L_2-set and the Box and Coutie set took roughly an order of magnitude less time.

9.2. Crash test example

A researcher is studying the relationship between vehicle design and an injury criterion based on head movement. Four design variables are under consideration: the first three binary and the fourth has five levels, giving 40 total combinations of variables. Expensive experiments were done by crashing nine test vehicles and a single set of conditions was repeated, so eight unique combinations were attempted. The researcher would then like to study all 40 possible combinations by using data from these eight attempted combinations. A finite element approximation to the crash event was built with unknown parameters.

The original users had a concern that, after finding the best parameter for their computer model by minimizing least squares ($\hat{l}(\text{data}, f_\theta)$), the residual plots with respect to the minimizer

of the square error exhibited problems; Fig. 3. The problems are especially clear with respect to the first design variable.

In total, 49 parameter combinations were identified as reasonable. So the parameter space is treated as this finite discrete space. The users were fairly confident that the discrepancy is slowly varying, so they were comfortable employing the kernel

$$\kappa(x, x') = \exp(-\|x - x'\|_2^2),$$

where all variables were scaled to the unit cube. Our task is then to choose from these 49 parameters those that best represent the response across all inputs. Because there is no derivative information, we cannot use the L_2- or OGP method. But we shall use the WSL method to compare with the CCS.

The numbers of parameters that were selected by each method are given in Table 2. Given the data to this point, the KO confidence set is difficult to trust. The WSL method appears to encourage extreme confidence in a single parameter, which should be viewed with some scepticism. We find that CCS is a more conservative option at every α-level.

Another attempt at creating a confidence set would be to select the linear kernel

$$\kappa(x, x') = 1 + x \cdot x',$$

which is a positive semidefinite kernel function. This kernel accounts for linear discrepancies so it might address the residual plots in Fig. 3. Then we could simply choose η to be quite large,

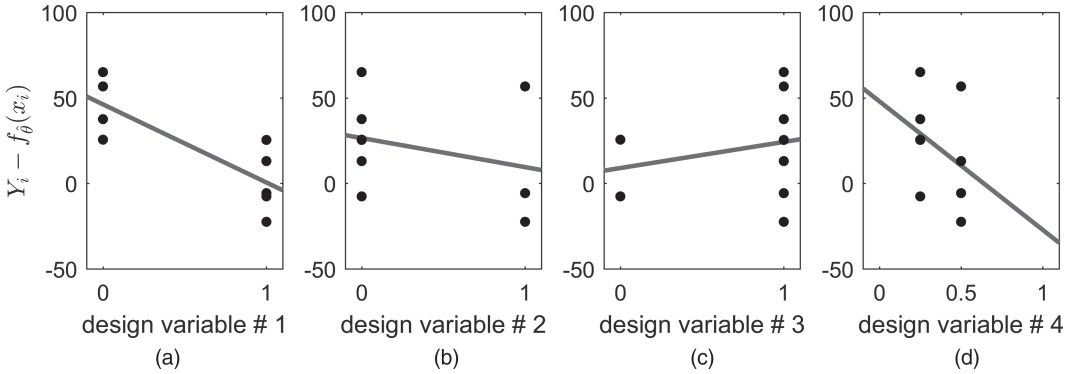

Fig. 3. Residual plots for the case-study in Section 9.2

Table 2. Number of parameters selected (out of 49) with different αs by using different methods for the example discussed in Section 9.2

Method	*Results for the following values of* α*:*			
	0.0001	*0.001*	*0.01*	*0.1*
$\widehat{\text{NS}}$	5	4	4	2
KO	20	13	5	2
WSL	1	1	1	1
CCS	20	18	17	10

say 10^4. At the $\alpha = 0.01$-level, this set selected nine parameters: nine fewer than the 18 in our baseline CCS. But this accounts only for linear discrepancy and thus might be a less robust set compared with the CCS.

Overall, the CCS at $\alpha = 0.01$ was used to form a model ensemble that includes the parametric uncertainty. This ensemble was used during the decision-making process on vehicle designs. In terms of computational speed, CCS was computed for $\alpha = 0.01$ in about a quarter of a second on the author's standard desktop computer.

10. Concluding remarks

This work has provided a framework for the production of sets that contain the best calibration parameter with at least some chosen probability. The results hold for both small and large samples. Moreover, these sets are shown to be consistent.

Some other major ideas from Bayesian calibration can be borrowed and placed in this framework. For functional responses (Bayarri *et al.*, 2007a; Higdon *et al.*, 2008), one could include the functional variable as an input and directly use the methods that are discussed in this work. Alternatively, it might prove fruitful to establish a set of basis functions to represent the functional response.

A full replacement for the existing Bayesian and large sample frequentist calibration methods would include the emulation of the computer model from a computer experiment and accounting for potential emulation error (Santner *et al.*, 2003). This can be accomplished, in theory, by placing the computer model in an appropriate function space and revising the optimization to include both the discrepancy and the computer model as decision variables. A computational mechanism to compute the set under this adjustment then is an open problem. The tricks from Section 4 do not have obvious analogues. Perhaps a more easy-to-compute approach would involve resampling (Wong *et al.*, 2017). Moreover, establishing consistency in the presence of emulation error appears beyond the tools that are used to establish consistency without emulation error. Thus there are theoretical questions in the presence of emulation error in addition to computational questions.

The idea of this paper is that a user would like to improve the computer model through parameter adjustment alone. Discrepancy correction was not explicitly explored in this paper. However, the constructions in this paper can be leveraged to build a set with $1-\alpha$ sufficient confidence for the response at x_0, $y(x_0)$. One example is if we define the region

$$U(\alpha) = \left\{ (\theta, d) \in \Theta \times G \text{ such that } d \in I(\theta) \text{ and } \hat{l}(\text{data}, f_\theta + d) \leqslant \frac{q_n(\alpha)}{n} \right\};$$

then a confidence interval is

$$\left[\min_{(\theta,d)\in U(\alpha)} f_\theta(x) + d(x), \max_{(\theta,d)\in U(\alpha)} f_\theta(x) + d(x) \right].$$

These adjusted confidence intervals will lack the salient features of the computer model, but they might be of interest to a user who would like intervals with sufficient confidence properties. We leave this as a potential topic for future research.

Acknowledgements

The author gratefully acknowledges National Science Foundation award 1833195 and the Issac Newton Institute for support during the production of this work. The author also thanks Dan

Apley, Barry Nelson, Anthony O'Hagan, J. P. Gosling, Judy Jin, Wenbo Sun, two reviewers, the Associate Editor and the Joint Editor for their helpful comments and ideas towards this research.

Appendix A: Proofs of results in Section 4

A.1. Proof of proposition 3

Each element d in the reproducing kernel Hilbert space can be decomposed into

$$d(\cdot) = \sum_{i=1}^{n} \beta_i \kappa(\cdot, x_i) + v(\cdot),$$

for some $v(\cdot)$ also in the reproducing kernel Hilbert space but is orthogonal to the first term. This, combined with the technical assumption on the kernel, implies that $v(x_i) = 0$ for all $i = 1, \ldots, n$. We have

$$\left\| \sum_{i=1}^{n} \beta_i \kappa(\cdot, x_i) + v \right\| = \left\| \sum_{i=1}^{n} \beta_i \kappa(\cdot, x_i) \right\| + \|v\|.$$

Thus setting $v = 0$ maximizes the feasible region and does not impact the objective function. Also, the norm of $\Sigma_{i=1}^{n} \beta_i \kappa(\cdot, x_i)$ is given by the constraint stated in the theorem by the definition of '$\|\cdot\|$' in equation (2).

A.2. Proof of proposition 4

Each element d in the reproducing kernel Hilbert space can be decomposed into

$$d(\cdot) = \sum_{i=1}^{n} \beta_i \kappa(\cdot, x_i) + \sum_{i=n+1}^{n+m} \beta_i \int \{f_{i-n}(x) - f_\theta(x)\} \kappa(\cdot, x) \mathrm{d}\mu(x) + v(\cdot),$$

where v is orthogonal to the previous two parts. This implies that $v(x_i) = 0$ for all $i = 1, \ldots, n$. This also implies that

$$\left\langle \int \{f_{i-n}(x) - f_\theta(x)\} \kappa(\cdot, x) \mathrm{d}\mu(x), v(\cdot) \right\rangle = 0$$

so that, by equation (3) with Fubini's theorem granted by the boundedness of κ and f_θ,

$$\int \{f_{i-n}(x) - f_\theta(x)\} v(x) \mathrm{d}\mu(x) = 0.$$

Thus setting $v = 0$ maximizes the feasible region and does not impact the objective function or the constraints that are generated by the parameters. Like the proof of proposition 3, the quadratic constraint is merely a restatement of norm.

A.3. Proof of theorem 3

If

$$l(f_\theta + d, f_\theta) \leqslant l(f_\theta + d, f_t) \qquad \text{for all } t \in \Theta,$$

then

$$l(f_\theta + d, f_\theta) \leqslant l(f_\theta + d, f_j) \qquad \text{for } j = 1, \ldots, m.$$

Thus CCS $\subset$ CCSD and theorem 2 completes the result.

A.4. Proof of theorem 4

Suppose that

$$l(f_\theta + d, f_\theta) \leqslant l(f_\theta + d, f_t) \qquad \text{for all } t \in \Theta.$$

Let $t = \theta + wb_i$, where w is a scalar and b_i is a length p vector with a 1 in the ith position and 0s elsewhere. Since Θ is open, for some sufficiently small $\epsilon > 0$, $|w| \leqslant \epsilon$ implies that $t \in \Theta$. Letting w go to 0 from above implies that

$$\int_X d(x) b_i^{\mathrm{T}} \nabla f_\theta(x) \mathrm{d}\mu(x) \leqslant 0,$$

and also letting w go to 0 from below implies that

$$\int_X d(x) b_i^{\mathrm{T}} \nabla f_\theta(x) \mathrm{d}\mu(x) \geqslant 0.$$

Noting this for all $b_1, \ldots, b_p$ gives if d is such that

$$l(f_\theta + d, f_\theta) \leqslant l(f_\theta + d, f_t) \qquad \text{for all } t \in \Theta$$

then

$$\int_X d(x) \nabla f_\theta(x) \mathrm{d}\mu(x) = 0.$$

Following the steps to rewriting the optimization problem for proposition 4, the following optimization problem has a result that is smaller than $\min_{d \in I(\theta)} \hat{l}(\text{data}, f_\theta)$,

$$\min_{\delta_1, \ldots, \delta_n, \delta_{n+1}, \ldots, \delta_{n+p}} \frac{1}{n} \sum_{i=1}^{n} \{ f_\theta(x_i) + \delta_i - Y_i \}^2$$

$$\text{such that } \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i \delta_j k_{ij} + 2 \sum_{i=1}^{n} \sum_{j=n+1}^{n+p} \delta_i \delta_j k_{ij} + \sum_{i=n+1}^{n+p} \sum_{j=n+1}^{n+p} \delta_i \delta_j k_{ij} \leqslant \eta$$

$$\delta_{n+j} = 0 \text{ for } j = 1, \ldots, p$$

where k_{ij} is the ijth element of the $(n + p \times n + p)$-sized matrix

$$\begin{pmatrix} K & q_\theta \\ q_\theta^{\mathrm{T}} & Q_\theta \end{pmatrix}^{-1}.$$

Using block matrix decomposition, the above programme is equivalent to that in the statement of theorem 4. Conclude that CCS $\subset$ CCSL and thus theorem 2 demonstrates the result.

Appendix B: Proofs for Section 5

To keep new notation separate from notation in the rest of the paper, several random variables will be introduced labelled $\mathcal{A}_n, \mathcal{B}_n, \ldots, \mathcal{M}_n$. The notation $\to^{\mathrm{p}}$ represents convergence in probability. Without loss of generality, we take $\eta = 1$.

B.1. Proof of lemma 1 and theorem 5

Let

$$\mathcal{A}_n := \frac{2}{n} \sum_{i=1}^{n} \{ y(x_i) - f_\theta(x_i) \} e_i - \max_{d \in D} \frac{2}{n} \sum_{i=1}^{n} d(x_i) e_i + \min_{d \in I(\theta)} \frac{1}{n} \sum_{i=1}^{n} \{ d(x_i) + f_\theta(x_i) - y(x_i) \}^2.$$

The condition $\mathcal{A}_n \leqslant (1/n)\{ q_n(\alpha) - \Sigma_{i=1}^{n} e_i^2 \}$ implies that $\theta \in \mathrm{CCS}_n$. Let

$$\mathcal{B}_n := \frac{1}{n} \sum_{i=1}^{n} \{ y(x_i) - f_\theta(x_i) \} e_i,$$

$$\mathcal{C}_n := \max_{d \in D} \frac{1}{n} \sum_{i=1}^{n} d(x_i) e_i$$

and

$$\mathcal{D}_n := \min_{d \in I(\theta)} \frac{1}{n} \sum_{i=1}^{n} \{ d(x_i) + f_\theta(x_i) - y(x_i) \}^2.$$

Lemma 2. $\mathcal{B}_n \to^{\mathrm{p}} 0$.

Lemma 3. $\mathcal{C}_n \to^{\mathrm{p}} 0$.

Lemma 4. For every $\theta^* \in \Theta^*$,

$$\mathcal{D} = \frac{\{l(y, f_\theta) - l(y, f_{\theta^*})\}^2}{\int \{f_\theta(x) - f_{\theta^*}(x)\}^2 \mathrm{d}\mu(x)/\mathrm{d}\nu\mathrm{d}\mu(x)}.$$

Then, for all $\epsilon > 0$, $\mathbb{P}(\mathcal{D}_n \leqslant \mathcal{D} - \epsilon) \to 0$.

These three results give lemma 1. Because

$$\frac{1}{n} q_n(\alpha) - \frac{1}{n} \sum_{i=1}^{n} e_i^2 \xrightarrow{\mathrm{p}} 0,$$

this implies theorem 5.

B.2. Proof of lemma 2

The proof of lemma 2 follows from assumption 3, where we know that, conditionally on $x_1, \ldots, x_n$, $\mathcal{B}_n$ is the average of variables with mean 0 and variance $\sigma^2 \int \{y(x) - f_\theta(x)\}^2 \mathrm{d}\nu(x)$. Thus the law of large numbers gives the result.

B.3. Proof of lemma 3

Maximizing a linear function over a quadric constraint (Wright and Nocedal (1999), chapter 16) gives

$$\mathcal{C}_n = \frac{1}{n} \sqrt{\left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa(x_i, x_j) e_i e_j \right\}}.$$

Then we can rewrite the kernel in terms of its eigenvalue expansion, with eigenvalues $\lambda_1, \lambda_2, \ldots$ and eigenfunctions $\psi_1(\cdot), \psi_2(\cdot), \ldots$:

$$\mathcal{C}_n = \frac{1}{n} \sqrt{\left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} e_i e_j \sum_{k=1}^{\infty} \psi_k(x_i) \psi_k(x_j) \lambda_k \right\}}.$$

or

$$\mathcal{C}_n = \sqrt{\left[\sum_{k=1}^{\infty} \lambda_k \left\{ \frac{1}{n} \sum_{i=1}^{n} \psi_k(x_i) e_i \right\}^2 \right]}.$$

This goes to 0 in probability if $\Sigma_{i=1}^{\infty} \lambda_k^2 < \infty$ (Grimmett and Stirzaker (2001), chapter 7). This is implied by technical assumption 1. See, for example, the beginning of the proof of theorem 3.1 in Koltchinskii and Ginée (2000).

B.4. Proof of lemma 4

The result is implied by finding a pair of random variables $\mathcal{E}_n$ and $\mathcal{F}_n$ such that

$$\mathcal{D}_n \geqslant \mathcal{E}_n / \mathcal{F}_n,$$

where $\mathcal{F}_n \to^{\mathrm{p}} \mathcal{F}$ in probability with $\mathcal{F} > 0$ and $\mathcal{E}_n \to^{\mathrm{p}} \mathcal{E}$ in probability for some $\mathcal{E} > 0$.

Out of the infinite number of constraints that define $I(\theta)$, consider only the one generated by θ^*:

$$0 \leqslant \int \{f_\theta(x) - f_{\theta^*}(x)\}^2 \mathrm{d}\mu(x) + 2 \int d(x) \{f_\theta(x) - f_{\theta^*}(x)\} \mathrm{d}\mu(x).$$

We also know that

$$l(y, f_\theta) - l(y, f_{\theta^*}) = \int \{f_{\theta^*}(x) - y(x)\}^2 \mathrm{d}\mu(x) - \int \{f_\theta(x) - y(x)\}^2 \mathrm{d}\mu(x)$$
$$= -\int \{f_\theta(x) - f_{\theta^*}(x)\}^2 \mathrm{d}\mu(x) + 2\int \{f_\theta(x) - y(x)\}\{f_\theta(x) - f_{\theta^*}(x)\}\mathrm{d}\mu(x).$$

Let $r = d + f_\theta - y$. A relaxation of $d \in I(\theta)$ is then that all r that meet the criterion

$$l(y, f_\theta) - l(y, f_{\theta^*}) \leqslant 2\int r(x)\{f_\theta(x) - f_{\theta^*}(x)\}\mathrm{d}\mu(x).$$

We need to relate the integral statements to approximations using data. To do this, consider

$$\mathcal{G}_n = \frac{1}{n}\sum_{i=1}^{n} r(x_i)\{f_\theta(x_i) - f_{\theta^*}(x_i)\}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x_i)$$

and

$$\mathcal{G} = \int r(x)\{f_\theta(x) - f_{\theta^*}(x)\}\mathrm{d}\mu(x),$$

where $\mathrm{d}\mu/\mathrm{d}\nu$ is the Radon–Nikodym derivative of μ with respect to ν which exists and is bounded per assumption 4.

Let

$$\mathcal{H}_n := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x_i)\{f_\theta(x_i) - f_{\theta^*}(x_i)\}\kappa(x_i, x_j)\{f_\theta(x_j) - f_{\theta^*}(x_j)\}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x_j)$$

and

$$\mathcal{I}_n := \frac{1}{n}\sum_{i=1}^{n}\int \{f_\theta(x) - f_{\theta^*}(x)\}\kappa(x, x_i)\{f_\theta(x_i) - f_{\theta^*}(x_i)\}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x_i)\mathrm{d}\mu(x).$$

Additionally define

$$\mathcal{J} := \int\int \{f_\theta(x) - f_{\theta^*}(x)\}\kappa(x, x')\{f_\theta(x') - f_{\theta^*}(x')\}\mathrm{d}\mu(x)\mathrm{d}\mu(x').$$

Clearly, $\mathcal{I}_n \to^{\mathrm{p}} \mathcal{J}$ by using condition 4. Also, using the eigenvalue decomposition from the proof of lemma 3,

$$\mathcal{H}_n = \sum_{k=1}^{\infty}\lambda_k\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x_i)\{f_\theta(x_i) - f_{\theta^*}(x_i)\}\psi_k(x_i)\right]^2,$$

and thus $\mathcal{H}_n \to^{\mathrm{p}} \mathcal{J}$.

The constraint that is implied by $d \in I(\theta)$ on r can be relaxed to

$$\min_{\|r - f_\theta + y\| \leqslant 1} \frac{\mathcal{I}_n}{\mathcal{H}_n}\frac{2}{n}\sum_{i=1}^{n} r(x_i)\{f_\theta(x_i) - f_{\theta^*}(x_i)\}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x_i) - 2\int r(x)\{f_\theta(x) - f_{\theta^*}(x)\}\mathrm{d}\mu(x)$$
$$\leqslant l(y, f_{\theta^*}) - l(y, f_\theta) + \frac{\mathcal{I}_n}{\mathcal{H}_n}\frac{2}{n}\sum_{i=1}^{n} r(x_i)\{f_\theta(x_i) - f_{\theta^*}(x_i)\}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x_i).$$

Leveraging the triangle inequality gives that $\|d\| \leqslant 1$ gives $\|r\| \leqslant 1 + \|y - f_\theta\|$, which is true only if

$$(\mathcal{G}_n \quad \mathcal{G})\begin{pmatrix}\mathcal{H}_n & \mathcal{I}_n \\ \mathcal{I}_n & \mathcal{J}\end{pmatrix}^{-1}\begin{pmatrix}\mathcal{G}_n \\ \mathcal{G}\end{pmatrix} \leqslant 1 + \|y - f_\theta\|.$$

Using the 2×2 matrix inversion formula, let $\mathcal{K}_n = \mathcal{J} - \mathcal{H}_n^2/\mathcal{I}_n$:

$$\begin{pmatrix} \mathcal{H}_n & \mathcal{I}_n \\ \mathcal{I}_n & \mathcal{J} \end{pmatrix}^{-1} = \begin{pmatrix} \mathcal{H}_n^{-1} + \dfrac{1}{\mathcal{K}_n}\dfrac{\mathcal{I}_n^2}{\mathcal{H}_n^2} & -\dfrac{1}{\mathcal{K}_n}\dfrac{\mathcal{I}_n}{\mathcal{H}_n} \\ -\dfrac{1}{\mathcal{K}_n}\dfrac{\mathcal{I}_n}{\mathcal{H}_n} & \dfrac{1}{\mathcal{K}_n} \end{pmatrix};$$

thus

$$\mathcal{G}_n^2\mathcal{H}_n^{-1} + \frac{1}{\mathcal{K}_n}\left(\mathcal{G} - \frac{\mathcal{I}_n}{\mathcal{H}_n}\mathcal{G}_n\right)^2 \leqslant 1 + \|y - f_\theta\|$$

and this implies that

$$\left| \int r(x)\{f_\theta(x) - f_{\theta^*}(x)\}\mathrm{d}\mu(x) - \frac{\mathcal{I}_n}{\mathcal{H}_n}\frac{1}{n}\sum_{i=1}^{n} r(x_i)\{f_\theta(x_i) - f_{\theta^*}(x_i)\}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x_i) \right| \leqslant \sqrt{(1 + \|y - f_\theta\|)}\sqrt{\mathcal{K}_n}.$$

Let

$$\mathcal{L}_n := 2\sqrt{(1 + \|y - f_\theta\|)}\sqrt{\mathcal{K}_n},$$

where we have shown that $\mathcal{L}_n \to^{\mathrm{p}} 0$.

Maximizing the quadratic objective with our single linear constraint (Wright and Nocedal (1999), chapter 16) gives

$$\mathcal{D}_n \geqslant (l(y, f_\theta) - l(y, f_{\theta^*}) - \mathcal{L}_n)_+^2 \Big/ \frac{\mathcal{I}_n^2}{\mathcal{H}_n^2}\frac{1}{n}\sum_{i=1}^{n}\{f_\theta(x_i) - f_{\theta^*}(x_i)\}^2\left\{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x_i)\right\}^2.$$

where '$(\cdot)_+$' is the positive part. Thus

$$\mathcal{F}_n := \frac{\mathcal{I}_n^2}{\mathcal{H}_n^2}\frac{1}{n}\sum_{i=1}^{n}\{f_\theta(x_i) - f_{\theta^*}(x_i)\}^2\left\{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x_i)\right\}^2 \xrightarrow{\mathrm{p}} \int \{f_\theta(x) - f_{\theta^*}(x)\}^2\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\mathrm{d}\mu(x) =: \mathcal{F},$$

and

$$\mathcal{E}_n := (l(y, f_\theta) - l(y, f_{\theta^*}) - \mathcal{L}_n)_+^2 \xrightarrow{\mathrm{p}} \{l(y, f_\theta) - l(y, f_{\theta^*})\}^2 =: \mathcal{E}.$$

Noting that assumption 4 gives $\mathrm{d}\mu(\cdot)/\mathrm{d}\nu < M$ finishes the result.

B.5. Proof of theorem 6

Let

$$\mathrm{IL}(\theta) = \left\{ d \in D \text{ such that } \int d(x)\nabla f_\theta(x)\mathrm{d}\mu(x) = 0 \right\}.$$

This result follows the starting arguments of the proof of lemma 1. The only exception is that $\mathcal{D}_n$ needs to be replaced with

$$\mathcal{M}_n := \min_{d \in \mathrm{IL}(\theta)} \frac{1}{n}\sum_{i=1}^{n}\{d(x_i) + f_t(x_i) - y(x_i)\}^2.$$

Lemma 5. There is some $\mathcal{M} > \sigma^2$ such that, for all $\epsilon > 0$,

$$\mathbb{P}(\mathcal{M}_n \leqslant \mathcal{M} - \epsilon) \to 0.$$

Proof. Let $\theta^w = w\theta^* + (1 - w)\theta$. For w sufficiently close to 0, this is always in set Θ by the openness condition. Convexity of the scoring metric gives that

$$l(y, f_{\theta^w}) \leqslant wl(y, f_{\theta^*}) + (1 - w)l(y, f_\theta)$$

which can be rewritten as

$$\frac{l(y, f_{\theta^w}) - l(y, f_\theta)}{w} \leqslant l(y, f_{\theta^*}) - l(y, f_\theta).$$

Call

$$g(x) = \nabla f_\theta(x) \frac{\theta - \theta^*}{\|\theta - \theta^*\|_2},$$

where '$\|\cdot\|_2$' represents the L^2-norm. Letting $w \to 0$ yields

$$2 \int \{y(x) - f_\theta(x)\} g(x) \mathrm{d}\mu(x) \geqslant l(y, f_\theta) - l(y, f_{\theta^*}).$$

Replacing $f_\theta - f_{\theta^*}$ with g in the proof of lemma 4 gives the result with

$$\mathcal{M} = \sigma^2 + \frac{\{l(y, f_\theta) - l(y, f_{\theta^*})\}^2}{\int g(x)^2 \mathrm{d}\mu(x)/\mathrm{d}\nu\mathrm{d}\mu(x)}.$$

The only chore is to make sure that $\int g(x)^2 \mathrm{d}\mu(x)/\mathrm{d}\nu\mathrm{d}\mu(x) \neq 0$, which is taken care of by assumption.

Appendix C: Proof of theorem 7

Let $\delta = y - f_\theta$ and $I(\theta, \eta)$ represent the function space $I(\theta)$ when η is used. Let $c_n = 1 - \hat{\eta}_n / \|\delta\|$; then $(1 - c_n)\delta(\cdot) \in I(\theta, \hat{\eta}_n)$ and we have that

$$\min_{d \in I(\theta, \hat{\eta}_n)} \hat{l}(\text{data}, f_\theta + d) \leqslant \sum_{i=1}^n c_n^2 \delta(x_i)^2 + 2 \sum_{i=1}^n c_n \delta(x_i) e_i + \sum_{i=1}^n e_i^2.$$

By our assumption on the bounded fourth moment on e_is, $n^{-1/2} \Sigma_{i=1}^n (e_i^2 - \sigma^2)$ converges in distribution to a normal distribution with mean 0 where $n^{-1/2}\{q_n(\alpha) - \sigma^2\}$ is the $(1 - \alpha)$-quantile. We also have that, by the law of large numbers,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n c_n^2 \delta(x_i)^2 + \frac{2}{\sqrt{n}} \sum_{i=1}^n c_n \delta(x_i) e_i$$

goes to 0 in probability. Slutzky's theorem then gives the result.

References

Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J. and Walsh, D. (2007a) Computer model validation with functional output. *Ann. Statist.*, **35**, 1874–1906.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-H. and Tu, J. (2007b) A framework for validation of computer models. *Technometrics*, **49**, 138–154.

Beale, E. M. L. (1960) Confidence regions in non-linear estimation (with discussion). *J. R. Statist. Soc.* B, **22**, 41–88.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A. and Ritov, Y. (1998) *Efficient and Adaptive Estimation for Semiparametric Models*, vol. 2. New York: Springer.

Box, G. and Coutie, G. (1956) Application of digital computers in the exploration of functional relationships. *Proc. IEE* B, **103**, 100–107.

Craig, P. S., Goldstein, M., Seheult, A. H. and Smith, J. A. (1997) Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments. In *Case Studies in Bayesian Statistics* (eds C. Gatsonis, J. S. Hodges, R. E. Kass, R. McCulloch, P. Rossi and N. D. Singpurwalla), pp. 37–93. New York: Springer.

Efron, B. (2005) Bayesians, frequentists, and scientists. *J. Am. Statist. Ass.*, **100**, 1–5.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Ass.*, **102**, 359–378.

Goldstein, M. and Rougier, J. (2006) Bayes linear calibrated prediction for complex systems. *J. Am. Statist. Ass.*, **101**, 1132–1143.

Grimmett, G. and Stirzaker, D. (2001) *Probability and Random Processes*. Oxford: Oxford University Press.

Han, G., Santner, T. J. and Rawlinson, J. J. (2009) Simultaneous determination of tuning and calibration parameters for computer experiments. *Technometrics*, **51**, 464–474.

Hettich, R. and Kortanek, K. O. (1993) Semi-infinite programming: theory, methods, and applications. *SIAM Rev.*, **35**, 380–429.
Higdon, D., Gattiker, J., Williams, B. and Rightley, M. (2008) Computer model calibration using high-dimensional output. *J. Am. Statist. Ass.*, **103**, 570–583.
Joseph, V. R. and Melkote, S. N. (2009) Statistical adjustments to engineering models. *J. Qual. Technol.*, **41**, 362–375.
Kennedy, M. C. and O'Hagan, A. (2001) Bayesian calibration of computer models (with discussion). *J. R. Statist. Soc.* B, **63**, 425–464.
Koltchinskii, V. and Giné, E. (2000) Random matrix approximation of spectra of integral operators. *Bernoulli*, **6**, 113–167.
Owen, A. and Zhou, Y. (2000) Safe and effective importance sampling. *J. Am. Statist. Ass.*, **95**, 135–143.
Plumlee, M. (2017) Bayesian calibration of inexact computer models. *J. Am. Statist. Ass.*, **112**, 1274–1285.
Plumlee, M., Joseph, V. R. and Yang, H. (2016) Calibrating functional parameters in the ion channel models of cardiac cells. *J. Am. Statist. Ass.*, **111**, 500–509.
Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989) Design and analysis of computer experiments. *Statist. Sci.*, **4**, 409–423.
Santner, T. J., Williams, B. J., Notz, W. and Notz, W. I. (2003) *The Design and Analysis of Computer Experiments*, 1st edn. New York: Springer.
Schölkopf, B., Herbrich, R. and Smola, A. J. (2001) A generalized representer theorem. In *Proc. Int. Conf. Computational Learning Theory* (eds D. Helmbold and B. Williamson), pp. 416–426. New York: Springer.
Scholkopf, B. and Smola, A. J. (2001) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge: MIT Press.
Tarantola, A. and Valette, B. (1982) Generalized nonlinear inverse problems solved using the least squares criterion. *Rev. Geophys.*, **20**, 219–232.
Tuo, R. and Wu, C. F. J. (2015) Efficient calibration for imperfect computer models. *Ann. Statist.*, **43**, 2331–2352.
Tuo, R. and Wu, J. C. (2016) A theoretical framework for calibration in computer models: parametrization, estimation and convergence properties. *J. Uncertnty Quant.*, **4**, 767–795.
Vernon, I., Goldstein, M. and Bower, R. (2014) Galaxy formation: Bayesian history matching for the observable universe. *Statist. Sci.*, **29**, 81–90.
Wahba, G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc.* B, **40**, 364–372.
Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.
Wendland, H. (2004) *Scattered Data Approximation*. New York: Cambridge University Press.
Wong, R. K. W., Storlie, C. B. and Lee, T. C. M. (2017) A frequentist approach to computer model calibration. *J. R. Statist. Soc.* B, **79**, 635–648.
Wright, S. J. and Nocedal, J. (1999) *Numerical Optimization*, 1st edn. New York: Springer.
Xie, X., Kou, S. and Brown, L. D. (2012) SURE estimates for a heteroscedastic hierarchical model. *J. Am. Statist. Ass.*, **107**, 1465–1479.