5-2022

Advancements in Gaussian Process Learning for Uncertainty Quantification

John C. Nicholson
jcnicho@clemson.edu

Advancements in Gaussian Process Learning for Uncertainty Quantification

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical Statistics

---

by
John Nicholson
May 2022

Accepted by:
Dr. Peter Kiessler, Committee Chair
Dr. Andrew Brown, Committee Co-chair
Dr. Robert Lund
Dr. Christopher Mcmahan

Abstract

Gaussian processes are among the most useful tools in modeling continuous processes in machine learning and statistics. The research presented provides advancements in uncertainty quantification using Gaussian processes from two distinct perspectives. The first provides a more fundamental means of constructing Gaussian processes which take on arbitrary linear operator constraints in much more general framework than its predecessors, and the other from the perspective of calibration of state-aware parameters in computer models.

If the value of a process is known at a finite collection of points, one may use Gaussian processes to construct a surface which interpolates these values to be used for prediction and uncertainty quantification in other locations. However, it is not always the case that the available information is in the form of a finite collection of points. For example, boundary value problems contain information on the boundary of a domain, which is an uncountable collection of points that cannot be incorporated into typical Gaussian process techniques. In this paper we construct a Gaussian process model which utilizes reproducing kernel Hilbert spaces to unify the typical finite case with the case of having uncountable information by exploiting the equivalence of conditional expectation and orthogonal projections. We discuss this construction in statistical models, including numerical considerations and a proof of concept.

State-aware calibration is a novel approach in describing the relationship be-

tween properties of a system and experimental data by allowing calibration parameters to vary across the input domain as functions rather than remaining constant. Typical formulations in literature on the subject assume that it is already known whether calibration parameters are state-aware, but this is likely not the case in practice. Making incorrect assumptions on whether parameters are state-aware can produce confounding which misrepresents properties of a system and increases prediction error. We propose a means of determining the state of parameters which leverages the effect of the covariance function of Gaussian processes on their variation throughout the parameter space. We then apply the methodology to the analysis of interphase properties of composite materials and compare our results with previous studies

Dedication

I would like dedicate this work to my parents, who provided me with a solid foundation of overwhelming support when it was most needed. I would also like to dedicate this work to my partner who inspired in me the courage and strength to embrace the unknown and the temporary discomfort that often followed in order to realize the goals for which I strove. Lastly, I would like to dedicate this work to my guitar instructor and friend David Stevenson, who provided for me an outlet from the difficulties of life in graduate school; an outlet which I will carry with me for the rest of my life.

Table of Contents

List of Tables

List of Figures

Introduction

Gaussian process metamodels are a very popular means of approximating high fidelity processes where limited information is available, or pointwise computation carries a high computational expense. For example, suppose one is interested in a function f which can be computed from the approximate solution to a partial differential equation based upon a collection of tuning parameters specified before solving the equation. Since partial differential equations carry a notoriously high computational expense, evaluating the function directly can be infeasible directly. In this scenario, Gaussian process metamodels can be employed by evaluating f on a grid of tuning parameters and training hyperparameters which represent properties of the process, e.g. smoothness, or variation. The application of Gaussian process regression to approximate functions dates back to the work of Daniel Krige in a geospatial application to determine the distribution in the amount of gold across mines in South Africa in 1951 [Krige, 1951], and the theory was later introduced more formally by Matheron in the 1960s [Matheron, 1963]. However, Gaussian processes were not introduced in a more broad context until the work of Sacks [Sacks et al., 1989b], who constructed the more modern formulation of Gaussian process regression seen today.

Computer experiments are a popular application for Gaussian process regression, due to their wide applicability for experimental design [Jones et al., 1998,

Williams et al., 2011], quantifying model bias [Kennedy and O'Hagan, 2001b, 1998], and effectiveness in regression settings to speed up complex computer simulations. Within the broad umbrella of computer experiments, uncertainty quantification is an important concept which one may consider as the characterization or process of reducing uncertainties that exist in a given system, and includes concepts such as model calibration and model inadequacy. The focus of this dissertation addresses both of these topics from two separate vantage points.

Part I of this dissertation addresses model inadequacy in Gaussian process regression, and provides a novel method of imposing complex constraints that standard kriging methods are not equipped to handle. Recent work in the area of constraining Gaussian processes has included imposing upper and lower bounds, monotonicity constraints, differential equation constraints, and boundary constraints [Swiler et al., 2020]. Our work was dedicated to providing a significantly more general framework for imposing linear operator constraints (e.g. differential equation, boundary, or otherwise) using theory from Reproducing Kernel Hilbert Spaces, which was then validated through a series of numerical examples. This part is broken into two separate chapters, the first of which provides a characterization of boundary constraints from the standpoint of conditional probability. This chapter was inspired by a Gaussian process regression application involving predicting material properties of composite materials [Arp et al., Submitted], where the solution of the material model was explicitly known on one particular boundary of the parameter space. The second chapter pf Part I is dedicated to providing a much more general framework which extends to arbitrary linear operator constraints, and also provides a more computationally accessible methodology for the practitioner.

While Part I addresses the incorporation of constraints upon Gaussian processes in computer models, Part II is dedicated to aiding in providing a more general

framework in which computer models can be calibrated. Computer model calibration involves the estimation of unobservable parameters of a system in a way which maximizes the predictive capability of computer code to reality. This idea dates back to the work of Berman and Nagy Berman and Nagy [1983] in 1983, but was not popularized by statisticians until the seminal work of Kennedy and O'Hagan Kennedy and O'Hagan [2001b] in 2001. More recent research in the domain of computer model calibration has involved allowing calibration parameters to vary as functions of input parameters, rather than being constant. These parameters are said to be "state-aware". There are many cases in which the phenomenon of non-constant calibration parameters have been observed and accounted for, resulting in great improvements in predictive capability Atamturktur et al. [2015], Chodora et al. [2020], Plumlee et al. [2015], Pourhabib et al. [2015], Ezzat et al. [2018].

However, all research performed using state-aware calibration parameters has made the assumption that the constant parameters and the state-aware parameters are known *a priori*. The research in Part II aims to remove this requirement in state-aware calibration models, and perform a selection routine which identifies each parameter as state-aware or constant in conjunction with performing calibration. Included in this work are several numerical examples, as well as a case study involving the inference of material properties of composite materials Arp et al. [Submitted] illustrating the utility of this methodology in practice.

Lastly, conclusions are made regarding the applications of the research in Parts I and II, and future directions of research are discussed.

Part I

Functionally Constrained Gaussian Processes

Chapter 1

A Kernel-Based Approach for Modelling Gaussian Processes with Functional Information

1.1 Introduction

Gaussian processes [Rasmussen et al., 2006] are popular tools among statisticians and engineers for modeling complex problems because of their flexibility, simplicity, and their ability to quantify uncertainty. As Gaussian processes have become more popular in practice, there is an increased demand to modify Gaussian processes to possess certain characteristics. Swiler et. al. Swiler et al. [2020] give several such possibilities to implement bound constraints, monotonicity constraints, differential equation constraints, and boundary condition constraints.

In differential equations, boundary constraints on the actual values of the solution are called Dirchlet boundary conditions (as opposed to, e.g., Neumann boundary conditions which specify values of the derivatives). This is a common setting for mod-

eling GPs. In a more general scenario, however, one may simply have knowledge of a process on a subset of the domain. This does not necessarily fit under the umbrella of "boundary conditions", as the knowledge of the process may not be on the boundary and/or the process may not to be known to satisfy a differential equation. In this paper we propose a novel adaptation of a large class of Gaussian processes which have known, fixed values on an arbitrary subset of the domain. For simplicity, we will refer to this notion throughout the paper as "boundary constraints" while recognizing that the methodology is not limited to the boundary.

As motivation, consider the following materials science application. Finite element models can be used to predict the strength of composite materials consisting of a polymer matrix and a filler material consisting of embedded spherical particles [Arp et al., Submitted]. There are seven parameters contributing to variations in strength, six of which determine properties of the filler and interactions between the filler and the matrix. The code to run the finite element model is too expensive to run directly, so Gaussian process models can serve as an approximation of the model given model runs throughout the domain. However, when there is no filler in the material, the strength of the composite is simply the strength of the polymer, which is a control parameter. Therefore, the strength of the composite is already known on a six-dimensional subset of the seven-dimensional domain. In an ideal setting one would be able to use that information in totality to improve the Gaussian process model. This information, though, cannot be captured via conditioning on a finite-dimensional multivariate Gaussian distribution.

Given that infinitely many points are available in this scenario, one may suggest selecting a sufficient number of discrete points so that prediction error on this subset is below a certain threshold. For instance, the standard rule of thumb for choosing the sample size in a computer experiment is $10d$ where d is the dimension

of the domain Loeppky et al. [2009]. However, this rule is given in the context of computer experiments, where computing computer model runs can be very time consuming. Given the application stated above where there is no computational cost associated with the information, this may not be the best approach. A more tailored approach to choosing the necessary sample size given an error threshold is given by Harari et al. [2017], who consider sample size as a a random variable whose distribution is determined by the prior distribution on the parameterization of the covariance kernel family used. Though useful from a theoretical perspective, practically this would require strong prior knowledge of the parameter values, which is not likely to be known. Ultimately, one may simply check the prediction accuracy based upon various sample sizes and choose an appropriate sample size based upon trial and error. However, this still raises the question of how these points are distributed throughout the domain. Our interest here is thus a method for using Gaussian processes to capture information on an arbitrary subset of the domain in a more principled way.

There exist in the literature several proposed approaches for solving simplified versions of this problem. Solin and Särkkä [2019] suggested modifying an analytic stationary covariance function by approximation with a collection of functions which vanish on the boundary of the domain. The basis functions used were solutions to the eigenvalue problem for the homogeneous Laplace equation. Lange-Hegermann [2020] used pushforward mappings to modify Gaussian processes to satisfy homogeneous linear operator constraints, including boundary constraints. One particular pushforward of a Gaussian process $\mathbb{X}$ is of the form $\rho\mathbb{X}$, where $\rho : \mathbb{R}^d \to [0, 1]$. The author suggested choosing ρ so that $\rho \equiv 0$ on the boundary as a means of satisfying the constraint. Tan [2016] several years earlier developed an explicit construction representative of the reasoning from Lange-Hegermann [2020], and developed a mean function which permits nonzero constant boundary conditions. Though these meth-

ods have proven reasonable and effective under certain circumstances, none are able to handle truly general boundary conditions.

The reasoning behind our construction follows from a more probabilistic perspective, in which fixing the value of a Gaussian process at certain points can be considered as computing the conditional distribution. For Gaussian distributions, computing conditional distributions is very straightforward in finite dimensions. But, for cases in which the value is assumed to be fixed on an uncountable subset containing infinitely many points, it is not straightforward to compute the conditional distribution. Our approach is to consider conditional expectation as an orthogonal projection, and so computing the conditional distribution reduces to explicitly identifying the form of the projection, which we are able to do.

As an illustration, consider the following example. Let $T \subset \mathbb{R}^d$, and define $\mathbb{X}^0 = \{X_s^0;\, s \in T\}$ to be a Gaussian field with mean function μ and covariance kernel k. Define $T_0 \subset T$ to be a finite collection of points, $T_0 = \{t_1, ..., t_n\}$. It is well known that the stochastic process $\mathbb{X}^n = \{X_s^n;\, s \in T\}$ where $X_s^n = X_s^0 | (X_{t_1} = x_{t_1}, ..., X_{t_n} = x_{t_n})$ is a Gaussian process with mean function μ

$$\mu_0(s) = \mu(s) + k(s, \mathbf{t})k(\mathbf{t}, \mathbf{t})^{-1}(\mathbf{x} - \mu(\mathbf{t})), \tag{1.1}$$

and covariance kernel

$$k_0(s, s) = k(s, s) - k(s, \mathbf{t})k(\mathbf{t}, \mathbf{t})^{-1}k(\mathbf{t}, s), \tag{1.2}$$

where $\mathbf{t} = (t_1, \ldots, t_n)^\top$ and $s \in T$. This can be shown using orthogonal projections

and properties of Hilbert spaces. Define

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\Big(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_{22} \end{pmatrix} \Big),$$

Recalling that conditional expectation is an orthogonal projection, we can write $X_1 = (X_1 - PX_2) + PX_2$, for some linear operator P so that $\mathrm{Cov}(X_1 - PX_2, X_2) = 0$. In the finite dimensional case, P is simply a matrix. Expanding this covariance out, we see

$$0 = \Sigma_{12} - P\Sigma_{22},$$

Thus, P is the solution to $\Sigma_{12} = P\Sigma_{22}$. In the finite dimensional case, assuming X_2 is nondegenerate, we see $P = \Sigma_{12}\Sigma_{22}^{-1}$. Then, it follows that

$$\begin{aligned} E[X_1|X_2 = x_2] &= E[X_1 - PX_2|X_2 = x_2] + PE[X_2|X_2 = x_2] = \mu_1 - P\mu_2 + Px_2 \\ &= \mu_1 + P(x_2 - \mu_2), \\ V(X_1|X_2 = x_2) &= V(X_1 - PX_2) = \mathrm{Cov}(X_1, X_1 - PX_2) - \mathrm{Cov}(PX_2, X_1 - PX_2) \\ &= \mathrm{Cov}(X_1, X_1 - PX_2) = \Sigma_{11} - \Sigma_{12}P^\top. \end{aligned}$$

In the finite dimensional case, projection matrices typically can be computed explicitly. However, for infinite dimensional function spaces, projections are not as tractable. Therefore, our goal is to identify the distribution of a Gaussian process $\mathbb{X}^0$ conditional on $\mathbb{X}^0|_{T_0} = g_0$ with an orthogonal projection from one function space to another, describe the projection operator in a more meaningful way, and use it to compute the conditional distribution. Then, we discuss how one might derive this result from conditioning on a representative set of points, providing an avenue for showing that our results do indeed represent the conditional distribution.

This paper is organized as follows: Section 2 introduces some of the relevant information and notation that will be used throughout the paper, while Section 3 describes the construction of the mean and covariance of the process and illustrates how it can be derived by limits. Section 4 provides some probabilistic credence to the derivation in Section 3 including the connection to conditional expectation, and Section 5 is dedicated to illustrating how one might actually employ this approach in the context of more complex statistical models, as well as the notion of inexact or noisy information on T_0. Lastly, Section 6 discusses computational implementation, including several examples. We draw on several fundamental results from probability, functional analysis, and Reproducing Kernel Hilbert space theory that can be found in Kallenberg [1997], Lax [2002], and Paulsen and Raghupathi [2016] respectively.

1.2 Preliminaries

Construction of a conditional distribution revolves around the covariance function, which for the case of Gaussian processes will be studied as an element of a function space. As conditional expectation is an orthogonal projection in a Hilbert space, we need the covariance function to satisfy more properties than simply continuity or continuous differentiability. In this section we breifly review reproducing kernel Hilbert spaces (RKHS) and universal kernels, which play fundamental roles in our proposed construction. We use K to denote the integral operator in $L^2(T)$ associated with k, defined by

$$Kx(t) = \int_T k(s,t)x(s)ds,$$

where $T \subset \mathbb{R}^d$, denote the range of K as $R(K)$, and define $\langle \cdot\,, \cdot \rangle$ to be the standard inner product on L^2.

1.2.1 Reproducing Kernel Hilbert Spaces

For $t \in T$, define δ_t to be the Dirac functional which maps a function f to $f(t)$. The collection $\{\delta_t\}_{t\in T}$ are known as the evaluation functionals. These are commonly seen defined on the continuous functions $(C(T), ||\cdot||_\infty)$ where $||\cdot||_\infty$ denotes the supremum norm. As elements of the dual space, the evaluation functionals correspond to Dirac measures. The motivation behind reproducing kernel Hilbert spaces (RKHS) is to construct a Hilbert space so that the evaluation functionals are bounded, and thus identify uniquely with an element of the space itself. This is different from an L^2 space that contains congruence classes of functions in which two classes are equal if their representives are equal almost surely. Under this construction, the evaluation functionals are not even well-defined. Thus, to guarantee these functionals exist and are bounded, clearly the Hilbert space must contain only continuous functions. Therefore, a RKHS on T is defined to be a collection of functions $(\mathcal{H}(T), \langle\cdot,\cdot\rangle_{\mathcal{H}(T)})$ such that the evaluation functionals are bounded.

A kernel k defined on $T \times T$ has the reproducing property on $\mathcal{H}(T)$ if the representation of δ_t in $\mathcal{H}(T)$ is $k_t := k(\cdot, t)$ for each $t \in T$. Thus, it follows that the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}(T)}$ satisfies $f(t) = \langle f, k_t\rangle_{\mathcal{H}(T)}$, for any $f \in \mathcal{H}(T)$, $t \in T$. By the Moore-Aronszajn Theorem, each RKHS is identified uniquely with a kernel [2.14 Paulsen and Raghupathi, 2016, Theorem]. The space $\mathcal{H}(T)$ is constructed via closing the span of the functions $\{k_t\}_{t\in T}$ under $||\cdot||_{\mathcal{H}(T)}$, which implies $\{k_t\}_{t\in T} \subset \mathcal{H}(T)$. In addition, the norm of k_t can be calculated explicitly by

$$||k_t||_{\mathcal{H}(T)} = \langle k_t, k_t\rangle^{1/2}_{\mathcal{H}(T)} = k(t,t)^{1/2}.$$

Furthermore, for $s, t \in T$,

$$||k_s - k_t||^2_{\mathcal{H}(T)} = \langle k_s - k_t, k_s - k_t \rangle_{\mathcal{H}(T)} = k(s,s) - k(s,t) - k(t,s) + k(t,t)$$

Using this, we may note that if k is γ-Hölder continuous, then $||k_s - k_t||^2_{\mathcal{H}(T)} \leq B|s-t|^\gamma$, for some constant $B > 0$. This fact plays an important role in showing weak convergence of Gaussian processes to a limit in Section 4.

Mercer's theorem [Lax, 2002, pp. 343-344] plays a fundamental role in the theory of RKHS, which states that if k is a continuous kernel, then for any $s, t \in T$, there exists a non-negative sequence $\{\lambda_n\}$ and an orthonormal basis $\{e_n\}$ such that

$$k(s,t) = \sum_{n=1}^{\infty} \lambda_n e_n(s) e_n(t),$$

which as a series converges absolutely and uniformly. In addition, it can be shown that for $f, g \in \mathcal{H}(T)$,

$$\langle f, g \rangle_{\mathcal{H}(T)} = \sum_{n=1}^{\infty} \frac{\langle f, e_n \rangle \langle g, e_n \rangle}{\lambda_n},$$

and thus any $f \in \mathcal{H}(T)$ must satisfy $\sum_{n=1}^{\infty} \frac{\langle f, e_n \rangle^2}{\lambda_n} < \infty$. Therefore, we can generalize this to write $\mathcal{H}(T) = \{\sum_{n=1}^{\infty} a_n e_n : \{\frac{a_n}{\sqrt{\lambda_n}}\} \in \ell^2\}$.

Consider the square root operator $K^{1/2}$ of the integral operator K. Observing that $\sum_{n=1}^{\infty} \lambda_n = \int_T k(s,s) ds < \infty$, it follows that $K^{1/2}$ is a bounded, compact, self-adjoint operator [Lax, 2002], and can be represented by

$$K^{1/2} x = \sum_{n=1}^{\infty} \lambda_n^{1/2} \langle x, e_n \rangle e_n.$$

Since for $x \in L^2(T)$,

$$||K^{1/2}x||^2_{\mathcal{H}(T)} = \langle K^{1/2}x, K^{1/2}x\rangle_{\mathcal{H}(T)} = \sum_{n=1}^{\infty} \frac{\langle K^{1/2}x, e_n\rangle^2}{\lambda_n} = \sum_{n=1}^{\infty} \frac{(\sqrt{\lambda_n}\langle x, e_n\rangle)^2}{\lambda_n}$$
$$= \sum_{n=1}^{\infty} \langle x, e_n\rangle^2 \leq ||x||^2_{L^2},$$

where the last inequality is an application of Bessel's inequality, we see that $\text{im}(K^{1/2}) \subset \mathcal{H}(T)$. Thus, $K^{1/2}$ is bounded with respect to $||\cdot||_{\mathcal{H}(T)}$. In particular, if K has a trivial nullspace, the eigenvectors $\{e_n\}$ span $L^2(T)$, which allows us to substitute the inequality with an equality. If this is the case, $K^{1/2}$ is an isometric isomorphism between $L^2(T)$ and $\mathcal{H}(T)$. Hence, $K^{-1/2}$ exists and is bounded, and for $f, g \in \mathcal{H}(T)$,

$$\langle f, g\rangle_{\mathcal{H}(T)} = \langle K^{-1/2}f, K^{-1/2}g\rangle.$$

As motivated in the previous section, the projection occurs in both the mean and the covariance, meaning that the mean function should be an element of the RKHS. If the mean function is zero, this is trivially the case. Otherwise, it is difficult to check if a function is an element of $\mathcal{H}(T)$. As stated before, $\mathcal{H}(T) \subset C(T)$, but the converse is not true in general. For example, it has been shown that the RKHS associated with the square exponential kernel given by $k(s,t) = \exp\{-|s-t|^2\}$ does not contain any constant functions or polynomials in general [Ha Quang, 2010]. Ideally, the mean function is an element of the RKHS, but in the case which it is not, it is important that it can be well approximated by an element of the RKHS. The notion of universality is an important concept which describes the "coverage" of a kernel with respect to the continuous functions.

1.2.2 Universal Kernels

Since the space of uniformly continuous functions does not form a Hilbert space, there cannot exist a kernel such that $\mathcal{H}(T) = C(T)$. Thus, the universality of a kernel refers to the ability of the associated RKHS to approximate continuous functions. In particular, a kernel is said to be *universal* if $\mathcal{H}(T)$ is dense in $C(T)$ under the supremum norm $||\cdot||_\infty$, i.e. if any continuous function can be approximated to arbitrary precision by an element of $\mathcal{H}(T)$. Uninversal kernels were covered extensively by Micchelli et al. [2006], and our insight stems from this paper.

In statistics and machine learning, it is typical for one to use translation-invariant or stationary kernels when defining Gaussian processes, i.e kernels $\tilde{k}$ such that $\tilde{k}(s,t) = k(s-t)$ for some function k. Bochner's theorem [Lax, 2002, pp. 141-147] provides that $\tilde{k}$ is a kernel if and only if there exists a unique Borel measure ν on $\mathbb{R}^d$ satisfying for any $s \in \mathbb{R}^d$,

$$k(s) = \int_{\mathbb{R}^d} e^{i(s,t)} \nu(dt),$$

where $(\cdot\,,\cdot)$ denotes the dot product on $\mathbb{R}^d$. Defining ϕ to be so that $\phi(s)(t) = e^{i(s,t)}$, we see that

$$k(s_1 - s_2) = \int_{\mathbb{R}^d} e^{i(s_1,t)} e^{-i(s_2,t)} \nu(dt) = \int_{\mathbb{R}^d} \phi(s_1)(t)\overline{\phi(s_2)}(t)\nu(dt) = \langle \phi(s_1), \phi(s_2) \rangle.$$

Since ϕ does not depend on k, the properties of universality are completely determined by the measure ν.

Micchelli et al. [2006] show that if ν is absolutely continuous with respect to Lebesgue measure, then $\tilde{k}$ is universal. In this sense, any characteristic function of a continuous, symmetric probability distribution is universal. This fact alone provides

that since the square exponential kernel is the characteristic function of a zero mean Gaussian distribution, and the Matérn kernel is the characteristic function of the t-distribution, any square exponential and Matérn kernel is universal. Furthermore, any kernel of the form

$$k(s-t) = C\exp\left(-\sum_{i=1}^{d}\ell_i|s_i - t_i|^{p_i}\right),\ C, \ell_i, p_i > 0$$

is universal, as these are the characteristic functions of a subclass of symmetric stable distributions. Furthermore, non-stationary universal kernels may be constructed using the idea presented below.

Proposition 1.2.1. *Suppose k is a universal kernel, and q is a kernel of the form*

$$q(s,t) = \sigma(s)\sigma(t)k(s,t),$$

where σ is a continuous function on T satisfying $0 < m \leq \sigma(s) \leq M < \infty$ for some m and M for each $s \in \mathbb{R}^d$. Then, q is universal.

Proof. Since k is universal, $R(K)$ is dense in $C(T)$. Now, define $Q : L^2(T) \to L^2(T)$ by

$$Qx(t) = \int_X q(s,t)x(s)ds = \sigma(t)\int_X k(s,t)\sigma(s)x(s)ds.$$

Thus, we observe that $I_\sigma = \{\sigma f : f \in R(K)\} \subset R(Q)$. Therefore, it suffices to show that I_q is dense in $C(T)$ under $||\cdot||_\infty$. So, let $g \in C(T)$. Then, $\frac{g}{\sigma} \in C(T)$. So, for $\epsilon > 0$, choose $f \in R(K)$ so that $||f - \frac{g}{\sigma}||_\infty < \epsilon/M$. Then, for any $s \in T$,

$$|\sigma(s)f(s) - g(s)| = |\sigma(s)|\left|f(s) - \frac{g(s)}{\sigma(s)}\right| < \epsilon.$$

□

Thus, one may combine translation invariant kernels such as those given above with non-homogeneous variance conditions to generate a general class of non-stationary universal covariance kernels. In practice, working with a universal kernel is important since it is often not realistic to assume the function one is interested in estimating is in $\mathcal{H}(T)$. In the next section, the importance of universal kernels will become clear, as the solution relies on the computation of an RKHS inner product.

1.3 Deriving the Mean and Covariance

In this section, we define a mean and covariance for a Gaussian process $\mathbb{X}$ that results from the limit of mean and covariance functions obtained via conditioning on finitely many points in a subset of the domain. Section 4 will discuss the implications these results from a probabilistic perspective.

1.3.1 Derivation

Let $T \subset \mathbb{R}^d$ be compact, and $T_0 \subset T$ be an arbitrary set on which we assume information about a particular function g is known. Any Gaussian process which is fixed on T_0 must have a covariance function k_0 satisfying $k_0(s,t) = 0$, if one of $s, t \in T_0$. Denote by $\mathcal{H}(T)$ to be the RKHS associated with continuous and universal kernel k, and define

$$\mathcal{H}_0 = \{f \in \mathcal{H}(T) : \, f|_{T_0} \equiv 0\}.$$

It can be verified that $\mathcal{H}_0$ is a closed subspace of $\mathcal{H}(T)$, which implies that there exists an orthogonal projection $P : \mathcal{H}(T) \to \mathcal{H}_0$. $\mathcal{H}_0$ is also a RKHS with reproducing kernel $k_0(s,t) = (Pk)(s,t) = \langle Pk_s, k_t \rangle_{\mathcal{H}(T)}$ [Paulsen and Raghupathi, 2016, Theorem 2.5]. Furthermore, by properties of orthogonal projections, any function $f \in \mathcal{H}(T)$ which

satisfies $f = g$ on T_0 must be of the form

$$f = h_0 + g_\perp,$$

where $g_\perp \in \mathcal{H}_0^\perp$ is so that g has the unique representation $g = g_0 + g_\perp$, where $g_0, h_0 \in \mathcal{H}_0$. The Kolmogorov existence theorem permits the existence of a Gaussian process given a mean μ and kernel function k provided that the k is symmetric and positive semi-definite [Kallenberg, 1997, pp. 92]. As a corollary, we have the following result.

Theorem 1.3.1. *For a continuous covariance function k given, and $\mu \in \mathcal{H}(T)$, there exists a Gaussian process $\mathbb{X} = \{X_t;\, t \in T\}$ with mean $\mu_0 = P\mu + g_\perp$ and covariance Pk. In addition, $X_t = g_\perp(t)$ a.s. for each $t \in T_0$.*

Though such processes are guaranteed to exist, this result by itself is not very useful from a practical standpoint since it is unclear how one might compute Pf for arbitrary $f \in \mathcal{H}(T)$. Note that

$$\mathcal{H}_0^\perp = \overline{\text{Span}}(\{k_s;\, s \in T_0\}).$$

Hence, in the remainder of this section, we use $\mathcal{H}_0^\perp$ for computations, as the elements of this RKHS are more naturally described.

Let $k_\perp$ be the reproducing kernel for $\mathcal{H}_0^\perp$. Since $\mathcal{H}(T) = \mathcal{H}_0 \oplus \mathcal{H}_0^\perp$, it follows that $k = k_0 + k_\perp$ [Paulsen and Raghupathi, 2016, Corollary 5.5], and therefore $k_0 = k - k_\perp$. Naturally, one may compute $k_\perp(s,t) = \langle (I-P)k_s, k_t \rangle_{\mathcal{H}(T)}$. However, in this section, we will find a more tractable expression for $k_\perp$ which does not require the use of a projection operator.

First, suppose $T_0 = \{t_1, \ldots, t_n\}$, and define Q to be the orthogonal projection

onto $\mathcal{H}_0^\perp = \text{Span}(\{k_{t_1}, \ldots, k_{t_n}\})$. Although computing the conditional distribution in this case is trivial, we provide an alternative derivation which extends directly to a more general setting. Without loss of generality, assume that $\{k_{t_1}, \ldots, k_{t_n}\}$ is a linearly independent set so that the matrix $k(\mathbf{t}, \mathbf{t}) = (k(t_i, t_j))_{i,j=1}^n$ has full rank. Then, any $f \in \mathcal{H}(T)$ can be decomposed uniquely as $f = f_0 + Qf$, where $Qf_0 \equiv 0$, and

$$Qf = \sum_{i=1}^n a_i(f) k_{t_i},$$

where $Qf(t_i) = f(t_i)$ for each $i = 1, \ldots, n$ [Paulsen and Raghupathi, 2016, Corollary 3.5]. In turn, this implies the vector $\mathbf{a}(f) = (a_1(f), \ldots, a_n(f))^\top$ satisfies

$$\mathbf{a}(f) = k(\mathbf{t}, \mathbf{t})^{-1} f(\mathbf{t}).$$

Therefore, for $f_1, f_2 \in \mathcal{H}(T)$, the inner product on $\mathcal{H}_0^\perp$ for Qf_1, Qf_2 is computed as

$$\begin{aligned}\langle Qf_1, Qf_2 \rangle_{\mathcal{H}(T)} &= \Big\langle \sum_{i=1}^n a_i(f_1) k_{t_i}, \sum_{j=1}^n a_j(f_2) k_{t_j} \Big\rangle_{\mathcal{H}(T)} = \sum_{i=1}^n \sum_{j=1}^n a_i(f_1) a_j(f_2) \langle k_{t_i}, k_{t_j} \rangle_{\mathcal{H}(T)} \\ &= \sum_{i=1}^n \sum_{j=1}^n a_i(f_1) a_j(f_2) k(t_i, t_j) = \mathbf{a}(f_1)^(\mathbf{t}, \mathbf{t}) \mathbf{a}(f_2) = f(\mathbf{t})^\top k(\mathbf{t}, \mathbf{t})^{-1} f_2(\mathbf{t}).\end{aligned}$$

Using this formula, we see for $s_1, s_2 \in T$ that

$$k_\perp(s_1, s_2) = \langle Qk_{s_1}, Qk_{s_2} \rangle_{\mathcal{H}(T)} = k(s_1, \mathbf{t}) k(\mathbf{t}, \mathbf{t})^{-1} k(\mathbf{t}, s_2),$$

which implies that

$$k_0(s_1, s_2) = k(s_1, s_2) - k(s_1, \mathbf{t}) k(\mathbf{t}, \mathbf{t})^{-1} k(\mathbf{t}, s_2).$$

By setting $P = I - Q$, noting that $g_\perp = Qg$, and using the reproducing property, we

may write

$$\mu_0(s) = [(I-Q)\mu](s) + g_\perp(s) = \mu(s) + [Q(g-\mu)](s) = \mu(s) + \langle Qk_s, Q(g-\mu)\rangle_{\mathcal{H}(T)}$$
$$= \mu(s) + k(s,\mathbf{t})k(\mathbf{t},\mathbf{t})^{-1}(g(\mathbf{t}) - \mu(\mathbf{t})).$$

Note the formulae for μ_0 and k_0 correspond with those for the conditional distribution of $X_s|(X_{t_1} = g(t_1), \ldots, X_{t_n} = g(t_n))$, as expected.

Define the mapping $\psi : \mathcal{H}_0^\perp \to \mathbb{R}^n$ by $\psi(f) = f(\mathbf{t})$. Equipping $\mathbb{R}^n$ with the inner product

$$\langle \mathbf{f_1}, \mathbf{f}_2\rangle_0 = \mathbf{f_1}' k(\mathbf{t},\mathbf{t})^{-1}\mathbf{f_2},$$

it is clear that ψ is an isometry. This observation is emphasized because of the fact that even though elements of $\mathcal{H}_0^\perp$ are functions on T, they are completely determined by their values on T_0. In fact, $(\mathbb{R}^n, \langle\cdot,\cdot\rangle_0)$ is itself an RKHS with kernel $k(\mathbf{t},\mathbf{t})$, which is congruent to $k|_{T_0\times T_0}$. Therefore, in some sense one can think of ψ as a restriction to the set T_0. This is a key feature of our construction, one that holds true in the general case.

Now suppose T_0 is an arbitrary subset of T, and define $\mathcal{H}(T_0)$ to be the RKHS generated by k of functions defined on T_0. Although this space is different than $\mathcal{H}_0^\perp$, one can also write

$$\mathcal{H}(T_0) = \overline{\text{Span}}(\{k_s|_{T_0} : s \in T_0\}),$$

so in some sense $\mathcal{H}(T_0)$ and $\mathcal{H}_0^\perp$ are generated by the same functions, which leads to an important result.

Theorem 1.3.2. *There exists an isometric isomorphism between $\mathcal{H}_0^\perp$ and $\mathcal{H}(T_0)$.*

Proof. Define $\psi : \text{Span}(\{k_s;\, s \in T_0\}) \to \mathcal{H}(T_0)$ by $f \mapsto f|_{T_0}$. Clearly ψ is well-defined

and linear. Additionally, for arbitrary $n \geq 1$, $\{t_1, \ldots, t_n\} \subset T_0$, and $f = \sum_{i=1}^{n} a_i k_{t_i}$, we have

$$\begin{aligned}\langle f, f\rangle_{\mathcal{H}(T)} &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \langle k_{t_i}, k_{t_j}\rangle_{\mathcal{H}(T)} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \langle \psi(k_{t_i}), \psi(k_{t_j})\rangle_{\mathcal{H}(T_0)} \\ &= \left\langle \psi\Big(\sum_{i=1}^{n} a_i k_{t_i}\Big), \psi\Big(\sum_{j=1}^{n} a_j k_{t_j}\Big)\right\rangle_{\mathcal{H}(T_0)} = \langle \psi(f), \psi(f)\rangle_{\mathcal{H}(T_0)}.\end{aligned}$$

Therefore ψ is an isometry. Since $\mathrm{Span}(\{k_s;\, s \in T_0\})$ is dense in $\mathcal{H}_0^{\perp}$, there exists an isometry $\tilde{\psi} : \mathcal{H}_0^{\perp} \to \mathcal{H}(T_0)$ [Rudin, 1973, pp. 205] which is defined by limits, and therefore must also map $f \mapsto f|_{T_0}$. Clearly $\tilde{\psi}$ is one-to-one since $\tilde{\psi} f \equiv 0$ implies that $f|_{T_0} \equiv 0$, meaning that $f \in \mathcal{H}_0$. Since $f \in \mathcal{H}_0^{\perp}$, $f \equiv 0$.

Now, suppose $h \in \mathcal{H}(T_0)$. Then, there exists a Cauchy sequence $\{h_n\} \subset \mathrm{Span}(\{k_s|_{T_0};\, s \in T_0\})$ which converges to h. One may define $\{f_n\} \in \mathcal{H}_0^{\perp}$ so that $\tilde{\psi} f_n = h_n$. Since $\tilde{\psi}$ is an isometry, $\{f_n\}$ is Cauchy and therefore has a limit $f \in \mathcal{H}_0^{\perp}$. Then,

$$\tilde{\psi} f = \tilde{\psi}\Big(\lim_n f_n\Big) = \lim_n \tilde{\psi} f_n = \lim_n h_n = h,$$

which completes the proof. □

Thus, defining Q to be the projection from $\mathcal{H}(T)$ to $\mathcal{H}_0^{\perp}$, we have

$$Qf(s) = \langle Qf, k_s\rangle_{\mathcal{H}(T)} = \langle f|_{T_0}, k_s|_{T_0}\rangle_{\mathcal{H}(T_0)}.$$

Therefore, in the more general case, for $s_1, s_2 \in T$, one may write

$$\mu_0(s_1) = \mu(s_1) + \langle k_{s_1}|_{T_0}, (g-\mu)|_{T_0}\rangle_{\mathcal{H}(T_0)}, \tag{1.3}$$

$$k_0(s_1, s_2) = k(s_1, s_2) - \langle k_{s_1}|_{T_0}, k_{s_2}|_{T_0}\rangle_{\mathcal{H}(T_0)}. \tag{1.4}$$

Referring back to series representation of the RKHS inner product, the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}(T_0)}$, is much more tractable than the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0^\perp}$ due to the fact that the kernel on $\mathcal{H}(T_0)$ is known explicitly, whereas the kernel $k_\perp$ for $\mathcal{H}_0^\perp$ is computed via a projection which is less tractable from a numerical perspective. In the Section 1.6, we show that this formulation can be used in a numerical setting.

1.3.2 Limits

As mentioned in Section 1, one potential method of approximating the distribution of a Gaussian process conditional on all of T_0 is by conditioning on a representative finite subset of T_0. We will now show that the conditional mean and covariance computed from this method converge to μ_0 and k_0 given by (3-4) as the number of points conditioned on increases. By Theorem 3.2, it is acceptable to consider functions on T_0. Assume any function defined in this section is done so on T_0 unless otherwise specified. Let $D = \{t_n\}$ be a countable dense subset of T_0, and consider $\mathcal{K}_D := \overline{\text{Span}}(\{k_t;\, t \in D\})$. Note that since D is dense, for arbitrary $s \in T_0$, there exists a subsequence $\{t_{n_j}\} \subset D$ so that $k_s = \lim_{j\to\infty} k_{t_j}$. Therefore,

$$\{k_s;\, s \in T_0\} \subset \mathcal{K}_D \subset \mathcal{H}(T_0).$$

which implies that $\mathcal{K}_D = \mathcal{H}(T_0)$. As a consequence, for a given $f \in \mathcal{H}(T_0)$ and for $\epsilon > 0$, there exists an N_0 so that any interpolating approximation f_N by $\{k_{t_n}\}_{n=1}^N$ of f satisfies

$$||f_N - f||_{\mathcal{H}(T_0)} < \epsilon, \text{ if } N \geq N_0.$$

By defining Q_N as the orthogonal projection on $\text{Span}(\{k_{t_n}\}_{n=1}^N)$, this is statement is equivalent to saying that $Q_N f \to f$ for any $f \in \mathcal{H}(T_0)$.

Now, define μ_0^N and k_0^N as the mean and covariance resulting from conditioning on $\{t_1, \ldots, t_N\}$. Recalling the derivation of $\langle Q_N f_1, Q_N f_2 \rangle_{\mathcal{H}(T_0)}$, and noting that

$$\langle Q_N f_1, Q_N f_2 \rangle_{\mathcal{H}(T_0)} \to \langle f_1, f_2 \rangle_{\mathcal{H}(T_0)},$$

it follows that for $s_1, s_2 \in T$

$$\mu_0^N(s_1) = \mu(s_1) + \langle Q_N k_{s_1}, Q_N(g - \mu) \rangle_{\mathcal{H}(T_0)} \to \mu(s_1) + \langle k_s, g - \mu \rangle_{\mathcal{H}(T_0)} = \mu_0(s_1) \tag{1.5}$$

$$k_0^N(s_1, s_2) = k(s_1, s_2) - \langle Q_N k_{s_1}, Q_N k_{s_2} \rangle_{\mathcal{H}(T_0)} \to k(s_1, s_2) - \langle k_{s_1}, k_{s_2} \rangle_{\mathcal{H}(T_0)} = k_0(s_1, s_2). \tag{1.6}$$

Observe also that

$$\begin{aligned} k_0^N(s_1, s_2) - k_0(s_1, s_2) &= \langle k_{s_1}, k_{s_2} \rangle_{\mathcal{H}(T_0)} - \langle Q_N k_{s_1}, Q_N k_{s_2} \rangle_{\mathcal{H}(T_0)} \\ &= \langle (I - Q_N) k_{s_1}, (I - Q_N) k_{s_2} \rangle_{\mathcal{H}(T_0)}, \end{aligned}$$

which implies that $k_0^N - k_0$ is a positive kernel. In the sense of stochastic processes, this property implies that k_0 is a further reduction of variance from k_0^N. In fact, equations (5-6) correspond directly to equations (1-2) respectively. The next section we address the question of stochastic convergence.

1.4 Weak Convergence and a Probabilistic Perspective

One of the highlights of the previous section was showing that the finite dimensional distributions of a Gaussian process conditioned on N points converges to a limiting process provided that the mean function μ is in the RKHS associated with the covariance kernel and the dense set of points defines a function g which is also contained in the RKHS. Define the sequence of Gaussian processes $\{\mathbb{X}^N\}$ so that $\mathbb{X}^N$ has mean and covariance μ_0^N and k_0^N, and define $\mathbb{X}$ to be a Gaussian process with mean and covariance μ_0 and k_0. To show that the limit of the finite dimensional distributions defines a Gaussian process $\mathbb{X}$ such that $\mathbb{X}^N \Rightarrow \mathbb{X}$, it remains to show that $\{\mathbb{X}^N\}$ is tight.

As the setting for many applications desires continuous processes, it is important to ensure that sample paths of $\{\mathbb{X}^N\}$ are almost surely continuous for each $N \geq 0$.

Lemma 1.4.1. *Suppose that $\mathbb{X}$ is a Gaussian process with mean μ and covariance kernel k. If μ is continuous and k is $\gamma-$Hölder continuous on $\mathbb{R}^d \times \mathbb{R}^d$, then there is a version of $\mathbb{X}$ which almost surely continuous.*

Proof. We will use the Kolmogorov-Chentsov theorem [Kallenberg, 1997, pp. 35-36] which states that $\mathbb{X}$ has a continuous version on $\mathbb{R}^d$ taking on values in a complete metric space (S, ρ) if there exists $a, b > 0$ such that

$$E[\rho(X_s, X_t)^a] \leq |s - t|^{d+b},\ s, t \in \mathbb{R}^d.$$

Assume that $\mathbb{X}$ has zero mean and covariance as specified above. Define ρ to be the

Euclidean norm on $\mathbb{R}$, and recall that for any zero mean Gaussian random variable Z and any even integer a,

$$E[Z^a] = C_a E[Z^2]^{a/2},$$

where $C_a = \prod_{i=1}^{a/2}(2i-1)$. Defining a to be the smallest even integer strictly larger than $\frac{2d}{\gamma}$, we see for any $s, t \in \mathbb{R}^d$,

$$\begin{aligned} E[\rho(X_t, X_s)^a] = E[(X_t - X_s)^a] = C_a E[(X_t - X_s)^2]^{a/2} = C_a \big[k(t,t) - 2k(t,s) + k(s,s)\big]^{a/2} \\ \leq C|s-t|^{\gamma a/2} = C|s-t|^{d+(\gamma a/2 - d)}. \end{aligned}$$

Thus, selecting $b = \gamma a/2 - d$, and scaling ρ appropriately, we get the result for a zero mean process. Lastly, the non-zero mean process can be achieved by translating the process by the mean, repeating the procedure above, and noting that the sum of continuous functions is continuous.

□

It is indeed the case that $\{\mathbb{X}^N\}$ is tight if the conditions for the Kolmogorov-Chentsov theorem stated above are met uniformly on N [Kallenberg, 1997, pp. 35-36]. The theorem below provides conditions for the tightness of $\{\mathbb{X}^N\}$ to a Gaussian process $\mathbb{X}$ with mean function μ_0 and covariance kernel k_0

Theorem 1.4.2. *If the covariance kernel k is $\gamma-$Hölder continuous, k is universal on T_0 and $g|_{T_0}, \mu|_{T_0} \in \mathcal{H}(T_0)$, then $\{\mathbb{X}^N\}$ is tight in $(C(T), ||\cdot||_\infty)$.*

Proof. Recall the remark in Section 3 in which the mean and covariance of $\mathbb{X}^N$ written

μ^N and k^N can be defined as

$$\mu_0^N(s) = \mu(s) + \langle Q_N k_s, Q_N(g-\mu)\rangle_{\mathcal{H}(T_0)},$$
$$k_0^N(s,t) = k(s,t) - \langle Q_N k_s, Q_N k_t\rangle_{\mathcal{H}(T_0)}.$$

Now, observe that for $s_0 \in T$,

$$\begin{aligned}
|k_0^N(s_0,s) - k_0^N(s_0,t)| &\leq |k(s_0,s) - k(s_0,t)| + |\langle Q_N k_{s_0}, Q_N(k_s - k_t)\rangle_{\mathcal{H}(T_0)}| \\
&\leq C|s-t|^\gamma + ||Q_N k_{s_0}||_{\mathcal{H}(T_0)}||Q_N(k_s-k_t)||_{\mathcal{H}(T_0)} \\
&\leq C|s-t|^\gamma + ||k_{s_0}||_{\mathcal{H}(T_0)}||k_s-k_t||_{\mathcal{H}(T_0)} \\
&\leq C|s-t|^\gamma + ||k_{s_0}||_{\mathcal{H}(T_0)}||k_s-k_t||_{\mathcal{H}(T)} \\
&\leq C|s-t|^\gamma + C'|s-t|^{\gamma/2} \leq \tilde{C}|s-t|^{\gamma/2},
\end{aligned}$$

where the first inequality follows frome the triangle inequality, the final inequality follows form the boundedness of T, and $\tilde{C}$ does not depend on s_0 or N. Since k itself is $\gamma-$Hölder continuous, it follows that k_0^N is $\gamma/2-$Hölder continuous on $T \times T$ uniformly in N. Furthermore, $\mu^N \to \mu$ uniformly where we again use the fact that $\tilde{K}$ is uniformly $\gamma/2-$Hölder continuous on $\{Q_N(g-\mu)\}$. Therefore, $\{\mathbb{X}^N\}$ is tight.

□

Therefore, it follows that $\mathbb{X}^N \Rightarrow \mathbb{X}$ if the original mean function is continuous, and the covariance kernel is Hölder continuous. In particular, $\mathbb{X}$ is the Gaussian process $\mathbb{X}^0$ conditioned on $\mathbb{X}^0|_D = g$. One would like to extend this result to say that $\mathbb{X}$ is the Gaussian process conditioned on $\mathbb{X}^0|_{T_0} = g$. Since conditional expectation is determined by the $\sigma-$fields generated by the random elements, it suffices to show

that

$$\sigma\big(\{X_t^0\,;\, t \in D\}\big) = \sigma\big(\{X_t^0\,;\, t \in T_0\}\big).$$

This follows directly from the fact that for any sequence $\{t_{n_j}\}$ such that $t_{n_j} \to t$,

$$X_{t_{n_j}} \to X_t, \text{ a.s.}$$

Furthermore, since measurability under limits of functions is preserved, for any $t \in T_0$, X_t is $\sigma\big(\{X_t^0\,;\, t \in D\}\big)$-measurable. Thus, $\mathbb{X}$ is a version of the original stochastic process conditioned on $\mathbb{X}^0|_{T_0} = g$. To more aptly discuss the significance of this result, denote $\mathcal{F}_0 = \sigma\big(\{X_t^0\,;\, t \in T_0\}\big)$. Then, defining $F_g = \{X_t^0 = g(t);\, t \in T_0\} \in \mathcal{F}_0$, we may simply define $\mathbb{X}$ by $X_t = X_t^0|F_g$.

Now, speaking in more broad terms, suppose we define $X_t = E[X_t^0|\mathcal{F}_0]$. Since $\mathbb{X}^0$ is continuous, there exists a unique process up to a nullset $\mathcal{N}$ whose elements are defined above [Kallenberg, 1997, pp. 34]. Furthermore, $\mathbb{X}$ is an $\mathcal{F}_0$-measurable process which can be thought of as an predictor of $\mathbb{X}^0$ rather than g, which allows us to discuss the notion of optimality in prediction.

Theorem 1.4.3. *For any $\mathcal{F}_0$-measurable process $\hat{\mathbb{X}}$, it follows that for any $t \in T$,*

$$E\big[(X_t^0 - \hat{X}_t)^2\big] \geq E\big[(X_t^0 - X_t)^2\big].$$

The proof of this follows directly from the definition of conditional expectation. To illustrate the value of this observation, consider a simple Gaussian process model given by $X^0(x) = \mu(x) + W(x)$, where W is a centered Gaussian process with covariance kernel k, where it is of interest to predict X^0. Then, given the information of $\mathbb{X}^0$ on any subset of its domain, the predictive process containing the prior information of $\mathbb{X}^0$ which minimizes the mean square prediction error is $\mathbb{X}$.

1.5 Inexact Solutions and Noise

Throughout the past two sections, it has been assumed that g and μ constricted to T_0 are contained in $\mathcal{H}(T_0)$. Though necessary for our computations, this is actually a very limiting assumption as in general $\mathcal{H}(T_0)$ is very small relative to $C(T_0)$ [Vaart and Zanten, 2011]. We will see in the next section that this does not play much of a factor in a more practical setting provided that $\mathcal{H}(T_0)$ is dense in $C(T_0)$. Nevertheless, the model presented in the previous two sections is confined to a very basic Gaussian process model, and it is unclear based upon the previous sections how one may apply our method to more involved statistical models. This section is dedicated to showing how one might modify our approach when complexity is added into a Gaussian process model, illustrated through several different examples.

It is commonplace for Gaussian process computer models to have more than one source of uncertainty. For example, one may model correlated data y^0 as

$$y^0(x) = \mu(x) + \delta^0(x) + \varepsilon^0(x),$$

μ is a deterministic computer model, δ^0 refers to zero mean model bias [Kennedy and O'Hagan, 2001b], and ε^0 refers to zero mean error associated with collecting data, with δ^0 and ε^0 independent Gaussian processes. Suppose that the output y^0 is known explicitly on T_0 and is described by the function g. This would correspond to $\varepsilon^0 = 0$ and $\delta^0 = g - \mu$ on T_0 with zero variance. If ε^0 represents uncorrelated error, then one may use this information to update y^0 so that

$$y(x) = \mu_0(x) + \delta(x) + \varepsilon(x),$$

where μ_0 is defined as in Section 3, $\delta(x)$ has mean zero and covariance k_0, and ε

is a zero mean white noise Gaussian process whose variance on T_0 is zero. If ε^0 is correlated error, then one may perform the same modification on ε^0 as done for δ^0 provided that the covariance function for ε^0 is continuous.

As one can see, considering slight alterations in the overall structure of the model does not significantly alter our methodology if one assumes that the information on T_0 is known exactly. Now, we will consider more complicated case where information on T_0 is known less explicitly.

1.5.1 Handling information up to a white noise

Now, suppose the information on T_0 is known up the white noise ε at each point which is independent of $\mathbb{X}^0$. In other words, we want to compute the distribution of $\mathbb{X}$ where $\mathbb{X}|_{T_0} = \tilde{g} = g + \varepsilon$. There are several reasons for adding the white noise term, with the first being that it may not be the case that information is known completely on T_0. Another common reason to consider is that covariance matrices constructed from very smooth kernels (e.g. square exponential) can be very ill-conditioned, and so one adds a "nugget" term ensure stable computations [Ranjan et al., 2011]. Using this formulation, one may derive very similar results as in Section 3.

In either case, the covariance function becomes $\tilde{k}(s,t) = k(s,t) + \sigma^2 \mathbb{I}(s = t)$. Since this kernel is not continuous, the theory of RKHS cannot apply here in the sense that has been described in the previous sections. For $\mathbf{t} = (t_1, \ldots, t_n)$, the covariance matrix generated by $\tilde{k}$ is of the form $k(\mathbf{t}, \mathbf{t}) + \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix. One may naturally extend this to L^2 by defining the operator $\tilde{K} = K + \sigma^2 I$, where K is the standard integral operator and I is the identity operator, which are both defined on $L^2(T_0)$. However, here it is important to note that $\tilde{K}$ maps to $L^2(T_0)$

rather than a RKHS. Now, recall the representation of the RKHS inner product as

$$\langle f, g\rangle_{\mathcal{H}(T_0)} = \langle K^{-1/2} f, K^{-1/2}\tilde{g}\rangle_{T_0},$$

where $\langle\cdot,\cdot\rangle_{T_0}$ denotes the L^2 inner product on T_0. Using previous notation, eigenvalues and eigenvectors of $\tilde{K}$ are $\{\lambda_n + \sigma^2\}$ and $\{e_n\}$, and so one may represent $\tilde{K}$ as

$$\tilde{K}(\cdot) = \sum_{n=1}^{\infty} (\lambda_n + \sigma^2)\langle\,\cdot\,, e_n\rangle_{T_0} e_n.$$

Therefore, $\tilde{K}^{-1/2}$ can be represented by

$$\tilde{K}^{-1/2}(\cdot) = \sum_{n=1}^{\infty} \frac{1}{\sqrt{\lambda_n + \sigma^2}}\langle\,\cdot\,, e_n\rangle_{T_0} e_n.$$

Replacing $K^{-1/2}$ with $\tilde{K}^{-1/2}$, we may define a new inner product for $f_1, f_2 \in L^2(T_0)$ by

$$\langle f_1, f_2\rangle_{\tilde{K}} = \langle \tilde{K}^{-1/2} f_1, \tilde{K}^{-1/2} f_2\rangle_{T_0} = \sum_{n=1}^{\infty} \frac{\langle f_1, e_n\rangle_{T_0}\langle f_2, e_n\rangle_{T_0}}{\lambda_n + \sigma^2}.$$

Since any continuous function defined on T_0 is also an element of L^2, this definition is valid. Using this, it follows that $\mathbb{X}$ is Gaussian with posterior mean $\tilde{\mu}_0$ and posterior covariance $\tilde{k}_0$, which are defined in the same way as μ_0 and k_0, but replacing $\langle\cdot,\cdot\rangle_{\mathcal{H}(T_0)}$ with $\langle\cdot,\cdot\rangle_{\tilde{K}}$. Therefore, we define $\tilde{\mu}_0$ and $\tilde{k}_0$ by

$$\tilde{\mu}_0(s_1) = \mu(s_1) + \langle k_{s_1}, \tilde{g} - \mu\rangle_{\tilde{K}},$$
$$\tilde{k}_0(s_1, s_2) = k(s_1, s_2) - \langle k_{s_1}, k_{s_2}\rangle_{\tilde{K}}.$$

Note here that $\tilde{g}$ is a stochastic process, so in fact this definition is not only conditional on $\mathbb{X}^0|_{T_0}$, but on ε as well.

1.5.2 Handling Stochastic Information

Lastly, we consider the more general case where the information on T_0 is known up to a zero mean Gaussian process δ with covariance kernel q, which is again independent of $\mathbb{X}^0$. One may write this as finding the distribution of $\mathbb{X}$ where $\mathbb{X}|_{T_0} = g_\delta = g + \delta$. Then, again the covariance matrix in the finite case is given by $k(\mathbf{t}, \mathbf{t}) + q(\mathbf{t}, \mathbf{t})$, and the associated RKHS with $k + q$ is the sum $\mathcal{H}_{k+q}(T_0) = \mathcal{H}(T_0) + \mathcal{H}_q(T_0)$, where $\mathcal{H}_q(T_0)$ is the RKHS associated with q. In general $\mathcal{H}(T_0) \cap \mathcal{H}_q(T_0) \neq \{0\}$, so the sum is not direct, which makes determining the inner product on $\mathcal{H}_{k+q}(T_0)$ as the sum of its constituents nontrivial. However, it is the case that any element of $\mathcal{H}(T_0)$ or $\mathcal{H}_q(T_0)$ is also an element of $\mathcal{H}_{k+q}(T_0)$, and therefore the mean and covariance are again defined as in (5-6), but replacing $\langle \cdot, \cdot \rangle_{\mathcal{H}(T_0)}$ with $\langle \cdot, \cdot \rangle_{\mathcal{H}_{k+q}(T_0)}$. Therefore, we define μ_0 and k_0 by

$$\mu_0(s_1) = \mu(s_1) + \langle k_{s_1}, g_\delta - \mu \rangle_{\mathcal{H}_{k+q}(T_0)},$$
$$k_0(s_1, s_2) = k(s_1, s_2) - \langle k_{s_1}, k_{s_2} \rangle_{\mathcal{H}_{k+q}(T_0)}.$$

As mentioned in Section 5.1, this definition also is conditional on δ as well as $\mathbb{X}^0|_{T_0}$.

1.6 Numerical Implementation

The previous sections have shown that one may construct a Gaussian process which has zero variation on an arbitrary select subset T_0 of the domain, and define its mean and covariance functions in terms of an RKHS inner product. However, in practice, the RKHS inner product in the general case cannot be computed exactly. Here we discuss techniques for computing the inner products, followed by examples.

1.6.1 Computation of RKHS Inner Product

Recall that the RKHS norm is given in terms of the spectral decomposition $\{(\lambda_n, e_n)\}$ of the integral operator K_{T_0}, which in general must be computed numerically. Then, the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}(T_0)}$ is approximated via the bilinear form $a_N(\cdot, \cdot)$, given by

$$a_N(f, g) = \sum_{n=1}^{N} \frac{\langle f, e_n \rangle_{T_0} \langle g, e_n \rangle_{T_0}}{\lambda_n}.$$

Naturally, the form of $a_N(\cdot, \cdot)$ does not permit a convergence independent of the selection of arbitrary $f, g \in \mathcal{H}(T_0)$. However, a uniform-type convergence can be established for the family $\mathcal{K} := \{k_t : t \in T\}$.

Proposition 1.6.1. *The collection of bilinear forms $\{a_N\}$ converge uniformly to $\langle \cdot, \cdot \rangle_{\mathcal{H}(T_0)}$ on $\mathcal{K} \times \mathcal{K}$.*

Proof. Define $F_N, F : T \times T \to \mathbb{R}$ by $F_N(s, t) = a_N(k_s, k_t)$ and $F(s, t) = \langle k_s, k_t \rangle_{\mathcal{H}(T_0)}$. It is clear that $F_N \to F$ pointwise, so it suffices to show that $\{F_N\}$ is equicontinuous. Defining Q_N to be the projection from $\mathcal{H}(T_0)$ to $\text{Span}(\{e_n\}_{n=1}^N)$, it is clear that

$$F_N(s, t) = \langle Q_N k_s, Q_N k_t \rangle_{\mathcal{H}(T_0)},$$

and so equicontinuity follows directly from the fact that F is Hölder continuous $\{Q_N\}$ are uniformly bounded by the identity operator.

□

Thus, given a function $\mu \in \mathcal{H}(T_0)$, and a tolerance ϵ, one may select N so that

$$|a_N(f, g) - \langle f, g \rangle_{\mathcal{H}(T_0)}| < \epsilon,$$

for $f, g \in \mathcal{K} \cup \{\mu\}$, which suggests that using this methodology in an application

setting is indeed stable. Naturally, $\{(\lambda_n, e_n)\}_{n=1}^N$ need to be computed, and are done so by solving the eigenvalue problem

$$Kf = \lambda f.$$

Oya et al. [2009] discuss various methods of computing RKHS inner products using this formulation, and suggested using a Ritz-Rayleigh (RR) approach to compute the approximate spectral decomposition of K and inserting the approximate values $\{(\tilde{\lambda}_n, \tilde{e}_n)\}$ to compute the inner product. To summarize this approach, suppose that $A \in \mathbb{R}^{n\times n}$ is positive semidefinite, and $V \in \mathbb{R}^{p\times n}$ has orthonormal row vectors $\{v_1, \ldots, v_p\}$, where $p < n$. Then, the matrix

$$A_V = VAV^* \in \mathbb{R}^{p\times p}$$

is a positive semidefinite matrix, which can be written as $A_V = UD_\alpha U^*$ for an orthonormal matrix U and a diagonal matrix of eigenvalues D_α. This matrix has the property that if an eigenvalue e_i of A is in the span of V, then there is a corresponding eigenvector u of A_V such that

$$e_i = V^*u$$

with $u^*A_Vu = e_i^*Ae_i$. This algorithm also applies in an arbitrary Hilbert space, and is the basis for many numerical methods in applied mathematics. Naturally, the effectiveness depends upon the function basis used.

In the case of symmetric kernels, one may actually define the RKHS inner product in terms of Fourier transforms. Let $\tilde{k}(s-t) = k(s,t)$, and define $\mathcal{F}$ to be the Fourier operator. Then, for $f, g \in \mathcal{H}(T_0)$ Berlinet and Thomas-Agnan [2004] define

the RKHS inner product by

$$\langle f, g \rangle_{\mathcal{H}(T_0)} = \frac{1}{(2\pi)^{d_0/2}} \int_{\mathbb{R}^{d_0}} \frac{\mathcal{F}[f](\omega)\overline{\mathcal{F}[g](\omega)}}{\mathcal{F}[\tilde{k}](\omega)} d\omega.$$

Direct computations of $\langle f, g \rangle_{\mathcal{H}(T_0)}$ using this approach can potentially be expensive, but discrete Fourier approximations may prove useful in this scenario.

1.6.2 Numerically Verifying the Reproducing Property

Although the RKHS inner product cannot be explicitly calculated for arbitrary functions, the accuracy of any approximation method can be verified by utilizing the reproducing property. For example, it is always the case that for $f \in \mathcal{H}(T)$,

$$\langle f, k_t \rangle_{\mathcal{H}(T)} = f(t).$$

As shown in Section 3, one may approximate the inner product by computing the mean function of a Gaussian process conditioned on its value at several points in the domain. So, it also of interest to know how more spectral approaches such as those given in Section 6.1 compare with the interpolation method of reproducing f. As previously mentioned, it is unlikely that any continuous function f is an element of $\mathcal{H}(T)$. Thus, it is worth considering the effects of reproducing functions which are not elements of $\mathcal{H}(T)$ as well as those which are.

Assume that $T = [-1, 1]$, and $k(x, x') = \exp\{-|x - x'|^2\}$. Define $f_1, f_2 \in C[-1, 1]$ to interpolate the points $\{(x_j, y_j)\}_{j=1}^{J}$ (which are assumed to be unknown), where f_1 does so using the kernel as a basis, and f_2 does so using a polynomial basis. Thus, f_1, f_2 should have a similar appearance, but $f_1 \in \mathcal{H}(T)$, whereas f_2 is not. J is selected to be 6, $\{x_j\}$ are selected to be equidistant on $[-1, 1]$, and $\{y_j\}$

are randomly selected in $[-1, 1]$. Figure 1.1 indicates, as one may expect, that the difference between f_1 and f_2 in this type of setup is negligible. However, Figure 1.2 indicates that the RR method described in Section 6.1 significantly outperforms the standard interpolation method for f_2, suggesting that this method perhaps is better for reproducing functions which are not necessarily in the RKHS.

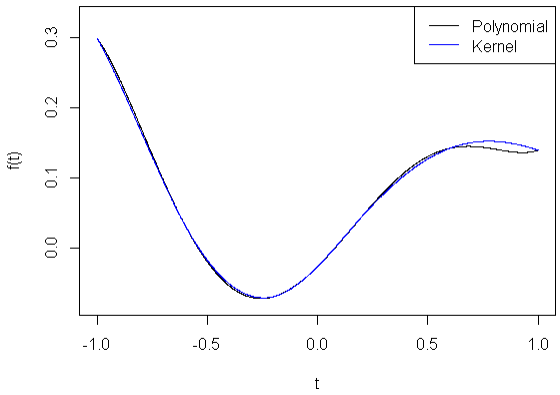


Figure 1.1: Plot showing f_1, f_2 constructed as described above. As one can see, the difference between the two functions is negligible.

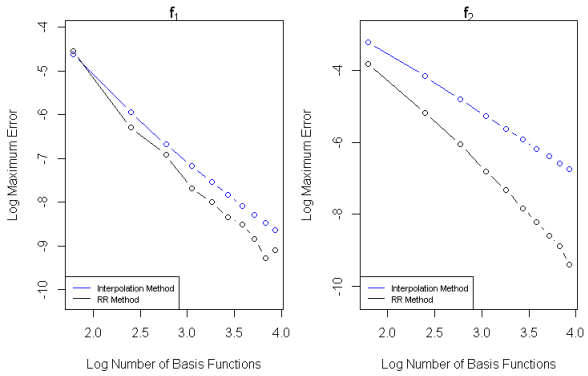


Figure 1.2: Plot showing the convergence rates of reproducing f_1 and f_2 via the RR method and the typical interpolation method seen in Gaussian process regression.

1.6.3 Numerical Examples

1.6.3.1 Boundary Conditions

As a basic application, let $T = [-1, 1]^2$ and define f by

$$f(t_1, t_2) = \frac{1}{2} e^{.2(t_1 - .5)^2} \sin\left(\frac{\pi t_1}{2}\right) + e^{-t_2^2} \cos\left(\frac{\pi t_2}{2}\right).$$

Assume that the value of f is known at M points of the domain, as well as on

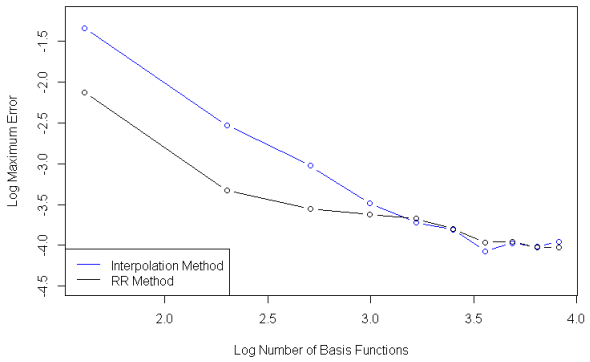
$$T_0 = \partial T.$$

Since T_0 has dimension one, one may define a parameterization $\ell : [-1, 1) \to T_0$ so that computations may be performed in one dimension. One practical issue with this however is that the function f is continuously differentiable on T, whereas the function $f \circ \ell$ is not differentiable on $\{-1, -1/2, 0, 1/2\}$.

To assess the accuracy of the method for different numbers of basis functions, the test data used is a collection of points on the set $\mathcal{T} = \{(.9t, .9s) : (t, s) \in \partial T\}$ and measure discrepancy based upon the the loss function

$$L(f, g) = \max_{t \in \mathcal{T}} |f(t) - g(t)|.$$

We select $M = 10$, where the points on the interior are chosen via a Latin Hypercube sampling scheme. Figure **??** shows the error as the number of basis functions increases. Observe that the log error flattens out, unlike what is observed in Figure 1.2 from reproducing the function. This can be thought of as a phenomenon where essentially all of useful information from the boundary has been extracted, leading to diminishing returns on predictive power with additional basis functions.



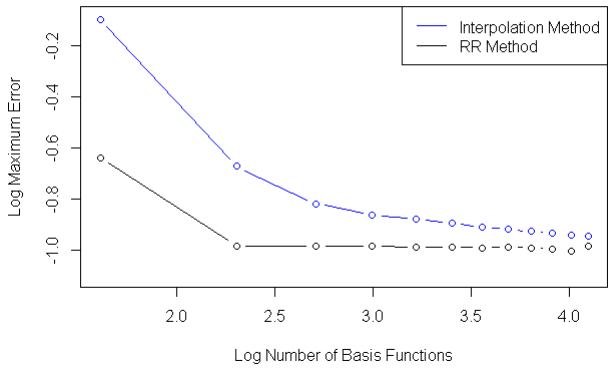
1.6.3.2 Diagonal Conditions

As mentioned previously, T_0 is not limited to the boundary, and can be any subset of T. In this example, assume that $T = [-1, 1]^2$, and let T_0 be the diagonal of T, i.e. $T_0 = \{(t, t) : t \in [-1, 1]\}$. Define f by

$$f(t_1, t_2) = t_2\sqrt{1 + t_1}\cos(\pi t_2)\sin\left(\frac{\pi(t_1 - t_2)}{2} + 1\right)e^{.5(t_1+t_2)^2}$$

Selecting $M = 10$ as before, and choosing test points on the set $\mathcal{T} = T \cap \big(\{(t, t \pm .1) : t \in [-1, 1]\}\big)$, we again compute the maximum predictive error as a function of the number of basis functions for each method. Figure **??** suggests that all of the information from the diagonal is extracted very quickly using the RR method, whereas the convergence is much slower using the standard interpolation method to approximate the RKHS norm. This is likely due to the fact that a parameterization of the diagonal is differentiable whereas a parameterization of the boundary is not.

The results from these two examples indicate that our adopted approach in computing RKHS inner products has proven effective for incorporating information from more general subsets of the domain into a predictive Gaussian process model. Additionally, the method appears to be even more valuable in the case where the information available does not exist in the RKHS generated by the covariance kernel,



which is certainly the case for the parameterized boundary in the first example, and likely the case in the complicated example given in the second example.

1.7 Conclusions and Future Directions

The goal of this paper was to construct Gaussian processes which are capable of using information from arbitrary connected subsets of the domain in a way which required minimal assumptions to be made. Using the theory of Reproducing Kernel Hilbert Spaces, we were able to explicitly define the conditional mean and covariance of Gaussian processes via orthogonal projections in an RKHS, prove that such processes exist, and show that the processes are optimal in the sense of minimizing pointwise mean square error given the initial assumptions made. In addition, we provided several numerical examples to exhibit the practical nature of our construction, which included evidence that one need not assume the functional information available is an element of a RKHS. Future work in this area includes extending the theory to more naturally handle the case where functional information is available on disjoint subsets of the domain. Another interesting avenue to extend this research is to provide a similar framework for including more general linear operator constraints,

e.g. differential operator constraints.

Chapter 2

A Feature Space Characterization of Gaussian processes with Functional Information

2.1 Introduction

As noted in the previous chapter, the method of computing the conditional distribution where T_0 does not possess a differentiable parameterization may not lead to as accurate results. In addition, the case where T_0 is a disconnected set written by

$$T_0 = \bigsqcup_{\ell=1}^{L} T_0^\ell,$$

where T_0^ℓ are disjoint despite being valid under our construction, is not considered directly, and may pose a problem under our method of computing conditional distributions from a practical standpoint.

Defining $\mathcal{H}_\ell(T_0) = \overline{\text{Span}}\{k_s|_{T_0};\, s \in T_0^\ell\}$, it is not the case that $\{\mathcal{H}_\ell(T_0)\}$ is an

orthogonal set of subspaces. Therefore, defining the kernel of $\mathcal{H}(T_0)$ from $\{\mathcal{H}_\ell(T_0)\}$ is not at all straightforward. Breaking this down into orthogonal subsets would be very complex, requiring iterated projections, and would not be by any means practical even in the case where $L = 2$. However, suppose one can find orthogonal Hilbert spaces $\{H_\ell\}$ which identify with each RKHS. Then, one may potentially perform calculations on this less complex collection of Hilbert spaces and use them to describe elements of $\{\mathcal{H}_\ell(T_0)\}$. For example, consider the subspaces of $L^2(T_0)$ denoted $L^2_\ell(T_0)$ which contains elements only having support on T_0^ℓ. These subspaces are orthogonal, which makes reconstructing elements in $L^2(T_0)$ straightforward. Therefore, if one can define an isometric mapping between $L^2_\ell(T_0)$ and $\mathcal{H}_\ell(T_0)$, the simplicity of the L^2 spaces may be exploited for easier computations. Thus, we will now explore the connection between RKHS', Gaussian processes, and arbitrary Hilbert spaces.

2.2 Derivation via Isonormal Gaussian Processes

Let $T \subset \mathbb{R}^d$, and k a kernel function on T. In addition, let H be a separable Hilbert space such that one can find a mapping $\gamma : T \to \mathcal{B}(\mathbb{R}, H)$ such that

$$\langle \gamma_s^*, \gamma_t^* \rangle_H = k(s,t),\ \forall s,t \in T,$$

where γ_s^* refers to the Riesz representation of the adjoint of γ_s. The adjoint operator in this context will refer to the Hilbert adjoint unless otherwise specified. Let $\mathcal{H}(T)$ denote the RKHS generated by k on T, and $\mathcal{H}_0 \subset \mathcal{H}(T)$ a closed subspace. Define the linear operator $A_T : H \to \mathcal{H}(T)$ by

$$A_T(h)(t) = \langle \gamma_t^*, h \rangle_H.$$

One may note that A_T is well-defined since A_T can be extended from the inverse image of $\mathcal{H}(T)$ in A_T by defining $A_T h = 0$ for any $h \in \left(A_T^{-1}\mathcal{H}(T)\right)^{\perp}$. Then, A_T is a well-defined partial isometry, and $A_T^* A_T$ is the orthogonal projection onto $(\ker A_T)^{\perp}$ [Carmeli et al., 2006, Proposition 2.4]. Thus, it follows that $\langle h_1, h_2\rangle_H = \langle A_T(h_1), A_T(h_2)\rangle_{\mathcal{H}(T)}$, which implies that $A_T(\gamma_t^*) = k(t,\cdot)$, which we will denote k_t. In particular, $h \in \mathcal{N}(A_T)$ iff $\langle \gamma_t^*, h\rangle_H = 0$, $\forall t \in T$, where $\mathcal{N}(\cdot)$ denotes the nullspace.

Without loss of generality, let us assume that $\mathcal{N}(A_T) = \{0\}$, or in other words that A_T is an isometric isometry. Now, define P_0 to be the orthogonal projection from $\mathcal{H}(T)$ onto $\mathcal{H}_0$, and define $A_0 = P_0 A_T : H \to \mathcal{H}(T)$.

Theorem 2.2.1. *A_0 is a partial isometry from H to $\mathcal{H}_0$, with feature map $\phi : T \to \mathcal{H}_0$ which satisfies*

$$\phi_t^* = A_0^* A_0 \gamma_t^*, \ \forall t \in T.$$

Moreover, $^{\exists} H_0 \leq H$, a projection operator $Q_0 : H \to H_0$, and a isometric isometry $U_0 : H_0 \to \mathcal{H}_0$ such that the diagram below commutes.

$$\begin{array}{ccc} H & \xrightarrow{A_T} & \mathcal{H}(T) \\ \downarrow{\scriptstyle Q_0} & \overset{A_0}{\searrow} & \downarrow{\scriptstyle P_0} \\ H_0 & \xrightarrow{U_0} & \mathcal{H}_0 \end{array}$$

Proof. Define $\{\phi_t\} \subset H$ so that for each $t \in T$,

$$\phi_t^* = A_T^* P_0 A_T \gamma_t^* = A_T^* P_0 k_t = A_0^* k_t.$$

Therefore, for $h \in H$, $t \in T$,

$$\begin{aligned}(A_0 h)(t) = (P_0 A_T)h(t) = \langle (P_0 A_T)(h), k_t\rangle_{\mathcal{H}(T)} = \langle A_T(h), P_0 k_t\rangle_{\mathcal{H}(T)} = \langle h, (A_T^* P_0)k_t\rangle_H \\ = \langle h, (A_T^* P_0 A_T)\gamma_t^*\rangle_H = \langle h, \phi_t^*\rangle_H.\end{aligned}$$

Therefore, A_0 is a partial isometry onto the RKHS with reproducing kernel given by

$$k_0(s,t) = \langle \phi_s^*, \phi_t^* \rangle_H = \langle P_0 k_s, P_0 k_t \rangle_{\mathcal{H}(T)}.$$

By Theorem 2.5 from Paulsen and Raghupathi [2016], k_0 is the reproducing kernel for $\mathcal{H}_0$. It follows from Proposition 2.4 [Carmeli et al., 2006, Proposition 2.4] that $Q_0 = A_0^* A_0$ which is an orthogonal projection. Furthermore, by defining $U_0 = A_0 Q_0^*$, we see that U_0 is a partial isometry, since A_0 is an isometry on H_0, and is injective since Q_0 is self adjoint. This completes the proof. □

Thus, one may note that any subspace of $\mathcal{H}(T)$, has an associated feature space which is obtained via an equivalent orthogonal projection on H. Equivalently, one may define a projection on the feature space and obtain an equivalent subspace of $\mathcal{H}(T)$. Since no assumption was made on the form of P_0, the same holds for $A_T - A_0$ and $\mathcal{H}_0^\perp$. Now, define $\mathbb{W} = \{W(h);\, h \in H\}$ to be a zero mean iso-normal Gaussian process so that

$$E[W(h_1)W(h_2)] = \langle h_1, h_2 \rangle_H.$$

Then, the process $\mathbb{X} = \{X_t;\, t \in T\}$ defined by $X_t = W(\gamma_t^*)$ is a zero mean Gaussian process with zero mean and covariance kernel k, and define $\mathcal{F}_0$ by

$$\mathcal{F}_0 = \sigma\big(\{W(h);\, h \in (\ker A_0)^\perp\}\big) = \sigma\big(\{W(h);\, h \in H_0\}\big).$$

This definition makes sense as a σ-algebra, but it may not necessarily be clear how to describe the elements of $\mathcal{F}_0$. However, we may note that for almost all $\omega \in \Omega$, one can consider $W(\cdot)(\omega)$ a bounded linear functional on H, and therefore is equivalent to some function $h_\omega \in H$ by the Riesz representation theorem. This indeed is the

case provided that

$$\langle h_\omega, h\rangle_H = W(h)(\omega),\ \forall h \in H.$$

Therefore, for $A \in \mathcal{B}(\mathbb{R})$, one has $\{W(h) \in A\} = \{\omega \in \Omega : \langle h_\omega, h\rangle_H \in A\}$.

The conditional expectation $E[W(h)|\mathcal{F}_0]$ is actually quite simple to compute. Given that $W(h) \perp \mathcal{F}_0$ for $h \in \mathcal{N}(A_0)$, and the fact that $W(h)$ has zero mean, one may compute

$$E[W(h)|\mathcal{F}_0] = E[W(h_0)|\mathcal{F}_0] + E[W(h_\perp)|\mathcal{F}_0] = W(h_0) + E[W(h_\perp)] = W(h_0),$$

where $h = h_0 + h_\perp$, with $h_0 \in H_0$ and $h_\perp \in H_0^\perp$. This result is not surprising given the connection between orthogonality in H and independence in $\mathbb{W}$.

Now, let $g \in \mathcal{H}_0$, and let $h_g \in H_0$ satisfy $A_0(h_g) = g$, and $||h_g||_H = ||g||_{\mathcal{H}(T)}$. Denote

$$F_g = \{W(h) = \langle h_g, h\rangle_H;\ h \in (\ker A_0)^\perp\} \in \mathcal{F}_0,$$

which is measurable due to the fact that H is separable.

Theorem 2.2.2. *The distribution of $W(h)|F_g$ is Gaussian with mean and covariance*

$$\mu_g^W(h) = \langle Q_0 h, h_g\rangle_H,$$
$$k_g^W(h_1, h_2) = \langle (I - Q_0)h_1, (I - Q_0)h_2\rangle_H = \langle (I - Q_0)h_1, h_2\rangle_H$$

In particular, for $s, t \in T$, we have

$$\mu_g^W(\gamma_t^*) = \langle k_t, g\rangle_{\mathcal{H}_0},$$
$$k_g^W(\gamma_s^*, \gamma_t^*) = k(s,t) - \langle k_s, k_t\rangle_{\mathcal{H}_0}.$$

Proof. First, note that for any $h_0 \in H_0$, $W(h_0)(\omega) = \langle h_0, h_g \rangle_H$ for any $\omega \in F_g$. As before, decompose $h \in H$ as $h = h_0 + h_\perp$, where $h_0 = Q_0 h \in H_0$, $h_\perp = (I - Q_0)h \in H_0^\perp$. Now, observe

$$\begin{aligned} E[\exp\{i\alpha W(h)\}|F_g] &= E[\exp\{i\alpha(W(h_0) + W(h_\perp)\}|F_g] \\ &= E[\exp\{i\alpha(\langle h_0, h_g \rangle_H + W(h_\perp))\}|F_g] \\ &= \exp\{i\alpha\langle h_0, h_g \rangle_H\} E[\exp\{i\alpha W(h_\perp)\}] \\ &= \exp\{i\alpha\langle h_0, h_g \rangle_H - \alpha^2/2\langle h_\perp, h_\perp \rangle_H\}. \end{aligned}$$

Thus, $W(h)|F_g$ is Gaussian with mean $\langle h_0, h_g \rangle_H$, and variance $\langle h_\perp, h_\perp \rangle_H$. The multivariate case is tedious, but follows the same steps as the univariate case, so it is omitted.

□

Thus, we are able to recover the process $\tilde{\mathbb{X}} = \mathbb{X}|F_g$ by setting $\mu_g : T \to \mathbb{R}$ by $t \mapsto \mu_g^W(\gamma_t^*)$ and $k_g : T \times T \to \mathbb{R}$ by $(s,t) \mapsto k_g^W(\gamma_s^*, \gamma_t^*)$.

2.3 Approximating Processes

In the earlier sections, we set $\mathcal{H}_0 = \overline{\text{Span}}(\{k_s;\, s \in T_0\})$, showed that $\mathcal{H}_0 = \overline{\text{Span}}(\{k_s;\, s \in D\})$, where $D = \{t_n\}$ is a dense subset of T_0, and proved that the finitely conditional processes from $\{k_s;\, s \in T_0\}$ converge weakly to the process $\mathbb{X}|\{X_t = g(t) : t \in T_0\}$ under non-constrictive regularity assumptions. We can rephrase this solution in terms of what was done in the previous section by noting that $\{\gamma_t^* : t \in D\}$ is a spanning set for $(\ker A_0)^\perp$. This follows from the fact that every element in $\{\gamma_t^* : t \in D\}$ identifies uniquely through A_0 with an element in $\{k_t : t \in D\}$, and the fact that A_0 is a partial isometry.

Let $\{h_n\} \subset H$ satisfy $\overline{\text{Span}}(\{h_n\}) = (\ker A_0)^\perp$ from Theorem 2.7.1. Then, $\overline{\text{Span}}(\{A_0(h_n)\}) = \mathcal{H}_0$. Now, define $P_N : \mathcal{H}_0 \to \mathcal{H}_0$ to be the orthogonal projection onto $\text{Span}(\{A_0(h_n)\}_{n=1}^N)$. Then, clearly $P^N \to I$ strongly. Now define H_{NN} to be the matrix with ijth element $\langle h_i, h_j \rangle_H$, and $H_N(h) = (\langle h_n, h \rangle_H)_{n=1}^N$, for $h \in H$.

For $f, g \in \mathcal{H}_0$, let h_f, h_g be the inverse images of f, g respectively in $(\ker A_0)^\perp$. Then,

$$\langle P_N f, P_N g \rangle_{\mathcal{H}_0} = \langle P_N f, g \rangle_{\mathcal{H}_0} = H_N(h_f)' H_{NN}^{-1} H_N(h_g).$$

This can also be written in terms of the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$, but the notation used above is more practically convenient if one selects H to be a Hilbert space with a more tractable inner product. Using the same reasoning as in Section 2.3 and 2.4, it is easy to see the same limit results shown originally hold in this more general case, as well as the weak convergence and conditional expectation results.

This grants us the ability to approximate the RKHS inner product from a more variational perspective without having to concern ourselves with approximating the RKHS inner product directly. However, this does not exclude the notion of conditioning on particular points. To see this, simply select $h_n = \gamma_{t_n}^*$ for $\{t_n\} \subset \mathcal{H}_0$. Then, $(H_{NN})_{ij} = \langle \gamma_{t_i}^* \gamma_{t_j}^* \rangle_H = k(t_i, t_j)$, and we get a more typical approach to Gaussian process regression.

2.4 A Natural Variational Approach

Suppose one has a collection of elements $\{f_n\} \overset{dense}{\subset} L^2(T_0)$. Assuming $\mathbb{X}$ has continuous sample paths, one may note that $X_t = g(t)$ for $t \in T_0$ iff $\int_{T_0} X_t f_n(t) dt = \int_{T_0} g(t) f_n(t) dt$, for each n. Thus, by defining $Z_n = \int_{T_0} X_t f_n(t) dt$, one may derive the conditional distribution of $X_t | \{Z_n\}_{n=1}^N$ directly for any N. This approach allows a

lot of flexibility and convenience from a practical standpoint, since it removes the necessity of computing the solution of an integral equation. Our goal is to show that this formulation fits into the framework given above.

Define $L^2(\mathbb{P}) = L^2(\Omega, \mathcal{A}, \mathbb{P})$ to be a probability space. Let $\mathbb{W}$ be defined as above, $H = L^2(T_0)$ and define $\mathbb{I} = \{I(h);\, h \in H\}$ for $h \in H$ by

$$I(h) = \int_{T_0} h(t) X_t dt.$$

One may treat $\mathbb{W}$ and $\mathbb{I}$ not only as stochastic processes, but as mappings from H to $L^2(\mathbb{P})$. Thus, we may discuss the connection between the two mappings.

Proposition 2.4.1. *Let K be the integral operator on T_0 associated with k. Then, $\mathbb{I}$ and $\mathbb{W}K^{1/2}$ are indistinguishable.*

Proof. By the separability of H, it suffices to show that for $h \in H$,

$$E\Big[\big(I(h) - W(K^{1/2}h)\big)^2\Big] = 0.$$

First, observe that one can write $K^{1/2}h(t) = \langle \gamma_t^*, h\rangle_{T_0}$, for $h \in H$. Now, note that

$$E[I(h)^2] = \langle Kh, h\rangle_{T_0} = E[W(K^{1/2}h)^2].$$

Thus, we have

$$\begin{aligned} E\big[I(h)W(K^{1/2}h)\big] &= E\bigg[\int_{T_0} h(t) W(\upsilon_t^*) W(K^{1/2}h) dt\bigg] = \int_{T_0} h(t) \langle \upsilon_t^*, K^{1/2}h\rangle_{T_0} dt \\ &= \int_{T_0} h(t) Kh(t) dt = \langle Kh, h\rangle_{T_0}, \end{aligned}$$

which completes the proof. □

The proof is elementary, but if one assumes that k is universal, one may define conditions in terms of $\mathbb{I}$ rather than $\mathbb{W}$, which allows one to compute inner products using k rather than the feature maps on $L^2(T_0)$, thus alleviating the need to perform a spectral decomposition of K on T_0. Considering the spanning set $\{f_n\}$, to use the approach in 6.2, we need to set $\{h_n\}$ to be so that $h_n = K^{1/2} f_n$, which also forms a spanning set of H since K is injective (due to k being universal). This may seem problematic as employing $K^{1/2}$ would likely require a spectral decomposition, but when computing inner products,

$$\langle K^{1/2} f, K^{1/2} g \rangle_{T_0} = \langle Kf, g \rangle_{T_0}$$

by the self adjointness of K, so this detail does not pose an issue. Of course, one drawback is that if one chooses an orthogonal set as a starting point, it is likely that H_{NN} will be a fully dense matrix. Ideally, one would select $\{f_n\}$ to be a linearly independent set so that H_{NN} has full rank.

2.5 A Proof of Concept

Returning to the original problem introduced at the beginning of the section, we can see that rather than dealing with ∂T as a whole, we may deal with the individual sides and merely consider interactions between basis functions on differing sides, and note that since

$$L^2(T_0) \cong \bigotimes_{\ell=1}^{4} L^2(T_0^\ell),$$

a basis for $L^2(T_0)$ can be constructed from bases for each $L^2(T_0^\ell)$. Moreover, since the topology of each is the same, a reparameterization of the same basis functions can be used for each side. An additional important note is that under a Lebesgue measure

space on T, T_0 has measure zero, so an orthogonal projection from $L^2(T)$ to elements whose support is concentrated on T_0 is equivalent to the zero operator. Thus, a straightforward approach as done in 6.2 is not valid. However, in this scenario, a solution is to start with two separate isonormal processes $\mathbb{W}_T$, and $\mathbb{W}_{T_0}$ which are defined on $L^2(T)$ and $L^2(T_0)$ respectively. Define $\gamma_t^* \in L^2(T)$ for $t \in T$, and $\upsilon_s^* \in L^2(T_0)$ for $s \in T_0$ as in 6.2 so that for each $s \in T_0$,

$$W_T(\gamma_s^*) = X_t = W_{T_0}(\upsilon_s^*).$$

Hence we are able to connect $L^2(T)$ and $L^2(T_0)$ in some sense by mapping to $L^2(\mathbb{P})$. Therefore, we see that although $\mathbb{W}_T$ is the stochastic process of interest, we are able to do most of the work on $\mathbb{W}_{T_0}$ which is more convenient.

Note that the construction

2.6 An Application

Consider the problem given in Section 6.4. In this section, the goal is to show that this can be implemented numerically using the approach in Section 6.3 to approximate a function f defined on T whose information on ∂T is known. So, let $T = [-1, 1]^2$ and define f by

$$f(t) = t_2\sqrt{1 + t_1}\cos\left(\frac{\pi t_2}{2}\right)\sin\big(\pi(t_1 - t_2)\big).$$

We assume that the value of f is known at M interior points of the domain, where $N \geq 0$. The choice of basis functions is likely to be dependent upon the smoothness of f. If f is smooth, using a polynomial or trigonometric basis may be preferable, whereas if f has a lot of variation, a wavelet basis may be preferable. In any case,

denote $\{h_n^1\}$ to be a basis of $L^2(T_0^1)$, where $T_0^1 = \{(x,-1) : x \in [-1,1]\}$. Let T_0^ℓ denote the other sides of T_0, and $\{h_n^\ell\}$ the same basis as $\{h_n^1\}$, but altered to be defined on T_0^ℓ. Note that h_n^ℓ can easily be extended to an L^2 function on T_0 by allowing it to have support only on h_n^ℓ. Then, using the approach from 6.3, and the notation from 6.2, one may approximate the covariance function k_0 as

$$k_0^N(s,t) = k(s,t) - H_N'(k_s)H_{NN}^{-1}H_N(k_t),$$

where

$$H_N(k_s)_{\ell N+n} = \int_{T_0^\ell} h_n^\ell(t)k(s,t)dt,$$

and $(H_{NN})_{\ell N+n,qN+j} = \int_{T_0^\ell}\int_{T_0^q} h_n^\ell(s)h_j^q(t)k(s,t)dtds$. Thus, given M points on the interior of the domain, one may use the updated mean function and updated covariance function as a prior distribution to perform Gaussian process regression.

Factors which influence the accuracy of the approximation to f are the type of kernel, the type of basis functions, and the number of basis functions used. In this example, the kernel $k(s,t) = \sigma^2 \exp\{-(s-t)'\Lambda(s-t)\}$ is chosen where Λ is a positive diagonal matrix. Several different bases are selected in order to compare efficiency.

It is assumed that each side of T_0 will have N basis functions, yielding a total of $4N$ basis functions. The bases which will be compared are Legendre and Chebyshev polynomials, trigonometric bases, and the feature maps (which is equivalent to simply conditioning on various values). Since in many cases it is important to report actual function values, the supremum norm is approximated as a measure of goodness of fit for each method, for values of N varying from 5 to 15, corresponding to 8 to 60 total basis functions describing the behavior of f on T_{0j}.

For the case of choosing the feature maps, the points are selected to be equidis-

tant for each value of N. Thus, for each N, there is a completely new set of points at which the feature map is evaluated, which is likely to be a sub-optimal procedure. The test points where selected to be close to the boundary in order to more directly observe the effect of the boundary information on the prediction accuracy. Figure 2.1 illustrates the superior efficiency of the two polynomial bases chosen over the feature map and trigonometric bases. However, it may be the case that strategic selection of feature maps would improve the performance of using a feature map basis.

One may observe that after a certain number of basis functions, the error plateaus, which suggests that no more information can be extracted from the boundary information given.

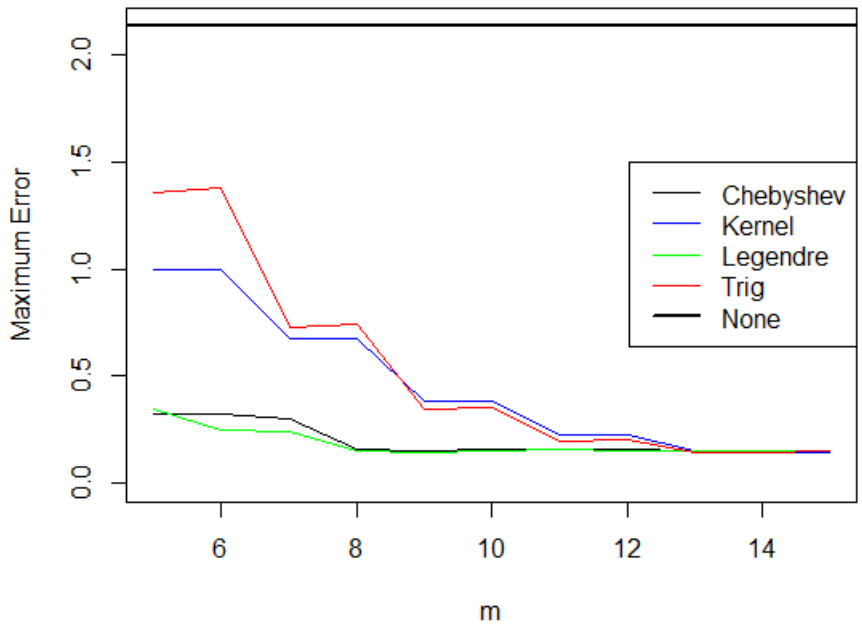


Figure 2.1: Comparison of the efficiency of various basis functions based upon the approximate supremum norm.

Part II

State-Aware Calibration

Chapter 3

Identification of State Aware Parameters in Computer Model Calibration

3.1 Introduction

In recent history, technology and computing power has played a significant role in understanding complex phenomena in which information is limited and observed data is the primary source of understanding. Complex scientific and engineering models are often insufficient in describing reality, or carry with them a high computational expense, and physical experiments can be expensive or potentially hazardous. Computer models have proven to be an excellent surrogate to correct observed bias induced by an incomplete model, or reduce the computational burden associated with a high fidelity deterministic model or physical experiment while minimizing information loss. In addition to providing a means of prediction of unknown behavior, computer models can be used for other tasks, such as design of experiments, or estimating unknown

qualities of a given system.

In the context of computer models, estimation of such unknown qualities of a system is referred to as calibration. Calibration parameters are components of a computer model which are generally considered unobservable and cannot be controlled as inputs. They play a role in tuning computer models for predictive purposes, but are often useful in gaining a further understanding of a physical system. Bayesian methods of calibrating parameters assume a prior probability distribution selected via background information, and typically assume parameters to be a collection of scalars which are then estimated via computer output and physical data. The posterior distribution of calibration parameters can then be utilized for determining credible regions in which a parameter lies, or determining point estimates.

However, treating calibration parameters as scalars independent of control inputs can often be insufficient and yield misleading results, which has been observed in models of plastic deformation [Atamturktur et al., 2015, Chodora et al., 2020], the study of ion channels for cardiac cells [Plumlee et al., 2015], buckypaper manufacturing [Pourhabib et al., 2015], and resistance spot welding, [Ezzat et al., 2018].

These works have indicated that allowing calibration parameters to vary as functions of control inputs naturally can improve the predictive power of computer models, and yield a more complete understanding of a system. However, in making this generalization, calibration parameters exist in a function space, which makes their estimation more difficult, particularly in non-linear models. The methods of Plumlee et al. [2015], Brown and Atamturktur [2016] suggest non-parametric approach in which parameters are defined *a priori* as Gaussian processes (GP). GP's are a popular choice for computer models because they do not assume any form of the function except for mean square continuity [Santner et al.], and have proven effective for smaller sample sizes, which is a common feature of experiments neces-

sitating computer models. However, incorrectly assuming a parameter is functional introduces greater computational costs as well as potential sources of confounding between variables, so naively treating all calibration parameters as state-aware is not the best approach. Of course, state-aware and constant parameters can be chosen prior to calibration, which was assumed in Brown and Atamturktur [2016]. These determinations can often be elicited via expert opinion on the subject matter, but in general such properties of calibration parameters are unknown. This paper will make contributions to the work in functional calibration by outlining a methodology for identifying state-aware calibration parameters.

The proposed model is similar in nature to variable selection, which is a well-researched topic. For instance, Spike-and-Slab regression is a well known Bayesian variable selection technique in which marginal posterior distributions of parameters have mass at zero with positive probability [Linkletter et al., 2006, Savitsky et al., 2011]. Alternatively, Stochastic Search Variable Selection (SSVS) [George and McCulloch, 1993] defines parameters as a weighted average of a Gaussian distribution tightly bound to zero and a more diffuse Gaussian distribution, and do not allow a point mass at zero. These methods are used as inspiration for exploring methods of identifying state-aware calibration parameters in this paper.

More recently, Barbillon et al. [2021] developed a framework which determines control settings contributing to model bias via an assessment of the correlation parameters which utilized the SSVS method. However, this is not the first attempt which has been made in mixing Gaussian processes with other models. Gramacy and Lee [2008] developed a model which determines whether a stochastic process follows a Gaussian process or a classical linear model with uncorrelated error through means similar to that of [George and McCulloch, 1993]. Combining this model with the Bayesian treed model [Chipman et al., 2002], Gramacy and Lee [2008] developed a

model which allowed output to take different levels of complexity in different locations of the domain. This was achieved via the correlation function, which in a classical linear model is simply white noise. Although the proposed model is similar in concept to that of Gramacy and Lee [2008], the model was designed as a surrogate, and not for calibration parameters. In addition, a constant Gaussian process with nonzero variance has the correlation function equivalent to 1, which is significantly different than a white noise error model over a continuous domain.

One concern involved with Gaussian processes with correlation function $\rho \equiv 1$ is the notion that any covariance matrix generated by ρ has rank one, and therefore is not invertible. Thus, a density does not exist, making likelihood based MCMC methods infeasible when approached directly. Section 2 will discuss strategies to overcome this obstacle so that likelihood based MCMC methods can be employed to produce samples of posterior distributions.

Additionally included in Section 2 is a full outline of the model which address further the strategies and concerns related to the formulation, and Section 3 will discuss the strategies in calculating the posterior distributions of the parameters. Section 4 will provide a basic proof of concept of the model and analyze the effectiveness of the method when considering sample size, and sensitivity of the parameters to the data. Section 5 involves a case study using data from an application in materials science [Arp et al., Submitted]. Lastly, section 6 discusses conclusions and potential future directions of research.

3.2 Model

3.2.1 Background

The process of calibrating computer models involves using physical data in conjunction with computer model output to select unobservable quantities associated with a given system so that the computer model most accurately reflects the physical data. This notion dates back at least as early as Berman and Nagy [1983], but was not studied extensively by statisticians until the work of Kennedy and O'Hagan [2001a]. The computer model is defined by the function η, and represents either evaluations from a physical model, or in the case of higher fidelity models, a Gaussian process emulator Sacks et al. [1989a] based upon training data. The framework provided by Kennedy and O'Hagan defined the computer model by η, taking in parameters $(\mathbf{x}, \mathbf{t})$, where $\mathbf{x}$ and $\mathbf{t}$ represent control parameter and calibration parameter settings respectively. The true response $y(\cdot)$, representing the physically measured data can then be written as

$$y(x) = \eta(\mathbf{x}, \theta) + \delta(\mathbf{x}) + \varepsilon(\mathbf{x}),$$

where δ represents the bias associated with the computer model with the true (not measured) physical model, and ε represents random noise associated with imperfectly measured data. This formulation is then compiled under a Bayesian framework, where y, δ are defined to be Gaussian processes, is typically defined to be a white noise Gaussian process, and θ is a vector of scalars whose prior distribution is elicited by subject matter expertise. Calibration is then carried out by maximizing the posterior distribution of θ while estimating δ as well.

A particular extension of interest of this framework involve calibration via the entire posterior distribution as opposed to the marginal distribution of θ and δ

[Higdon et al., 2004], which is how calibration will be performed in the methodology given in this paper.

The treatment of functional parameters in this paper is derived from the work of Plumlee et al. [2015], Brown and Atamturktur [2016]. Using the notation in the Kennedy O'Hagan framework, the model for the response will be given by

$$y(\mathbf{x}) = \eta(\mathbf{x}, \theta(\mathbf{x})) + \varepsilon(\mathbf{x}),$$

where $\theta(\mathbf{x})$ is a vector of Gaussian processes. In this particular application, concerns of poor identifiability has driven the decision to remove the model bias term. The model bias term in the case of constant calibration parameters results non-unique estimates from the data [Kennedy and O'Hagan, 2001a], and this concern needless to say would be exacerbated in the case of functional calibration parameters. This decision from a modeling perspective can be explained when considering the fact that the increased flexibility can potentially drive the model bias to zero.

3.2.2 Framework

To keep the framework relatively simple, we consider a model in which univariate data $\boldsymbol{y} = (y_1, ..., y_n)^T$ is available with control inputs $\boldsymbol{x} = (x_1, ..., x_n)^T$ and calibration parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$. As typically upper and lower bounds are set for each parameter based upon prior knowledge, we will assume without loss of generality that $\boldsymbol{x} \in [0, 1]^n$, $\boldsymbol{\theta} \in [0, 1]^p$. The goal is to design a statistical model under a Bayesian framework which determines whether the calibration parameters are functional. The formulation of the model is given below:

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\theta}, \lambda_y &\sim \mathrm{N}(\eta(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x})), \lambda_y^{-1}\boldsymbol{I}), \\
\lambda_y &\sim \mathrm{G}(\epsilon, \epsilon), \\
\theta_i(\cdot)|\rho_i, \lambda_i &\overset{indep}{\sim} \mathcal{GP}(0, \lambda_i^{-1} R_i(\cdot, \cdot)), i = 1, ..., p \\
\lambda_i &\sim \mathrm{G}(\omega, \omega),
\end{aligned}
$$

and we define R_i as

$$
(R_i(x, x'))_{j,k} = \rho_i^{|x_j - x_k|^r}, r \in [1, 2], \quad i = 1, ..., p,
$$

where r is fixed. Although it is commonly assumed for $r = 2$ due the nice L^2 properties it imposes on the Gaussian process such as infinite differentiability and is preferable for calibration due to its smoothness, $r = 1$ is selected for the particular purpose of this problem, as the end goal is not to estimate the process itself. In fact, selecting $r = 1$ will express the effect on different values of ρ_i more significantly since by our assumption, $|x_i - x_j| < 1$ which implies that the covariance matrix is more sensitive to small changes in ρ_i when ρ_i is sufficiently far away from zero. After determining the state of the parameters, one may select $r = 2$ to perform standard calibration as in Brown and Atamturktur [2016].

Note that as ρ_i approaches 1, the matrix $R_i(x, x)$ can become very ill-conditioned (particularly when $r = 2$), so it may be necessary to add a "nugget" to the matrix R_i, i.e. construct an approximation $\tilde{R}_i(x, x) = \epsilon I + R_i(x, x)$. This sets an upper bound for the condition number of the covariance matrix, ensuring that computations can be accurately performed [Ranjan et al., 2011].

η is assumed to be a continuous function from which calculations are inexpen-

sive. This can be achieved via either the form of η or an emulator. In many computer model settings, the function η is typically very accurate, so the prior distribution of λ_y is therefore selected to have high precision, which is important since θ needs to be estimated well to determine its state. Potential confounding effects in the design are the interplay between the hyperparameters ρ_i and λ_i. Since the primary interest is in the posterior distributions of ρ_i, it is advantageous to select the prior distribution for λ_i in a way that minimizes its effect on the model so that ρ_i can be the driving source of variation. Another possibility of mitigating the effect of λ_i is to marginalize the process θ_i over λ_i so that ρ_i is the only parameter affecting θ_i. The resulting distribution in this case is derived in Section 3.

The hierarchical structure of the model makes Markov Chain Monte Carlo (MCMC) a leading candidate as a method of drawing a sample from the posterior distributions. It is assumed that each calibration parameter is uniquely identifiable, which may be enforced by including strong prior information typically available via expert opinion. This suggests that the number of calibration parameters p considered should be reasonably small. If this is not the case, variable selection may need to be performed *a priori* to ensure all calibration parameters are needed in the model [see Savitsky et al., 2011].

It is also assumed that for a given sample size N, the function evaluation locations $\{x_i\}_{n=1}^N$ are dispersed in a way that a function of interest f is not misrepresented as constant. To illustrate the importance of this assumption, imagine that one is testing whether $f \equiv \sin(\cdot)$ is functional via a collection of points, but uses evaluations at the points $\{n\pi\}_{n=1}^N$. One would likely then incorrectly determine that $f \equiv 0$. Naturally, this assumption is not directly verifiable, but assuming that N is sufficiently large and the evaluation points are distributed non-sequentially will likely suffice for practical purposes. The effect of sample size on performance for several

test models will be demonstrated in Section 4.

Before describing the prior distribution for ρ_i, we will first explain our rationale for choosing ρ_i to be the driving force behind idenitfying a parameter as state aware. Let $y(\cdot)$ on $\mathcal{X}$ be a Gaussian process with constant mean and nonzero variance. Then, $y(\cdot)$ is constant *iff* $\text{Corr}(y(x_1), y(x_2)) = 1$, for any $x_1, x_2 \in \mathcal{X}$. In particular, $y(\cdot) \equiv Y$ almost surely for a scalar, Gaussian random variable Y. Defining k to be the covariance function of $y(\cdot)$, this is also equivalent to saying $k \propto 1$. Letting k be an element of the parameterized family of covariance functions defined in our model formulation, $k(x, x') = \lambda^{-1}\rho_i^{|x-x'|^r} \propto 1$ *iff* $\rho_i = 1$. Thus, one may determine if $y(\cdot)$ is state-aware by checking this condition.

The driving component in correctly identifying which parameters are state aware is the distribution of ρ_i. The design is similar to approaches used for variable selection [see Linkletter et al., 2006, Savitsky et al., 2011, George and McCulloch, 1993], as one can think of θ_i as having two distinct modes similar to how a variable can either be included or excluded from the model. Such methods can be consolidated into the following general framework:

$$\begin{aligned} \rho_i|\gamma_i &\sim \gamma_i \text{Beta}(\alpha_0, \beta_0) + (1-\gamma_i)F, \;\; i = 1, ..., p \\ \gamma_i &\overset{indep}{\sim} \text{Bernoulli}(\pi_i), \end{aligned}$$

where F is a distribution function with support on $[0, 1]$ which is designed to represent the case where $\rho_i = 1$. α_0 and β_0 are selected to be to be reasonably small as to encourage exploration but large enough to discourage values around zero and one. This represents a prior distribution selection in the case where θ_i is a Gaussian process. We would like for $\gamma_i = 0$ to correspond to $\rho_i = 1$, so therefore selecting F as the Dirac measure at 1 is a natural selection. This would correspond to a spike-and-slab

approach. The other option is to select F so that its weight is heavily concentrated at one, e.g. $F \sim \text{Beta}(\alpha, \beta)$, where $\alpha \gg \beta$, which corresponds to an SSVS approach.

Each approach has potential pitfalls. The covariance operator has full rank if $\rho_i < 1$, regardless of how close it becomes, but if $\rho_i = 1$, the covariance operator has rank one. In the latter case, any finite dimensional distribution of the process does not have a density, which makes performing likelihood-based MCMC techniques infeasible directly. An SSVS approach would require careful considerations when selecting F, as approximating very smooth functions with GP's may require ρ_i to be close to 1. Thus, state-aware calibration parameters having very smooth relationships with the input may be incorrectly determined as constants.

An alternative means of performing MCMC in this scenario under the spike-slab paradigm is to employ approximate Bayesian computation (ABC) methods, which do not require existence or evaluation of a likelihood. However, ABC methods generally don't consider properties such as smoothness [Wilkinson, 2014], which is a major disadvantage in the context of this problem. Therefore, it is not considered in the construction of this model.

In order to apply a spike-and-slab approach, one must slightly alter the distribution in order so that the density exist for the case of $\rho_i = 1$. One possibility is to reduce the Gaussian distribution to a univariate Gaussian, but the log-density of the univariate Gaussian random variable for $\rho_i = 1$ will be significantly smaller than the log-density for nearly constant vectors with $\rho_i \approx 1$, due to the fact that the smoothness of the kernel is taken into account in the multivariate case with $-\log|R_i|$ being considered, which diverges as $\rho_i \uparrow 1$. Thus, likelihood based MCMC methods would not be appropriate in this scenario.

Another possibility is to add a nugget term as previously mentioned to the singular covariance matrix, i.e. $\tilde{R}_i = 11^T + \epsilon I$ when $\rho_i = 1$ so that the matrix is

invertible. The resulting log-likelihood as a function of ρ_i does not yield desirable results. To illustrate this point, denote $f(\mathbf{y}|\rho_i) \sim \mathcal{MVN}(0, \tilde{R}_i(\mathbf{x}, \mathbf{x}))$, and suppose $\delta = 10^{-3}$. Figure 3.1 plots $f(\mathbf{y}|\rho_i)$ as a function of ρ_i for $\mathbf{y} \equiv .3$. Thus, even when the observed data is constant, it is not reflected in the log-likelihood resulting from $\tilde{R}_i$. Thus, we will proceed employing an SSVS approach, setting $F \sim \text{Beta}(\alpha, \beta)$ as described above.

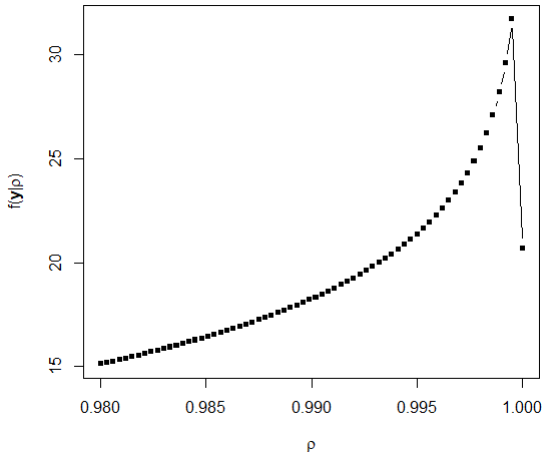


Figure 3.1: Plots $f(\mathbf{y}|\rho_i)$ as function of ρ_i.

3.3 Methodology

Before discussing the sampling from the posterior distribution of each component in our formulation, let us write down the sampling distribution

$$\pi(\theta(\mathbf{x}), \lambda, \rho_{\mathbf{i}}, \lambda_y|Y) \propto \pi(Y|\theta(\mathbf{x}), \lambda_y)\pi(\lambda_y) \prod_{i=1}^{2} \Big[\pi(\theta_i(\mathbf{x})|\rho_i, \lambda_i)\pi(\rho_i|\gamma_i)\pi(\gamma_i)\pi(\lambda_i)\Big],$$

where we assume $p = 2$ for the sake of simplicity. As the canonical decision metric under the SSVS paradigm is simply the posterior distribution of γ_i [George and McCulloch, 1993], the primary focus is effectively sampling from the posterior dis-

tribution of γ_i. Since $\gamma_i \sim \text{Bernoulli}(\pi)$, it's posterior distribution can be explicitly calculated. Indeed, we have

$$P(\gamma_i|Y, \theta(\mathbf{x}), ...) \propto p(\rho_i|\gamma_i, \theta_i(\mathbf{x}))p(\gamma_i),$$

and since the support is concentrated on two values we can simply calculate

$$P(\gamma_i = 1|Y, \theta(\mathbf{x}), ...) = \frac{p(\rho_i|\gamma_i = 1, \theta_i(\mathbf{x}))p(\gamma_i = 1)}{\sum_{j=0}^{1} p(\rho_i|\gamma_i = j, \theta_i(\mathbf{x}))p(\gamma_i = j)}.$$

As suggested in Savitsky et al. [2011], the joint posterior distribution of (γ_i, ρ_i) is considered, which is given by

$$p(\gamma_i, \rho_i|Y, ...) \propto p(\theta_i(\mathbf{x})|\rho_i, \lambda_i)p(\rho_i|\gamma_i)p(\rho_i).$$

From this calculation, it is clear that θ_i is a sufficient statistic for (γ_i, ρ_i). Therefore, it is important that θ_i is accurately estimated. It is also noteworthy that the prior distribution of $\rho_i|\gamma_i$ will play a significant role in the posterior distribution of γ_i. Figure 3.2 gives evidence of this, as the difference between the posterior distributions of $\gamma_i|\rho_i$ assumes completely different forms depending on different hyperparameters of the prior distribution of ρ_i. In this case, it is desired that the form is as given on the right, and so for this to occur, it is necessary that $\alpha_0 < \alpha$.

As inspiration from the "Add-Delete-Swap" method of Savitsky et al. [2011], the sampling strategy involved drawing (ρ_i, γ_i) jointly rather than separately. However, rather than selecting a uniform proposal for ρ_i in the case where $\gamma_i = 1$, the proposal distribution of ρ_i is posed as conditional on the proposed γ_i as follows:

$$\rho_i|\gamma_i \sim \gamma_i \text{Unif}(0, 1) + (1 - \gamma_i)\text{Unif}(a, 1), a \in (0, 1),$$

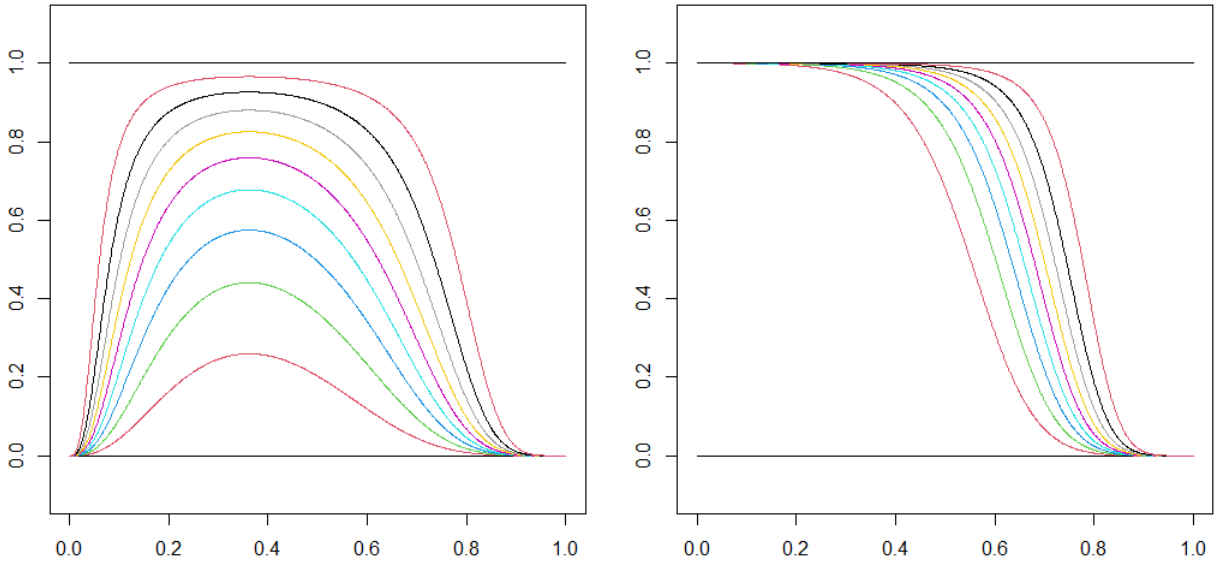


Figure 3.2: Above shows the two plots for the posterior distribution of γ_i, where each different curve corresponds to a different prior distribution on γ_i. The left plot the case where $\alpha_0 > \alpha$, whereas the right plot considers the case where $\alpha_0 < \alpha$. Clearly as ρ_i gets small, we would like for $\pi(\gamma_i = 1)$ to be close to 1. However, if $\alpha_0 < \alpha$, this is not the case. The plot on the right gives a much more desirable representation of an appropriate posterior distribution.

where a is selected to be close to 1. For the studies conducted in Section 4 and the application in Section 5, a is set to be 0.99. Note that the proposal distribution of γ_i depends upon the previous value of (γ_1, γ_2) [see Savitsky et al., 2011]. The probability of adding, deleting, or swapping is constructed to not favor a particular state or calibration parameter.

Given that ρ_i has compact support, a uniform mixture is capable of exploring the entire parameter space. Note that the marginal proposal distribution of ρ_i is discontinuous assuming that $\pi_i \notin \{0, .5, 1\}$ and $a \neq .5$, but this actually lends itself to the idea of an extension of a spike and slab proposal distribution, considering the Unif$(a, 1)$ converges to δ_1 as $a \to 1$. However, for $a \neq 1$, there is still no mass at one, so we bypass the issue of a singularity at one that would occur in a typical

spike-and-slab method.

As mentioned in the previous section, marginalizing λ_i out of the model will allow ρ_i to be the primary source of variation in the distribution of $\theta_i \equiv \theta_i(\mathbf{x})$, which will be shown to improve mixing of the posterior distribution of ρ_i. Under the assumption that $\lambda_i \sim \text{Gamma}(\omega, \omega)$, the marginalized prior $p(\theta_i|\rho_i)$ is computed as

$$\begin{aligned} p(\theta_i|\rho_i) &= \int_0^\infty p(\theta_i|\rho_i, \lambda)p(\lambda)d\lambda \propto \int_0^\infty \lambda^{\omega+n/2-1} \exp\left\{ -\lambda\Big[\omega + \frac{1}{2}\theta_i^T R_{\rho_i}^{-1}\theta_i\Big]\right\}d\lambda \\ &\propto \Big[\omega + \frac{1}{2}\theta_i^T R_{\rho_i}^{-1}\theta_i\Big]^{-(\omega+n/2)} \propto \Big[1 + \frac{1}{2\omega}\theta_i^T R_{\rho_i}^{-1}\theta_i\Big]^{-(\omega+n/2)}, \end{aligned}$$

which only depends on ρ_i. As one may expect, this distribution is in the multivariate t-distribution family, and thus $\theta_i|\rho_i$ as a stochastic process can be thought of as a Student's t-process [Shah et al., 2014]. However, since λ_i is no longer an element of the distribution, its prior parameters play a more significant role in the shape of the distribution, particularly for smaller sample sizes. Thus, priors must be carefully chosen. Each θ_i is sampled according to its respective full conditional distribution while marginalizing over λ_i (as shown above). This is achieved via a Metropolis Hastings algorithm, using a multivariate Gaussian random walk with a smooth covariance function. Additionally, the order in which θ_i is updated is randomized as to remove any potential dependency induced by the order in which the parameters are updated, which has been showed to not affect convergence of the Markov chain [Liu et al., 1995] (though it may affect the rate of convergence).

The white noise variance parameter is updated separately via a Gaussian random walk proposal distribution in a Metropolis Hastings scheme.

As mentioned before, the criterion which is used to determine whether a parameter is state aware is the posterior of distribution of γ_i. A simple way of estimating this parameter is by simply averaging all values of γ_i found via MCMC. However, not-

ing that

$$P(\gamma_i = 1|Y, \theta(\mathbf{x}), \rho_i) = \frac{p(\rho_i|\gamma_i = 1)p(\gamma_i = 1)}{\sum_{j=1}^{2} p(\rho_i|\gamma_i = j)p(\gamma_i = j)},$$

and recalling that (γ_i, ρ_i) are drawn jointly, one can approximate $P(\gamma_i = 1|Y, \theta(\mathbf{x}))$ with

$$\frac{1}{M}\sum_{m=1}^{M} P(\gamma_i = 1|\theta(\mathbf{x}), \rho_i^{(m)}),$$

where we use the fact that

$$P(\gamma_i = 1|Y, \boldsymbol{\theta}(\mathbf{x})) = E[\gamma_i|Y, \boldsymbol{\theta}(\mathbf{x})] = E\big[E[\gamma_i|Y, \boldsymbol{\theta}(\mathbf{x}), \rho_i]\big] = E_{\rho_i}\big[P(\gamma_i = 1|Y, \boldsymbol{\theta}(\mathbf{x}), \rho_i)\big]$$

Furthermore, given that the formulation does not permit $\rho_i = 1$ with positive probability, there is no true point at which $\gamma_i = 1$ with probability one *a posteriori*. Therefore, one may conclude that θ_i is constant provided that $P(\gamma_i = 1|Y, \theta(\mathbf{x}))$ is sufficiently small, and is state-aware otherwise. Since the criterion is dependent upon drawing good samples of ρ_i, it is very important that the samples $\{\rho_i^{(m)}\}$ have good mixing properties. In particular, as the series $\{P(\gamma_i = 1|Y, \boldsymbol{\theta}, \rho_i^{(m)})\}$ is of primary interest, the sample autocorrelation is used as a visual diagnostic for determining the strength of mixing. Figure 3.3 illustrates the improved mixing as a result of marginalizing over λ_i.

Although we assume throughout the paper that the control settings are one-dimensional, this may be easily extended to more than one dimension by simply modifying the correlation kernel R. As done in Savitsky et al. [2011], one may consider a product kernel

$$R_i(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^{d} \rho_{ij}^{|x_j - x_j'|^r}.$$

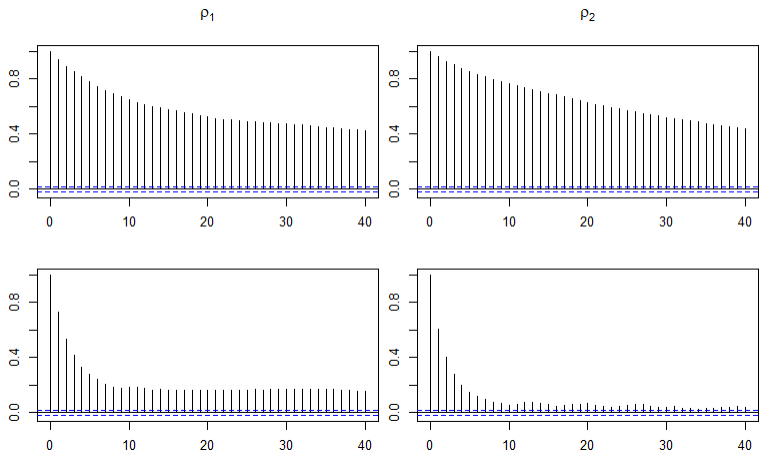


Figure 3.3: Autocorrelation functions for ρ_1, ρ_2 MCMC samples from the first example in Section 4. The top two figures show illustrate the poor mixing when not marginalizing out the variance parameter, whereas the bottom two figures show the improved mixing as a result of marginalization.

However, since R_i is constant only if $\rho_{i1} = \cdots = \rho_{id} = 1$, the individual values of each ρ_{ij} are not as relevant. In this sense, it may be more reasonable to consider an isotropic kernel of the form

$$R_i(\mathbf{x}, \mathbf{x}') = \rho_i^{||\mathbf{x}-\mathbf{x}'||_2^r}$$

where $||\cdot||_2$ denotes the Euclidean norm, since one is simply looking for dependence of any kind in the calibration parameter throughout the domain. Additionally, other distance functions might be useful in this context such as the maximum norm, but further exploration is necessary to make an evaluation of this proposition.

3.4 Simulation Studies

3.4.1 Model Validation

Before assessing the ability of the model to identify state aware parameters in a more practical setting with multiple calibration parameters possessing potential

interactions, the efficacy of the model is first illustrated in a simplified kriging setting, where η is the coordinate map associated with the calibration parameter $\theta_1(x)$. By validating the methodology in the simplest case, we may then attribute potential insufficiencies in posterior inference to more common issues in functional calibration and address those in a different setting.

For convenience, denote $R_{\rho i}$ to be a correlation matrix associated with ρ_i. In particular, for control inputs $\mathbf{x} = (x_i)_{i=1}^N$, select θ_1 so that

$$\theta_1(\mathbf{x}) = R_{\rho_{i0}}(\mathbf{x}, \mathbf{x})^{1/2}\mathbf{u},$$

where $\rho_{i0} \in (0, 1)$, and $\mathbf{u}$ is a realization from a $\mathcal{MVN}(0, I)$ distribution. Now, observe that

$$\begin{aligned} -\log \pi(\rho_i|\theta) &\propto \frac{1}{2}\log\left(|R_{\rho i}|\right) + (\omega + N/2)\log\left(1 + \frac{1}{2\omega}\theta_1(\mathbf{x})^* R_{\rho_i}^{-1}\theta_1(\mathbf{x})\right) \\ &\propto \frac{1}{2}\log\left(|R_{\rho i}|\right) + (\omega + N/2)\log\left(1 + \frac{1}{2\omega}\mathbf{u}^* R_{\rho_{i0}}^{1/2} R_{\rho_i}^{-1} R_{\rho_{i0}}^{1/2}\mathbf{u}\right), \end{aligned}$$

where we abuse notation by writing $R_{\rho i}$ in place of $R_{\rho i}(\mathbf{x}, \mathbf{x})$ for brevity. The work of Zaytsev et al. [2014] showed empirically that $-\log \pi(\rho_i|\theta)$ should be minimized near ρ_{i0}, which is not surprising given that $\theta_1(\mathbf{x})$ is select so that a posterior distribution generated by these points with the covariance $R_{\rho_{i0}}$ is in the RKHS generated by $R_{\rho_{i0}}$ for any N (as well as the limiting case). A simulation study was performed for $N = 5, 10, 15, 20, 25$, with $\theta_1(\mathbf{x})$ chosen as above for $\rho_i = 0.3$ for the state-aware case, and $\rho_i = 1$ for the state independent case, and credible intervals are given in Figure 3.4. Bias for the $\rho_i = 0.3$ case is likely due to the specification of the prior distribution, as well as the variability in θ_1, both of which dissipate as N becomes larger. Nevertheless, the method given is capable of differentiating between the two

classes of parameters. As the posterior distribution of γ_i is simply a function of the ρ_i, the estimates for γ_i will reflect the class separation seen from the estimates for ρ_i. (give a nod to Linkletter and Savitsky)

However, it is important to consider that the parameters considered above are contained in the RKHS generated by a kernel in the specified family. This is in general not the case for an arbitrary assumed continuous function of interest. Therefore, it very likely that the kernel used to approximate a parameter is misspecified. For example, it is a well known fact that if k is r times differentiable, all elements of its RKHS are as well [Zhou, 2008]. Thus, the RKHS generated by the square exponential kernel only contains analytic functions, which is a very small subset of the continuous functions, with the additional caveat that it does not contain all analytic functions. Ideally, it is desirable in the very least that the differentiability of the parameter corresponds to the differentiability of the covariance function to avoid cases where parameterizations could suggest the calibration parameter is constant when in reality the covariance family is not smooth enough to describe the behavior. Thus, alternative covariance functions such as Matern may prove to be preferable given their more flexibile nature in terms of differentiability compared to the Gaussian kernels. However, as the basis of this paper is to develop methodology, the appropriateness of various covariance families is not explored any further.

3.4.2 Examples

To effectively assess the validity of the model, it is important to test it with considerations to parameter sensitivity, parameter identifiability, and sample size. These components of our model are explored by constructing two toy examples, one of which is expected to have strong identifiability, but lower sensitivity, and the

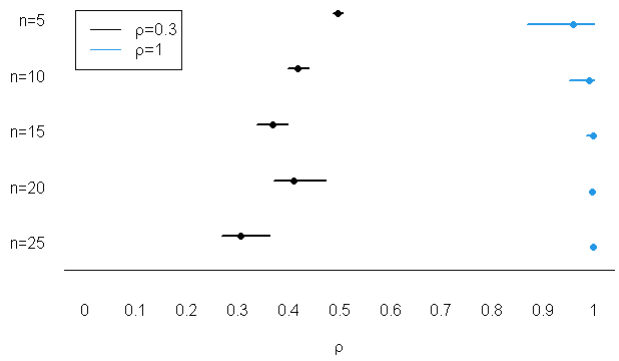


Figure 3.4: Credible intervals for posterior distribution of ρ_i for $\rho_i = 0.3$, and $\rho_i = 1$.

other which has poorer identifiability, but higher sensitivity, and each are tested at sample sizes $5, \ldots, 25$. These features are determined by computing the first and second order Sobol indices, where the `R` package `sensitivity` [Iooss et al., 2020] is used. Effectiveness is assessed by computing the posterior mean of γ_i for 100 separate simulations via the method described in Section 3 and observing 95% credible intervals constructed from the posterior means. In each example, no assumptions will be made *apriori* regarding which parameter is functional or constant.

For simplicity, only two calibration parameters and one control parameter are used in each example, but the model can easily be extended to an arbitrary number of calibration and control parameters. The first example considered is given with

$$\begin{aligned} y(x) &= \eta_1(x, \theta(x)) + \varepsilon \\ \eta_1(x, \mathbf{z}) &= z_1 \sin(2\pi x) + z_2 \cos(2\pi x), \\ \theta_1(x) &= 0.25, \\ \theta_2(x) &= \exp\{-(4x-1)/4\}. \end{aligned}$$

The reasoning for selecting η_1 of this form is that since $\sin(2\pi x)$, $\cos(2\pi x)$ are orthogonal in $L^2([0,1])$, the identifiability of the parameters is guaranteed. This

represents a very idealistic scenario in which the model should produce the correct results. The second example considered is given with

$$\begin{aligned} y(x) &= \eta_2(x, \theta(x)) + \varepsilon \\ \eta_2(x, \mathbf{z}) &= \exp\left\{\frac{xz_1}{1+z_2}\right\} z_1^{\sqrt{z_2}} - \log(1 + z_2^{-xz_1}) \\ \theta_1(x) &= x, \\ \theta_2(x) &= 0.8. \end{aligned}$$

This example is obviously more complicated due to the fact that each term in η_2 is not separable with respect to the calibration and control parameters, and the calibration parameters have much more interaction. This example is therefore not expected to perform as well.

Table 3.1 shows the predicted L^2 error for performing calibration considering all parameters as constant, and also considering all parameters as state-aware, for both examples provided above. The improved performance yielded from considering both θ_1 and θ_2 as state-aware is evident, and suggests that the proposed methodology should indicate at least one parameter is state-aware. A further exploration into the performance of the method is given later.

	Example 1		Example 2	
n	State-Aware	Constant	State-Aware	Constant
5	.037	.194	.117	.490
15	.017	.177	.120	.982
25	.010	.177	.067	.822

Table 3.1: L^2 error observed from calibrating with both state-aware and constant parameters for various sample sizes, for both examples provided above.

The first and second order Sobol indices for η_1 and η_2 are approximated in Table 3.2, as well as the total Sobol indices. As one can see, for η_1, the z parameters

have very little interaction, which suggests strong identifiability. For η_2, the z but their interaction is very large, so we predict that the results will be less consistent than those for η_1.

Component	η_1	η_2
x	0.776	0.133
z_1	0.019	0.699
z_2	-0.013	0.440
xz_1	0.096	-0.156
xz_2	0.122	-0.186
z_1z_2	-0.001	-0.190
Total	η_1	η_2
x	0.995	0.056
z_1	0.116	0.620
z_2	0.108	0.330

Table 3.2: Sobol indices of η_1 and η_2.

Plots of the HPD intervals of γ_1, γ_2 for $n = 5, 10, 15, 20, 25$ (first example then second example) are given in Figure 3.5. Observe that there appears to be some mean shift in the posterior distributions for the functional parameters. It is possibly the case that this is the result of the functions $\theta(x) = x$ and $\theta(x) = \exp\{-(4x-1)/4\}$ not belonging to the RKHS generated by any members of the covariance family given in Section 2 as suggested in Section 4.1, or in the case of η_2, a result of poor identifiability. Nevertheless, for η_1, the model clearly distinguishes between and correctly identifies state-aware and constant calibration parameters. For η_2, the identification is less strong, although it does appear to identify the constant parameter reasonably well as N gets larger.

To summarize, the results suggest that the methodology is indeed valid under some assumptions, but potentially lacks the robustness to handle less well-posed calibration problems well without a large amount of data relative to the size of the domain. In the next section, we will use the presented method to aid in determining

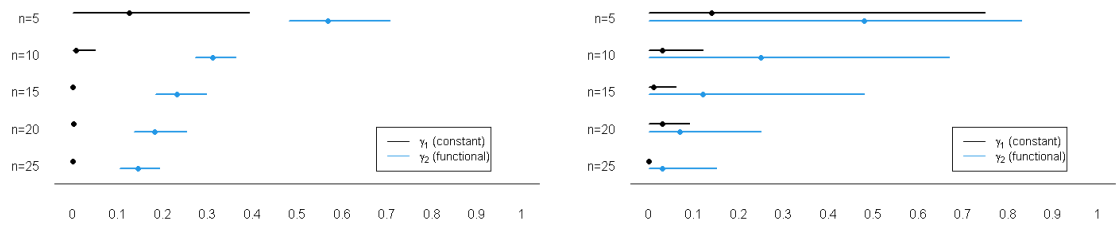


Figure 3.5: 95% credible intervals for each parameter considered at various sample sizes for the first example.

relationships between various parameters describing composite materials.

3.5 Application: Interphase Properties of Composite Materials

Composites are defined as materials consisting of at least two chemically distinct components which, when combined, result in a material with different characteristics than the individual materials, and have become widely popular due to their tunable characteristics. Composites typically consist of a matrix embedded with filler material, and properties of the material are determined by the interaction between the materials. These areas within the material are called the interphase, and are very difficult to observe accurately. However, high fidelity finite element software can be used to predict material properties of composites, but rely on interphase properties as inputs. Explicitly defined physics based models are available, but are not nearly as accurate as they do not incorporate as much information [Arp et al., Submitted].

Given a filler material and a matrix, the construction of a composite material still requires tunable properties such as the proportion of material composed of the filler (volume fraction), and the size of each filler particle which is parameterized by the radius. The effect of volume fraction on material properties (in this case tensile

strength) is well-studied, and included in the finite element software, but the effect of the filler particle size is not as well understood from a theoretical perspective, and it unclear whether it has an impact on interphase properties for larger particle sizes ($< 1\mu m$). Given that the finite element software provided is unit-less, particle size cannot be directly considered, and therefore its effect must be ascertained entirely from experimental data.

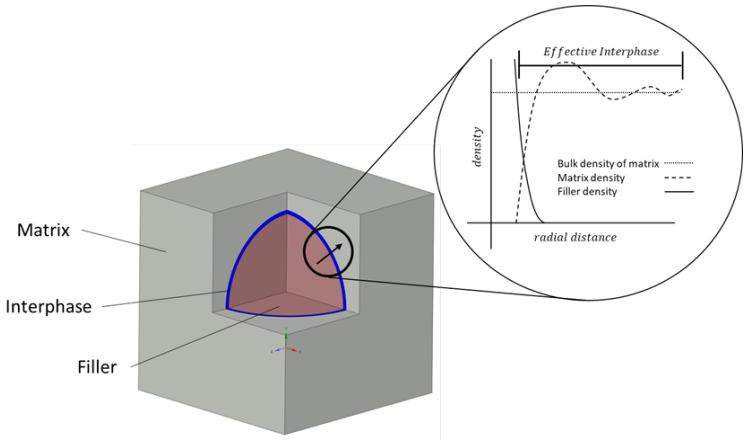


Figure 3.6: A visual representation of a section of composite material with spherical particles [Arp et al., Submitted].

So, the control parameters for the finite element model are volume fraction, interphase *relative* thickness, and interphase modulus, but the interphase *absolute* thickness is considered for the experimental model, as absolute thickness is a preferable measurement for comparison. In the work of Arp et al. [Submitted], interphase modulus was assumed to be a function of particle size and was treated independently for each size, and a simple linear regression model was constructed to describe the relationship between particle size and absolute thickness. Thus, in this formulation, determining whether interphase thickness depends upon particle size reduced to checking if zero was not a credible value for the posterior distribution of the slope parameter. Potential issues arise in this context if interphase modulus is actually

state-aware, or if interphase thickness does not linearly depend on particle size.

A subset of the data given in Vollenberg and Heikens [1989] is used which corresponds to three different matrix materials Polycarbonate (PC), Styrene-Acrylonitrile Copolymer (SAN), and Polypropylene (PP), with embedded spherical glass particles of sizes $4, 30$, and 100 microns (μm). In our analysis, we will treat both calibration parameters as potentially state-aware as given in Section 2, and determine the validity of the assumptions made for interphase modulus as well as compare conclusions for the interphase thickness. So, define $\theta_1(x)$ to be the interphase modulus at particle size x and $\theta_2(x)$ to be the absolute interphase modulus at particle size x, meaning that $\theta_2^*(x) = x^{-1}\theta_2(x)$ describes the relative interphase thickness at particle size x. Note that θ_2 being constant does not correspond to θ_2^* being constant, so one must be sure when sampling to draw from θ_2 and then scale by particle size.

The comparison of considering state-aware calibration is summarized in Table 3.3, where the WAIC [Gelman et al., 2013] as well as the RMSE are compared using a fully constant approach, to a fully state-aware approach. One may note an improvement in using our method for the Glass SAN and Glass PP systems, but not for the Glass PC system. Therefore, it is much more likely that our method selects constant parameters for the Glass PC system, and much more likely to select state-aware parameters for the other two systems.

Material	Constant		State-Aware	
Glass PC	-33.5	(.069)	-26.5	(.041)
Glass SAN	-15.1	(.087)	-26.9	(.059)
Glass PP	-7.8	(.164)	-15.0	(.098)

Table 3.3: This table compares the effect of selecting calibration parameters to be constant versus state-aware during calibration. The metrics given are WAIC on the left and RMSE in parentheses in each cell.

The results are summarized in Figure 3.7, where the first row of plots refer to

the posterior distribution of ρ_i for the Interphase Modulus calibration parameter, and the second row refers to the posterior distribution of ρ_i for the Absolute Interphase Thickness calibration parameter. The spike at one refers to the proportion of ρ_i samples which were associated with $\gamma_i = 0$, and the density refers to the approximate posterior density for the ρ_i samples associated with $\gamma_i = 1$.

One limitation in the methodology in this setting is the lack of variation in particle sizes. The impact of poor identifiability may be exponentiated in this context given that there are only three different points to distinguish between the two different parameters. This is evidenced from the simulation study performed in Section 4.2. However, it is notable that the results agree with those in Arp et al. [Submitted], depending on the threshold for concluding a parameter is state-aware or constant. The results suggest that Interphase modulus is state aware with respect to particle size in all of the systems, and that Interphase thickness is state-aware in each system other than Glass-Polycarbonate. Therefore, despite the potential identifiability issues associated with state-aware calibration, the results align with more standard techniques of determining variable dependence, and the framework additionally is much more robust.

3.6 Conclusions and Future Work

In this paper, we have presented a methodology for identifying state aware parameters in a nonparametric setting by exploiting the parameterization for a widely used family of covariance functions and utilizing strategies in variable selection within Gaussian process regression and otherwise. We have also illustrated under certain conditions that this methodology can be effectively employed for relatively small sample sizes, though its applicability is limited by the identifiability of parameters,

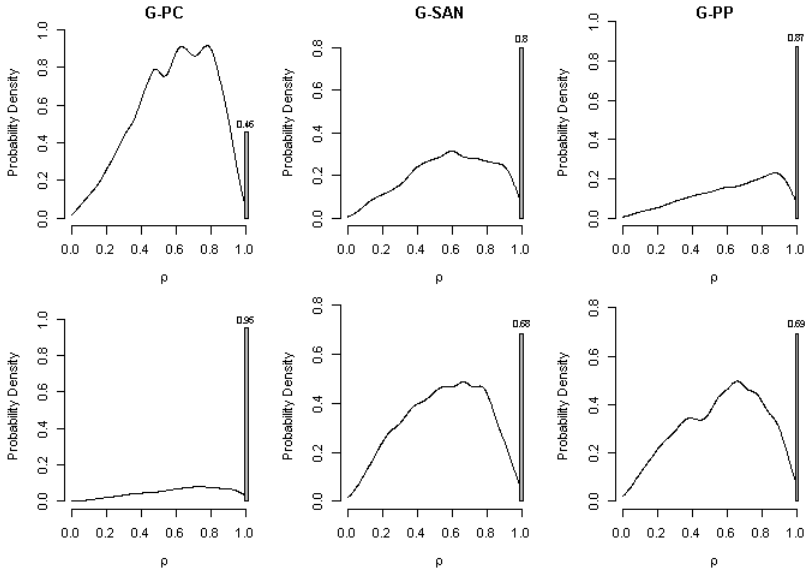


Figure 3.7: Posterior distributions of ρ_i for Interphase Modulus (top) and Interphase Thickness (bottom) for each composite material.

which is a major hurdle to overcome in the field of state-aware calibration in itself. Improvements in this methodology will likely include strategies of mitigating poor identifiability in the context of state-aware calibration. Additional considerations for future work may also include further exploration of the connection between sensitivity and results, and perhaps using sensitivity analysis to influence prior distributions of calibration parameters.

The framework we have constructed admittedly is not a final solution to this problem, but rather an initial step in the direction of performing more effective, and efficient analysis in the area of state-aware calibration. This may present itself in improving predictive power of computer models, or aiding in appropriately simplifying assumptions for running computationally expensive deterministic models to improve efficiency.

Conclusion

The research presented in the previous chapters addresses two distinct approaches for improving the quantification of uncertainty using Gaussian processes. The research involving constrained Gaussian processes previously was *ad hoc*, disjointed, and limited in the cases of linear operator constraints and boundary conditions. The research in chapters 1 and 2 provided a significantly more general approach for which imposing these types of constraints both from a theoretical and practical standpoint, and additionally made suggestions regarding implementation from a modeling and computational perspective. It is also noteworthy that the methodology fits naturally into a Bayesian formulation given that the constraints are imposed based upon conditional Gaussian distributions. The research in chapter 3 provides a new element to the state-aware calibration research which allows one to remove initial assumptions made on the state of calibration parameters, which can lead to a better understanding of a physical system.

As the research in chapters 1 and 2 are related directly to Gaussian process regression, the method of applying constraints has wide applicability. In the case of computer models, our method may be useful to improve the efficiency of simulations where the output is known on an arbitrary subset of the input space. Additional applications may include solving differential equations (including coupled differential equations) or integral equations, as well as providing uncertainty quantification for

each. The research in chapter 3 has wide applicability in the calibration of computer models, particularly in the presence of model bias with respect to experimental data. The ability to identify state-aware parameters in a computer model may also be used to assess initial assumptions made when the computer model was constructed. For example, if the computer model gives output to a partial differential equation using a collection of calibration parameters, our method can be used to identify which are state-aware, which may indicate where model simplifications can be made to improve efficiency. Additionally given the computational complexity of computing Gaussian process likelihoods, in the case of abundant experimental data, our method has shown to be effective when using a subset of the data (see 3.4.2), which may be useful in reducing computational costs. Lastly, in the case of a high dimensional calibration parameter space, our method can be useful in reducing computational costs associated with computing Gaussian process likelihoods for a large number of state-aware parameters.

As the two research projects presented are novel work in their respective fields, there are several future directions that will improve the applicability of each project. The primary weakness of the research in chapters 1 and 2 are the potential computation burden associated with numerical integration, which was required in both chapter 1 and 2. The examples given in each chapter both had two-dimensional domains, and the computations were actually performed in one-dimension, where numerical integration is straightforward. In higher dimensional or more complicated domains, numerical integration using Gaussian quadratures can require a large number of points for equivalent precision, and may be difficult to implement in the case of more complicated domains. It is possible that Monte Carlo sampling will be necessary as a surrogate, or perhaps using simplified basis functions rather than orthogonal polynomials in numerical integration is preferable. Other possibilities may include sampling

strategies which bypass the need for implementing numerical integration entirely. Future directions for the research in chapter 3 involve testing the methodology for larger calibration parameter spaces, as this has not been tested yet. A particularly complicated difficulty to overcome which is present not only in our methodology but in computer model calibration as a whole is identifiability of parameters. This can pose a concern in calibration with constant parameters, and the increased flexibility of state-aware models further complicates the identifiability of parameters. Possible avenues of overcoming this include parameter transformation and strong prior information.

Bibliography

Joshua Arp, John Nicholson, Joseph Geddes, D. Brown, Sezer Atamturktur, and Christopher Kitchens. Inferring effective interphase properties in composites by inverse analysis. *ACS Applied Materials & Interfaces*, Submitted.

S. Atamturktur, J. Hegenderfer, B. Williams, M. Egeberg, R. A. Lebensohn, and C. Unal. A resource allocation framework for experiment-based validation of numerical models. *Mechanics of Advanced Materials and Structures*, 22:641–654, 2015.

Pierre Barbillon, Anabel Forte, and Rui Paulo. Screening the discrepancy function of a computer model. 2021.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Space in Probability and Statistics*. 01 2004. ISBN 978-1-4613-4792-7. doi: 10.1007/978-1-4419-9096-9.

A. Berman and E. J. Nagy. Improvement of a large analytical model using test data. *AIAA Journal*, 21(8):1168–1173, 1983.

Derek Brown and Sez Atamturktur. Nonparametric functional calibration of computer models. *Statistica Sinica*, 28, 02 2016. doi: 10.5705/ss.202015.0344.

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 10:377–408, 10 2006. doi: 10.1142/S0219530506000838.

Hugh Chipman, Edward George, and Robert McCulloch. Bayesian treed models. *Machine Learning*, 48:299–320, 07 2002. doi: 10.1023/A:1013916107446.

Evan Chodora, Garrison Flynn, Trevor Tippetts, and Cetin Unal. Improving the interpretability of physics-based bias in material models. 05 2020. doi: 10.1115/VVS2020-8816.

A. A. Ezzat, A. Pourhabib, and Y. Ding. Sequential design for functional calibration of computer models. *Technometrics*, 60(3):286–296, 2018.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955. URL `https://books.google.com/books?id=ZXL6AQAAQBAJ`.

Edward George and Robert McCulloch. Variable selection via gibbs sampling. *Journal of The American Statistical Association - J AMER STATIST ASSN*, 88:881–889, 09 1993. doi: 10.1080/01621459.1993.10476353.

Robert B. Gramacy and Herbert K. H. Lee. Gaussian processes and limiting linear models, 2008.

Minh Ha Quang. Some properties of gaussian reproducing kernel hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32:307–338, 10 2010. doi: 10.1007/s00365-009-9080-0.

Ofir Harari, A. Dean, D. Bingham, and D. Higdon. Computer experiments: Prediction accuracy, sample size and model complexity revisited. *Statistica Sinica*, 2017.

D. Higdon, M. Kennedy, J. C. Cavendish, J. Cafeo, and R. D. Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004.

Bertrand Iooss, Sebastien Da Veiga, Alexandre Janon, Gilles Pujol, with contributions from Baptiste Broto, Khalid Boumhaout, Thibault Delage, Reda El Amri, Jana Fruth, Laurent Gilquin, Joseph Guillaume, Loic Le Gratiet, Paul Lemaitre, Amandine Marrel, Anouar Meynaoui, Barry L. Nelson, Filippo Monari, Roelof Oomen, Oldrich Rakovec, Bernardo Ramos, Olivier Roustant, Eunhye Song, Jeremy Staum, Roman Sueur, Taieb Touati, and Frank Weber. *sensitivity: Global Sensitivity Analysis of Model Outputs*, 2020. URL `https://CRAN.R-project.org/package=sensitivity`. R package version 1.21.0.

Donald Jones, Matthias Schonlau, and William Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 12 1998. doi: 10.1023/A:1008306431147.

Olav Kallenberg. *Foundations of Modern Probability*. Springer, 1997.

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B*, 63(3):425–464, 2001a.

Marc Kennedy and A O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87, 10 1998. doi: 10.1093/biomet/87.1.1.

Marc Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B*, 63:425–464, 02 2001b. doi: 10.1111/1467-9868.00294.

D.G. Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951. URL `https://journals.co.za/doi/abs/10.10520/AJA0038223X_4792`.

Markus Lange-Hegermann. Linearly constrained gaussian processes with boundary conditions. *ArXiv*, abs/2002.00818, 2020.

Peter Lax. *Functional Analysis*. Wiley, 2002.

Crystal Linkletter, Derek Bingham, Nicholas Hengartner, David Higdon, and Kenny Ye. Variable selection for gaussian process models in computer experiments. *Technometrics*, 48:478–490, 11 2006. doi: 10.1198/004017006000000228.

Jun Liu, W. Wong, and A. Kong. Covariance structure and convergence rate of the gibbs sampler with various scans. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 01 1995. doi: 10.2307/2346091.

Jason Loeppky, Jerome Sacks, and William Welch. Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, 51:366–376, 11 2009. doi: 10.1198/TECH.2009.08040.

G. Matheron. Principles of geostatistics. *Economic Geology*, 58:1246–1266, 1963.

Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 2006.

Antonia Oya, JesÚs Navarro-Moreno, and Juan Carlos Ruiz-Molina. Numerical evaluation of reproducing kernel hilbert space inner products. *IEEE Transactions on Signal Processing*, 57(3):1227–1233, 2009. doi: 10.1109/TSP.2008.2010424.

V. I. Paulsen and M. Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016.

Matthew Plumlee, V. Roshan Joseph, and Hui Yang. Calibrating functional parameters in the ion channel models of cardiac cells. *Journal of the American Statistical Association*, 111:1–31, 12 2015. doi: 10.1080/01621459.2015.1119695.

A. Pourhabib, J. Z. Huang, K. Wang, C. Zhang, B. Wang, and Y. Ding. Modulus prediction of buckypaper based on multi-fidelity analysis involving latent variables. *IIE Transactions*, 47:141–152, 2015.

Pritam Ranjan, Ronald Haynes, and Richard Karsten. A computationally stable approach to gaussian process interpolation of deterministic computer simulation data. *Technometrics*, 53(4):366–378, 2011. ISSN 00401706. URL http://www.jstor.org/stable/41714950.

C.E. Rasmussen, C.K.I. Williams, M.I.T. Press, F. Bach, and ProQuest (Firm). *Gaussian Processes for Machine Learning*. Adaptive computation and machine learning. MIT Press, 2006. ISBN 9780262182539. URL https://books.google.com/books?id=Tr34DwAAQBAJ.

W. Rudin. *Functional Analysis*. Higher mathematics series. McGraw-Hill, 1973. ISBN 9780070542259. URL https://books.google.com/books?id=ehzvAAAAMAAJ.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989a.

Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. Design and Analysis of Computer Experiments. *Statistical Science*, 4(4):409 – 423, 1989b. doi: 10.1214/ss/1177012413. URL https://doi.org/10.1214/ss/1177012413.

T.J. Santner, B.J. Williams, W.I. Notz, and W. Notz. *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. Springer. ISBN 9780387954202.

Terrance Savitsky, Marina Vannucci, and Naijun Sha. Variable selection for nonparametric gaussian process priors: Models and computational strategies. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 26:130–149, 02 2011.

Amar Shah, Andrew Gordon Wilson, and Zoubin Ghahramani. Student-t processes as alternatives to gaussian processes, 2014.

Arno Solin and Simo Särkkä. Hilbert space methods for reduced-rank gaussian process regression. *Statistics and Computing*, 30:419–446, Aug 2019.

Laura Swiler, Mamikon Gulian, Ari Frankel, Cosmin Safta, and John Jakeman. A survey of constrained gaussian process regression: Approaches and implementation challenges. 2020.

Matthias Tan. Gaussian process modeling with boundary information. *Statistica Sinica*, 10 2016. doi: 10.5705/ss.202015.0249.

Aad Vaart and Harry Zanten. Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119, 06 2011.

P.H.T. Vollenberg and D. Heikens. Particle size dependence of the young's modulus of filled polymers: 1. preliminary experiments. *Polymer*, 30(9):1656–1662, 1989. ISSN 0032-3861. doi: 10.1016/0032-3861%2889%2990326-1.

Richard Wilkinson. Accelerating ABC methods using Gaussian processes. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 1015–1023, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. URL `http://proceedings.mlr.press/v33/wilkinson14.html`.

Brian J. Williams, Jason L. Loeppky, Leslie M. Moore, and Mason S. Macklem. Batch sequential design to achieve predictive maturity with calibrated computer models. *Reliability Engineering System Safety*, 96(9):1208–1219, 2011. ISSN 0951-8320. doi: https://doi.org/10.1016/j.ress.2010.04.017. URL `https://www.sciencedirect.com/science/article/pii/S095183201100072X`. Quantification of Margins and Uncertainties.

Alexey Zaytsev, Evgeny Burnaev, and Vladimir Spokoiny. Properties of the bayesian parameter estimation of a regression based on gaussian processes. *Journal of Mathematical Sciences (United States)*, 203:789–798, 11 2014. doi: 10.1007/s10958-014-2168-5.

Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1):456–463, 2008. ISSN 0377-0427. doi: https://doi.org/10.1016/j.cam.2007.08.023. URL `https://www.sciencedirect.com/science/article/pii/S0377042707004657`.