UNCERTAINTY ANALYSIS FOR COMPUTER SIMULATIONS THROUGH VALIDATION AND CALIBRATION

By

John Milburn McFarland

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Mechanical Engineering

May, 2008

Nashville, Tennessee

Approved:

Professor Sankaran Mahadevan

Professor Prodyot K Basu

Professor Bruce Cooil

Professor Gautam Biswas

Dr. Laura Swiler

To my parents and teachers

ACKNOWLEDGEMENTS

This research would not have been possible without the financial support of the National Science Foundation through the Integrative Graduate Education, Research, and Traineeship (IGERT) program. I am also grateful for additional financial support from Sandia National Laboratories, and I would like to thank the project monitors, Dr. Thomas Paez, Dr. Laura Swiler, and Dr. Martin Pilch.

I am honored to have had Dr. Sankaran Mahadevan as my advisor. He has provided extensive professional guidance and support throughout my graduate career, and he is in no small way responsible for the changing attitudes within the engineering community regarding the importance of reliability, uncertainty, and non-deterministic analysis. Without his efforts, much of the work within these fields would not get the recognition it deserves.

I would also like to acknowledge the incredible amount of support that I have received from my mentors at Sandia National Laboratories, especially that of Dr. Laura Swiler, who has also served on my dissertation committee. Dr. Swiler has played an integral role in guiding my research both during my summer internships and during the academic year. I would also like to thank Dr. Tony Giunta for serving as a mentor at Sandia and providing a refreshing perspective on engineering research. Additionally, I am also very grateful to Dr. Vicente Romero for providing the data for the excellent thermally decomposing foam case study and for taking a strong interest in practical issues relating to uncertainty quantification. I also want to thank Dr. Charles Farrar for overseeing the *Los Alamos Dynamics Summer School* program, and for recognizing the importance of technical writing and communications skills.

Finally, I am grateful to all of those students in the Reliability and Risk Engineering and

Management IGERT program with whom I was fortunate enough to work, especially Dr. Ramesh Rebba, Barron Bichon, and Angel Urbina. I was fortunate enough to have Ramesh as a student mentor when I began the program, and I learned very much from him. I also feel lucky to have had the opportunity to work with Barron on several occasions, and I am inspired by the exemplary standards that he sets for his work. Last but not least, I am indebted to Angel for providing me the data for the "three-leg system" case study and for executing the simulator runs for that study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Overview

Modern science and engineering have seen a tremendous growth over the past decade in the use of complex computer simulations, and analysts have been placing increasingly larger demands on the simulation models. Computer codes are being developed to deal with complex fluid-structure interaction, transient behavior, collisions, micro-scale material behavior, and much more. As computers become more powerful, the scientific community has relied more and more heavily on these models. They are used for a variety of tasks, including parameter studies, design, and forecasting, and model predictions are often used to support high-consequence decisions. Simulations are often much less expensive to run than full-scale tests, and in many cases, full-scale tests are not possible at all.

If such importance is to be placed on modeling and simulation, what assurance is there that the results obtained from such models are trustworthy? Even further, is it possible to quantify the amount of error or uncertainty that is associated with model predictions? Such questions are fundamental to the study of model validation and uncertainty quantification.

One avenue for addressing these concerns is to conduct physical experiments. While the purpose of the models themselves is often to predict the behavior of a system that would be prohibitively expensive or impossible to observe empirically, it may be possible to observe the response of a similar system, having perhaps a reduced scale or less extreme loading conditions. The computational simulation in question can then be configured to predict the response

of the same system, and the results may be compared with those observed empirically. Comparing model predictions to observed responses in this manner for the purpose of assessing the suitability of a particular model constitutes what is known as *model validation*. The idea is that the "validation" of particular model outcomes (corresponding to specific, possibly multiple, realizations of the system configuration, geometry, and boundary conditions) lends support to the conclusion that the simulation itself is suitable for its intended purpose.

The process of attempting to validate simulation models in this manner has been of significant interest recently to practitioners and researchers, and in fact Sandia National Laboratories hosted a "model validation challenge workshop" in 2006 (Dowding et al., 2008; Red-Horse and Paez, 2008; Babuska et al., 2008) in which invited speakers were asked to address one of three hypothetical model validation "challenge problems" (two of these challenge problems are addressed in this dissertation, in Sections 6.1 and 6.2). Despite significant research efforts, there is still a lack of universally accepted procedures and approaches for validation assessment, especially with regards to particular "metrics" for developing quantitative measures of agreement between predictions and observations. Existing mathematical tools (such as statistical significance testing) have been adopted in many cases, but such tools are often used without a proper consideration of their relevance to the support of a well-defined conclusion or decision problem.

Uncertainty quantification for simulation models is not strictly limited to model validation, however. In fact, when experimental observations are available for validation assessment, analysts would often like to use the same observations for *model calibration*, which is the process of adjusting internal model parameters in order to improve the agreement between the model predictions and observations. But if internal model parameters are allowed to be adjusted in

this manner, this means that there is some amount of *uncertainty* associated with the true, or best, values of these parameters. And uncertainty associated with model inputs directly implies uncertainty associated with model outputs. Thus, the process of model calibration is actually an opportunity to quantify contributors to the total uncertainty associated with model predictions: if the model calibration process is capable of quantifying the amount of uncertainty in the corresponding parameter estimates, then this uncertainty can be propagated through the simulation in order to quantify the amount of uncertainty implied on the model output.

This dissertation strives to advance the state of the art with respect to the quantification of uncertainty in the modeling and simulation process. A strong focus is placed on the use of Bayesian inference as a tool that enables analysts to develop comprehensive, rigorous representations of uncertainty in parameter estimates in the model calibration process. Model validation assessment is considered as well, emphasizing the development of meaningful quantitative evidence that is pertinent to the model's intended use. Five case studies are presented to illustrate the appropriate use of all approaches discussed herein. Specific research objectives are outlined in the next section.

1.2 Research objectives

As a whole, the overall goal of this dissertation is to advance the current state of the art with respect to uncertainty quantification capabilities, while also producing experience through real world problems that future researchers and practitioners might draw upon whenever experimental observations or other data become available.

More specifically, the research presented in this dissertation can be divided into four distinct objectives, one of which deals with model validation, and three of which are related to model calibration, or parameter estimation. The objectives are as follows:

1. Develop a better understanding of how quantitative statistical decision making tools can be used to support the model assessment (validation) process. In particular, explore the appropriate use of significance testing, including the computation and interpretation of test power.

2. Investigate and extend the uncertainty quantification capabilities of the Bayesian approach for the calibration of computer simulations. Use case studies to illustrate this methodology when (a) the computer simulation is expensive to evaluate; (b) the output of the simulation is highly multivariate, perhaps a function of time and/or space; and (c) the number of calibration inputs is relatively large.

3. Compare different approaches to the estimation and uncertainty representation of internal simulation parameters. In particular, what are the practical differences between the classical nonlinear regression approach and the Bayesian approach? Also, how does the extended Bayesian calibration methodology proposed by Kennedy and O'Hagan, which incorporates a scenario-dependent model inadequacy function, compare to the more conventional Bayesian formulation?

4. When experimental data are available at various levels of modeling complexity (hierarchies; e.g. component, subsystem, system), how should calibration be approached? For example, the parameters governing low-level, constitutive models are typically estimated using data obtained from low-level experiments, but is it possible to obtain better predictions by calibrating the low-level model(s) with high (system)-level data?

1.3 Highlights of the research

In an effort to produce a largely self-contained document, this dissertation is written such that it provides comprehensive developments of several topics which, while not yet "textbook" material, are fairly well-established within the scientific community. As such, there is the potential for confusion regarding what material constitutes original research.

A case in point is Chapter III, which provides a comprehensive treatment from an applied perspective of the surrogate modeling technique known as Gaussian process interpolation. This material has been included in such depth because this technique plays a central role in much of the work discussed in this dissertation. Most of the material presented in Chapter III is not new work (although see below), but the chapter covers a wide variety of topics (including thorough closed-form expressions for the gradients of the likelihood and restricted likelihood functions, expressions for the gradients of the predicted response, modeling of multivariate output, and accounting for uncertainty in training data), a unified treatment of which is probably not otherwise available.

Thus, the highlighting of some of the specific contributions of this work may prove beneficial both to guide the knowledgeable reader towards those sections in which new ideas are discussed, and also to help the uninitiated reader distinguish between new work and established concepts. The following list points out those particular locations in the dissertations at which specific research contributions of interest may be found.

- A novel approach is developed in Section 3.4 that enhances and simplifies the use of Gaussian process interpolation to approximate response quantities that are functions of temporal and/or spatial coordinates.

- Two extensions to the Bayesian model calibration framework are developed:

  - Section 5.3.2 presents a framework whereby one can quantify the effect on the uncertainty analysis when additional model inputs are given prescribed uncertainty distributions. This framework is illustrated for a real-world modeling and simulation problem in Section 6.4.7.
  - Section 5.3.3 presents an approach in which the usual probabilistic model defining the calibration analysis may have multiple error terms. In addition to the usual Gaussian error term, additional error terms might be added to represent, for example, measurement uncertainty that is characterized with bounds. This approach is illustrated in Section 6.4.6.

- Two approaches for dealing with calibration of expensive simulators with highly multivariate output are illustrated:
  - Illustrated in Section 6.4, the surrogate model captures the response as a function of time, enabled by the point selection routine developed in Section 3.4.
  - Illustrated in Section 6.3.3, principal component analysis is used to achieve a lower-dimensional representation of the output, and separate, independent surrogate models are used to approximate each new output quantity.

- A variety of comparisons and discussions of statistical model calibration ideas are presented:
  - Section 5.4 provides a detailed discussion regarding the differences between the basic Bayesian calibration framework presented in Section 5.3 and the extended framework developed by Kennedy and O'Hagan (2001). The implications of the complete Kennedy and O'Hagan framework are poorly understood within the com-

munity because of the framework's complexity, so any effort towards an increased clarity of understanding is likely beneficial.

  - The conventional Bayesian calibration framework and the framework developed by Kennedy and O'Hagan (2001) are compared in terms of implementation and uncertainty quantification by illustrating their application to the calibration of a heat-transfer model in Sections 6.1.3 and 6.1.4.
  - Similarly, the classical nonlinear least-squares approach to parameter estimation uncertainty is compared to the Bayesian approach via the calibration of a model of a thermally decomposing foam element in Section 6.4.3 and 6.4.4.
  - Theoretical equivalencies among non-linear least squares, maximum likelihood, and Bayesian point estimates are considered in Section 5.5 for a broad class of parameter estimation problems.

- Detailed discussion and clarification regarding the use of statistical significance testing for model validation assessment is provided in Section 4.1.2. While statistical significance testing (a.k.a hypothesis testing) is a well-established science, its appropriate use for model validation assessment is often misunderstood. The concepts discussed in Section 4.1.2 are illustrated via a hypothetical model validation challenge problem in Section 6.1.2.

- A novel approach is developed that allows one to characterize and sample from a probability distribution for a high-dimensional, non-Gaussian random vector based on observed sample data. This approach, which is discussed in Section 6.2.2, brings together three established techniques: kernel density estimation (Section 4.2.2), principal component analysis (Section 4.2.3), and Markov Chain Monte Carlo sampling (Section 2.3).

- A "top-down" philosophy for parameter estimation in hierarchical simulation models is proposed and illustrated via the case study of Section 6.5.

- In Chapter VI, five case studies, three of which relate to real-world modeling and simulation projects, are used to illustrate a variety of model validation, calibration, and uncertainty quantification techniques.

1.4 Organization of the dissertation

The remainder of the dissertation is organized such that four chapters providing theory are presented first, followed by Chapter VI, which presents applications and case studies. Most, if not all, of the concepts discussed in the theory chapters are illustrated via case studies. As discussed above, the theory chapters provide a mix of new and existing ideas, but Section 1.3 provides quick guidance as to which constitute new work. Those readers interested primarily in applied research in uncertainty analysis may skip immediately to Chapter VI; however, it should be noted that some theoretical and/or methodological concepts are presented exclusively in Chapter VI (see, for example, Section 6.5), and that much of the theory (particularly the model validation assessment theory) is much more meaningful when considered in terms of an applied case study.

The first chapter dealing in theory is Chapter II, which presents a brief overview of the theory of Bayesian analysis. This chapter is necessary because a large part of the research makes use of the Bayesian framework. Since Bayesian inference provides a rigorous, mathematical treatment of uncertainty in parameter inference, it is particularly useful as a tool for uncertainty analysis in the model calibration setting (see Sections 5.3 and 5.4). While Bayesian inference has also been previously applied to model validation assessment (see Section 4.1.1), this ap-

proach is not pursued here. The use of Bayesian inference as a tool for uncertainty analysis in the model calibration process is illustrated in the applications of Sections 6.1, 6.3, 6.4, and 6.5.

Chapter III presents the theory for Gaussian process (GP) interpolation (which is also known as kriging interpolation). Like Bayesian inference, GP interpolation plays an important role in most of the research that makes up this dissertation. This is because Gaussian process interpolation provides a powerful approach for developing inexpensive surrogate models that allow for comprehensive uncertainty quantification techniques (such as Bayesian inference) to be applied to expensive computer simulations. Further, Section 5.3 discusses how the uncertainty introduced by the Gaussian process response approximation can be explicitly accounted for in the Bayesian calibration process. Gaussian process interpolation is employed for the case studies of Sections 6.2, 6.3, 6.4, and 6.5.

The next chapter, Chapter IV, presents theory for model validation assessment and uncertainty propagation. Model validation assessment is concerned with comparing model predictions with experimentally observed outcomes. Uncertainty propagation involves estimating the probability distribution of a model output, based on an assigned probability distributions for the model inputs. When variability and/or uncertainty are acknowledged, uncertainty propagation is typically a prerequisite for model validation, so they are presented in the same chapter.

Section 4.1 provides an overview of model validation ideas, with an emphasis on significance testing (significance testing is illustrated in Section 6.1, while Section 6.2 illustrates an error characterization approach for model validation assessment). Section 4.2 provides an overview of various uncertainty propagation concepts, including distribution characterization, sampling, and the use of response surface approximations. Principal component analysis and kernel density estimation are also presented here (principal component analysis is applied in

Sections 6.2 and 6.3, and kernel density estimation is applied in Section 6.2).

Next, Chapter V discusses the theory behind the calibration of computer simulation models, which topic constitutes the majority of the research discussed in this dissertation. Of most importance, the Bayesian framework for model calibration is presented in Section 5.3, and this framework is applied in Sections 6.1, 6.3, 6.4, and 6.5.

As mentioned above, Chapter VI presents applications and case studies that illustrate the previously developed concepts. In total, five case studies are presented. The first two case studies address two hypothetical "model validation challenge problems" developed at Sandia National Laboratories (Dowding et al., 2008; Red-Horse and Paez, 2008). The objectives for these two challenge problems are the same (to compare model predictions to experimental observations, and to use given models to predict a failure probability), but different techniques are employed to address each.

The next two case studies, given in Sections 6.3 and 6.4, present applications of the Bayesian calibration methodology discussed in Section 5.3 to real-world modeling and simulation projects at Sandia National Laboratories. The fifth and final case study is included to illustrate the use of a proposed "top-down" approach to the calibration of hierarchical simulation models. The approach is applied to the calibration of a model for a system of nonlinear bolted joints, and the Bayesian framework for parameter inference is again employed.

Finally, some summarizing remarks and recommendations for future work are given in Chapter VII.

Let it also be pointed out that those readers who are accessing an electronic *PDF* form of this document may find that the presence of clickable "hyperlinks" can aid significantly in the navigation of the document. Perhaps of most utility is the "bookmarked" table of contents

sidebar, which provides direct access to all of the sections in the document. Though not as obvious, equation, table, figure, section, and citation references within the text body also have links that direct the reader to the item being referenced.

1.5 Remarks on notation

Before beginning, some brief remarks on notation may prove useful. The content covered in this dissertation is broad enough in scope that it becomes somewhat difficult to develop a unified notation that is consistent with the traditional notation of each discipline that is discussed. In order to minimize confusion, the basic ideas behind the notations used in this document are mentioned here.

One of the few notational conventions that is strictly adhered to throughout is that scalar quantities are typeset in italics (as in x), whereas vector and matrix quantities are typeset in boldface (as in $\boldsymbol{x}$). An attempt is made to denote matrix quantities with uppercase letters or symbols and vector quantities in lowercase, but this convention is not strictly adhered to. A superscript T is used to denote vector and matrix transposition, as in $\boldsymbol{A}^T$. In addition, all vectors are treated as column vectors, so that the inner product of two vectors is written $\boldsymbol{a}^T\boldsymbol{b}$, and vector concatenation is written $\boldsymbol{c} = \left(\boldsymbol{a}^T, \boldsymbol{b}^T\right)^T$.

On a few occasions, a special notation is used to describe the construction of a matrix. The notation $\boldsymbol{A} = \left[f(i,j)\right]_{i,j}$ means that the matrix $\boldsymbol{A}$ can be constructed by employing the specified function $f(\cdot,\cdot)$ to compute each element as a function of the indices i and j.

It is common in the classical probability and statistics literature to denote random variables in uppercase and realizations of a random variable in lowercase (although this convention is not typically seen in the Bayesian literature). While this can be a useful convention, it is not adopted here (primarily because symbols are used to denote certain random variables,

specifically θ and ε, which can make this particular convention somewhat confusing).

The expression $x \sim N(\mu, \sigma^2)$ means that the random variable x follows a normal (a.k.a. Gaussian) distribution with mean μ and variance σ^2. The convention that the second argument for a normal distribution denotes the variance is not strictly adhered to, but should be clear from the context. When specific values are given, the second argument denotes the standard deviation of the distribution, which is the square-root of the variance. This is because the standard deviation has the same units as x. For example: $x \sim N(100, 10)$ inches means that the random variable x follows a normal distribution with a mean of 100 inches and a standard deviation of 10 inches. When a subscript is used, it specifies the dimension of the random variable: for example, $\boldsymbol{x} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ specifies that the three-dimensional random vector $\boldsymbol{x}$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

When referring to probability density functions, expressions such as $f(x) \propto \exp(-\frac{1}{2}x^2)$ are also used. This expression means that the probability density function for the random variable x is equal to $c\exp(-\frac{1}{2}x^2)$, where c is some constant that does not depend on x (in this case $c = (2\pi)^{-1/2}$ and $x \sim N(0, 1)$). This notation is common in the Bayesian literature, in which the constant of proportionality is typically not considered until the end of the analysis. In fact, this notation is particularly useful for expressing improper probability density functions (those that can not be made to integrate to one; see Section 2.2).

This dissertation addresses both Gaussian process modeling and model calibration, two areas whose standard notations can cause confusion when used together. In particular, it is common to use σ^2 to refer to the process variance for a Gaussian process model, while in model calibration σ^2 usually refers to the variance of the error term. To avoid confusion, λ will be used here instead to refer to the process variance in Gaussian process modeling. Also, m

will be used to refer to the number of training points for a Gaussian process model, while n will be used to denote the number of experimental observations available for model calibration.

Finally, in the parameter estimation literature, the most common notation is for the vector $\boldsymbol{\theta}$ to denote those unknown coefficients that are being estimated, and for the vector $\boldsymbol{x}$ to denote the independent variables, or covariates. However, because $\boldsymbol{x}$ is used so often to represent other quantities as well (for example, see Chapter III and Section 4.1.2), $\boldsymbol{s}$ is used to denote the independent variables when discussing parameter estimation.

CHAPTER II

BAYESIAN ANALYSIS

Bayesian inference can be a very powerful tool for quantifying uncertainty when using observed data for parameter estimation. In particular, it is used extensively in this work for quantifying the uncertainty in the model calibration process (see Section 5.3). A brief overview of the Bayesian framework is presented first, followed by a discussion of the "prior distribution," with an emphasis on the formulation of non-informative priors. Finally, the chapter concludes with a presentation of the numerical technique known as Markov Chain Monte Carlo sampling (MCMC). MCMC is one method for constructing the "solution" to a Bayesian inference problem when a more convenient analytical representation is not possible, and it is used exclusively in this dissertation for all Bayesian computations.

2.1 Introduction to Bayesian analysis

Bayesian statistical analysis differs from classical (or frequentist) statistics fundamentally by the two camps' interpretations of probability. In classical statistics, the meaning of probability is directly related to frequency of occurrence. What sets Bayesians apart is that they allow probability and probability distributions to connote belief or uncertainty about uncertain parameters. Thus, Bayesian analysis begins with what is known as a "prior" distribution for the uncertain parameters, denoted $\pi(\boldsymbol{\theta})$. Knowledge about the uncertain parameters is then updated by observations, $\boldsymbol{d}$, to arrive at what is called the "posterior" distribution of $\boldsymbol{\theta}$. This

process is expressed formally through what is known as Bayes' theorem:

$$f(\boldsymbol{\theta} \mid \boldsymbol{d}) = \frac{\pi(\boldsymbol{\theta}) f(\boldsymbol{d} \mid \boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta}) f(\boldsymbol{d} \mid \boldsymbol{\theta})\, d\boldsymbol{\theta}}, \tag{2.1}$$

where $f(\boldsymbol{d} \mid \boldsymbol{\theta})$ is known as the likelihood function of $\boldsymbol{\theta}$, and is commonly denoted $L(\boldsymbol{\theta})$ because the data in $\boldsymbol{d}$ hold a fixed value once observed.

The meaning of Bayes' theorem is that the posterior distribution of $\boldsymbol{\theta}$ is proportional to the prior times the likelihood (note that the integral in the denominator functions to normalize the posterior distribution so that is has a total area of one). It is worth noting that while many classical statisticians are not comfortable with the Bayesian philosophy because of the apparent subjectivity present in formulating prior distributions, there do exist guidelines for selecting appropriate vague "reference" prior distributions whose purpose is to represent the absence of prior knowledge (this is discussed in Section 2.2). In fact, there are many cases in which classical results can be derived using Bayesian analysis with reference prior distributions.[1]

The primary computational difficulty in applying Bayesian analysis is the evaluation of the integral in the denominator of Eq. (2.1), particularly when dealing with multiple unknowns. Unless the relationship between the data and the unknowns is very simple and a particular type of prior distribution is used (such that the prior and the likelihood form what is known as a "conjugate pair"[2]), then numerical methods will be needed.

[1] For example, a Bayesian linear regression analysis can be used to construct a posterior distribution for the regression coefficients. Based on this posterior distribution, one can derive the Bayesian equivalent of confidence intervals for the coefficients, and if the standard reference prior distribution is employed, the confidence intervals turn out to be the same as those obtained via classical analysis (not to mention the point estimates are also the same). See Lee (2004) for the Bayesian derivation, and Devore (2000), for example, for the classical derivation. Another example in which the requirement to formulate a prior distribution does not prevent one from reproducing more intuitive classical results is discussed in Section 5.5.

[2] The most common example of a conjugate pair is possible when the unknown parameter is the mean of a normal distribution. In this case, representing the prior distribution of the mean with a normal distribution results in the posterior also having a normal distribution. The term "conjugate" is used because both the prior and the posterior have the same distribution type.

Computation of the posterior distribution is essentially a numerical integration problem; while standard numerical integration techniques such as numerical quadrature are occasionally used, specialized techniques are generally preferable. In particular, the technique known as Markov Chain Monte Carlo (MCMC) sampling is especially widespread. This technique is popular because it is simple to implement, is effective for high-dimensional problems, and provides a convenient representation of the posterior distribution (in the form of random samples). MCMC sampling is presented in Section 2.3.

2.2 The prior distribution

In Bayesian analysis, the prior distribution, $\pi(\boldsymbol{\theta})$, is a representation of all knowledge and information about the unknowns, before accounting for the observed data $\boldsymbol{d}$. The fact that the use of a prior distribution is a requirement is often a point of contention for classical statisticians, who feel that the use of a prior distribution will "bias" the results. This section will briefly discuss appropriate use of the prior distribution, in particular providing guidance for the case in which one would like the prior distribution to convey a complete lack of information.

The prior distribution is simply a way to represent one's state of knowledge about the unknowns before observing the data $\boldsymbol{d}$. It is possible to use a prior distribution that represents a large amount of prior information, thus dominating the effect of observed data, and this is the primary reason why many statisticians are wary of its use. However, in most cases, logical choices also exist for prior distributions that capture the notion of a lack of prior information. Such distributions are often called "vague" prior distributions, and they are sometimes termed "reference" priors because they allow various analysts to compare results based on a standard prior distribution.

The simplest case of a vague prior distribution applies when dealing with a scalar unknown

that has support over the entire real line. The most common example is when making inference about the mean of a normal distribution. The standard reference prior for this case captures the notion that *a priori*, any value of the unknown is equally likely:

$$\pi(\theta) \propto 1. \tag{2.2}$$

Thus, the prior probability density is simply proportional to a constant, and is independent of the value of θ. It is easy to see that for this prior distribution, the posterior will be proportional to the likelihood only. This concept also extends naturally to inference about multiple unknowns, in which case the corresponding reference prior would be $\pi(\boldsymbol{\theta}) \propto 1$.

Some will notice that the prior of Eq. (2.2) is not a proper probability distribution because it does not integrate to one. However, it is still considered an acceptable choice because in many cases of interest (for example, a normal likelihood, which is almost ubiquitous in Bayesian model calibration) it will combine with the likelihood to form a proper posterior (Lee, 2004).

Another common vague reference prior is that for the variance of a normal distribution. The standard reference prior for this case is (Lee, 2004):

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}, \tag{2.3}$$

which is uniform in both $\log \sigma^2$ and $\log \sigma$. Clearly, this is also an improper distribution.

It is interesting to note that the reference prior distributions given by Eqs. (2.2) and (2.3) do have a theoretical basis. In fact, both of these prior distributions can be derived using what is known as Jeffreys' rule (Jeffreys, 1961), which was developed because the resulting prior distribution has the desirable property that it is invariant to the particular scale in which the

parameter is measured.

It is also often the case that there are parameters representing both "location" and "scale." In such cases, Lee (2004) suggests that it is reasonable to think of such parameters as being *a priori* independent, and that an appropriate reference prior distribution is the product of the reference prior distribution obtained for each parameter separately. For example, when making inference about the normal distribution with both the mean and variance unknown, the recommended reference prior distribution would be the product of Eqs. (2.2) and (2.3), which gives

$$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}. \tag{2.4}$$

One particularly useful feature of the prior distribution is its ability to enforce bound constraints on the unknowns. In fact, in many cases it is possible to use the prior distribution to represent any nonlinear constraint, subject only to the requirement that one can compute whether or not a given value $\boldsymbol{\theta}$ meets the constraint. For example, a very general extension of the vague prior of Eq. (2.2) is a prior which is proportional to a constant when the constraints are met and zero otherwise:

$$\pi(\boldsymbol{\theta}) \propto \begin{cases} 1, & \boldsymbol{\theta} \in \Omega, \\ 0, & \boldsymbol{\theta} \notin \Omega, \end{cases} \tag{2.5}$$

where Ω is the region containing feasible values for $\boldsymbol{\theta}$.

2.3 Markov Chain Monte Carlo simulation

Markov Chain Monte Carlo (MCMC) simulation is a numerical simulation method that is often used in Bayesian analysis to construct the posterior distribution when no analytical expression is available. MCMC simulation works by generating random samples from the target distribu-

tion (typically a Bayesian posterior), and is especially powerful when dealing with multivariate distributions. MCMC methods can be used whenever the target density is known at least up to proportionality constant, and are thus well suited for Bayesian analysis, since the complicated integral expression for the normalizing constant can often be very difficult to evaluate.

The idea behind the Markov Chain Monte Carlo method is to construct a Markov chain such that its stationary distribution is exactly the same as the distribution of interest (in this case the Bayesian posterior). The particular MCMC implementation used in this work is known as the Metropolis algorithm (Metropolis et al., 1953; Chib and Greenberg, 1995), and it is a form of rejection sampling. The algorithm is fairly simple to implement, and it can be used to generate samples from both univariate and multivariate densities. Consider that one wants to generate samples from a univariate density $f(x)$ that can be evaluated up to a proportionality constant, such that $\tilde{f}(x)$ is known, where $\tilde{f}(x) \propto f(x)$ (in Bayesian inference, $\tilde{f}(x) = \pi(x)L(x)$). In this case, the Metropolis algorithm can be implemented as follows:

1. Set $i = 0$ and choose a starting value, x_0.

2. Initialize the list of samples: $\boldsymbol{X} = \{x_0\}$.

3. Repeat the following steps many times:

   (a) Sample a candidate x^* from the proposal density function $q(x^* \mid x_i)$.

   (b) Calculate the acceptance ratio $\alpha = \min\left[1, \dfrac{\tilde{f}(x^*)}{\tilde{f}(x_i)}\right]$.

   (c) Generate a random number u from the uniform distribution on $[0, 1]$.

   (d) If $u < \alpha$, set $x_{i+1} = x^*$, otherwise set $x_{i+1} = x_i$.

   (e) Augment the list of sampled values, $\boldsymbol{X}$, by x_{i+1}.

   (f) Increment i .

4. After convergence is reached, the list of samples $\boldsymbol{X}$ can be be used to construct an approximation to the target density $f(x)$.

The proposal density $q(x^* \mid x_i)$ defines a probability density that generates random moves x^* based on the current point x_i. In theory, the only restriction on the choice of proposal density $q(\cdot \mid \cdot)$ is that it be symmetric with respect to its arguments, i.e. the probability of going from x_i to x^* is the same as that of going from x^* to x_i. An extension of this algorithm, known as the Metropolis-Hastings algorithm, allows the proposal density to have any form.

Convergence of the chain is generally achieved fairly quickly. However, poor choices for the starting value, x_0, may cause the chain to take many samples to reach its stationary distribution. A simple method for assessing convergence is to look at a trace plot of the samples.

In general, the user must only specify the starting value and the proposal density, $q(\cdot \mid \cdot)$. Unfortunately, the performance of the algorithm can be sensitive to both of these choices, particularly the choice of proposal density. The most commonly used proposal density is the random walk density, in which the candidate point is given by $x^* = x_i + \eta$, where η is a random variable chosen to be symmetric about the origin. The choice of the variance of η is critical to the performance of the algorithm. If the moves are very small and the acceptance probability is very high, most moves will be accepted but the chain will take a large number of iterations to converge. If the moves are large, they are likely to fall in the tails of the posterior distribution and result in a low value of the acceptance ratio. One wants to cover the parameter space in a computationally efficient fashion. Many studies have been done on optimal acceptance rates, and the results seem to indicate that 0.45–0.5 is the optimal acceptance rate for 1-dimensional problems, whereas 0.23–0.25 is the optimal acceptance rate for high-dimensional problems (Gilks et al., 1996).

When the target density is multivariate, there are two options for generating random samples: the candidate moves can be made in all dimensions simultaneously, or the moves can be made on one component at a time. Choosing a joint proposal density that is a good approximation to the target can be a difficult task, and the added complexity of working with multivariate densities often makes it undesirable to generate multi-dimensional candidate moves. The componentwise scheme discussed by Hastings (1970) and Chib and Greenberg (1995) allows candidate moves to be made on each component independently. A proposal density is specified for each component of $\boldsymbol{x}$, and the acceptance ratio for a particular move is given by $\alpha_i = \min\left[1, \frac{f(x_i^* \mid \boldsymbol{x}_{-i})}{f(x_i \mid \boldsymbol{x}_{-i})}\right]$, where $f(x_i \mid \boldsymbol{x}_{-i})$ denotes the full conditional density of the i^{th} component. Thus, the components of $\boldsymbol{x}$ are sampled sequentially from their respective full conditional densities. This method is similar to Gibbs sampling, and it is often referred to as Metropolis-within-Gibbs. Consider the acceptance ratio for a move on component i, $\alpha_i = \min\left[1, \frac{f(x_i^* \mid \boldsymbol{x}_{-i})}{f(x_i \mid \boldsymbol{x}_{-i})}\right]$. Note that the full conditional density of x_i is given by $f(x_i \mid \boldsymbol{x}_{-i}) = \frac{f(x_i, \boldsymbol{x}_{-i})}{f(\boldsymbol{x}_{-i})}$. In computing the acceptance ratio, the marginal density $f(\boldsymbol{x}_{-i})$ will cancel because only the i^{th} component of $\boldsymbol{x}$ is varying. Thus, the acceptance ratio becomes $\alpha_i = \min\left[1, \frac{f(x_i^*, \boldsymbol{x}_{-i})}{f(x_i, \boldsymbol{x}_{-i})}\right]$, which can be computed as long as the joint density $f(\boldsymbol{x})$ is known up to a proportionality constant.

The nature of MCMC sampling is that the samples obtained in this fashion will almost always show a strong degree of serial correlation, depending on the particular proposal distribution being used. For this reason, one generally makes inference about the posterior distribution using a very large number of samples (typically on the order of 10,000, or more). The nature of the resulting Markov chain should also be kept in mind if one wants to generate a "small" random sample from the posterior distribution. For example, if the posterior distribution is sim-

ulated using 20,000 MCMC samples, and 100 random samples from the posterior are needed, it would not be appropriate to take 100 consecutive samples from the chain. Instead, one might either choose the samples from the chain at evenly spaced intervals of 200, or choose the 100 samples randomly from the 20,000 available.

2.4 Summary

This chapter provides the theoretical background for Bayesian inference, including discussion of the prior distribution and the numerical sampling technique known as Markov Chain Monte Carlo sampling. Bayesian inference provides a mathematical theory for representing parameter uncertainty in terms of probability density functions. As such, it is employed extensively in this dissertation, especially as a means for constructing a rigorous framework for addressing uncertainty in the model calibration process (see Chapter V). Note that this is a review chapter, and none of the material presented herein constitutes original research.

CHAPTER III

GAUSSIAN PROCESS INTERPOLATION

3.1 Introduction

The capability to construct an efficient approximation to a complex functional relationship can be beneficial to many quantitative fields. Often times a computer simulation is developed to describe the relationship among a set of input and output quantities. Such a simulation may serve a variety of purposes, but time and/or cost constraints often prevent the use of the simulation itself to exhaustively explore the relationship between the inputs and outputs. In this case, a "surrogate model" or "response surface approximation" might be developed as an inexpensive approximation of the functional relationship that is described by the computer simulation. Being cheap to evaluate, the surrogate model could then be used to support a variety of traditionally expensive iterative procedures, such as optimization, uncertainty propagation, and calibration.

Gaussian process (GP) interpolation (which is in most cases equivalent to the family of methods that go by the name of "kriging" predictors) is a powerful technique based on spatial statistics that has recently gained interest in the engineering community for its potential as a surrogate modeling technique. GP modeling uses a set of observed inputs and outputs (the "training data"; for example the results from ten different runs of a computer simulation) to construct an approximation to the underlying relationship. In most cases, one wants the resulting approximation to directly interpolate the observed data (as in the case of a surrogate to a deterministic computer simulation), and GP models are typically constructed in this man-

ner, but the flexibility does exist to construct GP models that instead "smooth" or regress the observations (such models are discussed in Section 3.5).

One of the primary advantages of GP interpolation is that it is a non-parametric technique, which means that *a priori* assumptions about the functional relationship that exists between the inputs and the outputs (e.g., a linear relationship) are not required. However, the framework is still quite flexible: assumptions about smoothness properties can be reflected in the model, and large-scale variations can be captured via a parametric trend function.

The GP model has another significant feature, which is the ability to provide a direct representation of the uncertainty associated with its interpolative approximation. This uncertainty representation can be quite useful, and it has been used previously to improve the efficiency of both optimization (Jones et al., 1998) and reliability estimation (Bichon et al., 2008).

While Gaussian process modeling can be quite powerful, there is unfortunately a steep learning curve needed to obtain a working understanding of the methodology, and there are several potential pitfalls. This chapter provides a comprehensive coverage of the practical considerations relevant to GP modeling. Additional reference information can be obtained from Rasmussen (1996); Martin and Simpson (2005); Mardia and Marshall (1984); Santner et al. (2003). Section 3.2 presents the basic theory, while Section 3.3 provides a detailed description of the parameter estimation process, which is often the most challenging aspect of applying Gaussian process interpolation for surrogate modeling. Section 3.4 discusses the use of GP interpolation when the response is a function of temporal and/or spatial coordinates. Finally, Section 3.5 discusses GP models that do not directly interpolate the training data.

3.2 Gaussian process models

Consider that one wants to build an approximation to a function of a vector-valued input $\boldsymbol{x}$, based only on m observations of the inputs and outputs: $Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_m)$. Appropriate approaches to choosing those inputs $\boldsymbol{x}$ for which the simulator should be run, known as the design of computer experiments, are discussed by Sacks et al. (1989b); Simpson et al. (1997); Sacks et al. (1989a); McKay et al. (1979); Sacks and Schiller (1988); Welch (1983); Morris et al. (1993); Currin et al. (1991). The basic idea of the GP interpolation model is that the outputs, Y, are modeled as a Gaussian process that is indexed by the inputs, $\boldsymbol{x}$. A Gaussian process is simply a set of random variables such that any finite subset has a multivariate Gaussian distribution[1]. A Gaussian process is defined by its mean function and covariance function, which in this case are functions of $\boldsymbol{x}$. Once the Gaussian process is observed at m locations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, the conditional distribution of the process can be computed at any new location, $\boldsymbol{x}^*$, which provides both an expected value and variance (uncertainty) of the underlying function.

The key here is that the function describing the covariance among the outputs, Y, is a function of the inputs, $\boldsymbol{x}$. The covariance function is constructed such that the covariance between two outputs is large when the corresponding inputs are close together, and the covariance between two outputs is small when the corresponding inputs are far apart. As shown below, the conditional expected value of $Y(\boldsymbol{x}^*)$ is a linear combination of the observed outputs,

[1]The Gaussian model assumption is needed to derive the fundamental equations within the stochastic process framework. Interestingly, the same equations can be arrived at using different arguments (within a framework known as kriging), and it turns out the Gaussian assumption is in fact not needed to derive the equation for the predictor and its variance (mean-squared error). Nevertheless, the Gaussian assumption does come into play when one wants to construct confidence intervals for the true value of the function. O'Hagan (2006) has this to say about the Gaussian assumption: "Although the assumption of normality, implicit in the use of a GP, may seem to represent rather a strong limitation, in practice it has no impact if we can make enough runs of the computer code to produce an accurate emulation. In more complex problems, where it is not practical to make enough code runs to emulate with negligible code uncertainty, normality can matter." The use of transformations of the code output (such as a logarithmic transformation) in those cases when the uncertainty is significant and the assumption of normality does not seem appropriate is a subject of ongoing research.

$Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_m)$, in which the weights depend on how close $\boldsymbol{x}^*$ is to each of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ (one is reminded of radial basis functions). In addition, the conditional variance (uncertainty) of $Y(\boldsymbol{x}^*)$ is small if $\boldsymbol{x}^*$ is close to the training points and large if it is not.

Further, the GP model may incorporate a systematic, parametric trend function whose purpose is to capture large-scale variations. This trend function can be, for example, a linear or quadratic regression of the training points. It turns out that this trend function is actually the (unconditional) mean function of the Gaussian process. The effect of the mean function on predictions that interpolate the training data tends to be small, but when the model is used for extrapolation, the predictions will follow the mean function very closely as soon as the correlations with the training data become negligible.

To develop the theory, let $Y(\boldsymbol{x})$ denote a Gaussian process with mean and covariance given by

$$\mathrm{E}\big[Y(\boldsymbol{x})\big] = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} \tag{3.1}$$

and

$$\mathrm{Cov}\big[Y(\boldsymbol{x}), Y(\boldsymbol{x}^*)\big] = \lambda c(\boldsymbol{x}, \boldsymbol{x}^* \mid \boldsymbol{\xi}), \tag{3.2}$$

where $\boldsymbol{f}^T(\boldsymbol{x})$ defines q basis functions for the trend, and is given by 1 for a constant trend and $[1 \ \boldsymbol{x}^T]^T$ for a linear trend; $\boldsymbol{\beta}$ gives the coefficients of the regression trend; $c(\boldsymbol{x}, \boldsymbol{x}^* \mid \boldsymbol{\xi})$ is the correlation between $\boldsymbol{x}$ and $\boldsymbol{x}^*$; and $\boldsymbol{\xi}$ is the vector of parameters governing the correlation function. While the process variance is typically denoted by σ^2, λ is used instead throughout this dissertation in order to avoid confusion with the error variance in calibration analysis (Chapter V).

Consider that the process has been observed at m locations (the training or design points) $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ of a d-dimensional input variable, yielding the resulting observed random vector

$\boldsymbol{Y} = \big(Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_m)\big)^T$. By definition, the joint distribution of $\boldsymbol{Y}$ satisfies

$$\boldsymbol{Y} \sim \boldsymbol{N}_m \left(\boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta},\ \lambda \boldsymbol{R}\right), \tag{3.3}$$

where $\boldsymbol{R}$ is the $m \times m$ matrix of correlations among the training points. Under the assumption that the parameters governing both the trend function and the covariance function are known, the conditional expected value and variance (uncertainty) of the process at an untested location $\boldsymbol{x}^*$ are calculated as

$$\mathrm{E}\big[Y(\boldsymbol{x}^*) \mid \boldsymbol{Y}\big] = \boldsymbol{f}^T(\boldsymbol{x}^*)\boldsymbol{\beta} + \boldsymbol{r}^T(\boldsymbol{x}^*)\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta}) \tag{3.4}$$

and

$$\mathrm{Var}\big[Y(\boldsymbol{x}^*) \mid \boldsymbol{Y}\big] = \lambda \left(1 - \boldsymbol{r}^T\boldsymbol{R}^{-1}\boldsymbol{r}\right), \tag{3.5}$$

where $\boldsymbol{F}$ is an $m \times q$ matrix with rows $\boldsymbol{f}^T(\boldsymbol{x}_i)$ (the trend basis functions at each of the training points), and $\boldsymbol{r}$ is the vector of correlations between $\boldsymbol{x}^*$ and each of the training points. Further, the full covariance matrix associated with a vector of predictions can be constructed using the following equation for the pairwise covariance elements:

$$\mathrm{Cov}\big[Y(\boldsymbol{x}), Y(\boldsymbol{x}^*) \mid \boldsymbol{Y}\big] = \lambda \left[c(\boldsymbol{x}, \boldsymbol{x}^*) - \boldsymbol{r}^T\boldsymbol{R}^{-1}\boldsymbol{r}_*\right], \tag{3.6}$$

where $\boldsymbol{r}$ is the vector of correlations between $\boldsymbol{x}$ and each of the training points, and $\boldsymbol{r}_*$ is the vector of correlations between $\boldsymbol{x}^*$ and each of the training points.

When the coefficients of the trend function are not known, but are estimated using a generalized least squares procedure, or equivalently, maximum likelihood, the variance estimate of

Eq. (3.5) can be expanded as (Ripley, 1981): [2]

$$\mathrm{Var}\big[Y(\boldsymbol{x}^*) \mid \boldsymbol{Y}\big] = \lambda\Big\{1 - \boldsymbol{r}^T\boldsymbol{R}^{-1}\boldsymbol{r} + \Big[\boldsymbol{f}(\boldsymbol{x}^*) - \boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{r}\Big]^T\Big[\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{F}\Big]^{-1}\Big[\boldsymbol{f}(\boldsymbol{x}^*) - \boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{r}\Big]\Big\}, \quad (3.7)$$

which can also be written in matrix form as

$$\mathrm{Var}\big[Y(\boldsymbol{x}^*) \mid \boldsymbol{Y}\big] = \lambda - \Big[\boldsymbol{f}^T(\boldsymbol{x}^*),\ \boldsymbol{r}^T\Big]\begin{bmatrix}\boldsymbol{0} & \boldsymbol{F}^T\\ \boldsymbol{F} & \boldsymbol{R}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{f}(\boldsymbol{x}^*)\\ \boldsymbol{r}\end{bmatrix}. \quad (3.8)$$

Note that when using Eq. (3.5), the variance of Y takes values between 0 and λ. If $\boldsymbol{x}^*$ is equal to or very close to one of the training points, then the term $\boldsymbol{r}^T\boldsymbol{R}^{-1}\boldsymbol{r}$ will go to 1, and the variance of $Y(\boldsymbol{x}^*)$ will be 0 (because there is no uncertainty at the tested locations). The effect of a known trend function on this prediction variance manifests itself through the value of the parameter λ. If the trend function captures much of the variation of Y, then the "maximum likelihood" estimate of λ will be smaller (for more information on this, refer to Section 3.3).

When the trend function coefficients are not assumed known, the prediction variance of Eq. (3.7) still has a lower bound of 0, but there is no upper bound. In this case, the term $\big[\boldsymbol{f}(\boldsymbol{x}^*) - \boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{r}\big]$ also goes to zero at a location equal to one of the training points. When $\boldsymbol{x}^*$ interpolates the training points, this author has not found a significant difference in the prediction uncertainties given by Eqs. (3.5) and (3.7), although Eq. (3.7) is significantly more expensive to evaluate.

[2] There are some notational peculiarities that are specific to the models developed from the standpoint of "kriging," as opposed to Gaussian processes. In kriging, Eq. (3.4) is termed the "universal kriging predictor" when $\boldsymbol{\beta}$ is replaced by its generalized least squares estimate and the parameters λ and $\boldsymbol{\xi}$ are assumed known (although in practice the covariance parameters are rarely known). When $\boldsymbol{\beta}$, λ, and $\boldsymbol{\xi}$ are all estimated using the maximum likelihood procedure, the estimate given by (3.4) is sometimes referred to as the "unified universal kriging predictor" (Mardia and Marshall, 1984).

There are several different methods of parametrizing the correlation function. The form implemented by this author is the squared exponential form, given by

$$c(\boldsymbol{x}, \boldsymbol{x}^*) = \exp\left[-\sum_{i=1}^{d} \xi_i (x_i - x_i^*)^2\right], \tag{3.9}$$

where d is the dimension of $\boldsymbol{x}$, and the d parameters ξ_i must be non-negative. The exponent must lie in the range $[0, 2]$ in order for the covariance matrix to be positive definite, but the value 2 is usually chosen because it produces a function that is infinitely differentiable. This form of the correlation function dictates that the degree of correlation of the outputs depends on the closeness of the inputs.

Also, the relative magnitudes of the parameters $\boldsymbol{\xi}$ are related to the amount of importance each dimension of $\boldsymbol{x}$ has in predicting the output Y: a large value for ξ_i (i.e., a small correlation length) indicates a high amount of "activity" (and likewise a low amount of correlation) in that direction. For example, if the response is independent of one of the inputs, then that input will have an infinite correlation length (because the response does not change in that direction) and a ξ of 0.

If the Gaussian process predictor is to be used to estimate gradients of the response, it can be tempting to estimate the gradients using finite differencing because the surrogate is so cheap to evaluate. This can be dangerous, though, because for even modest finite difference step sizes, numerical round-off error can render such estimates useless. Fortunately, it is not difficult to derive the exact expressions for the gradients. Only the case of a constant trend function is considered here, but more general expressions for the gradients are available (see, for example, Vazquez and Walter, 2005).

For the constant trend case, the predictor is given by

$$\mathrm{E}\big[Y(\boldsymbol{x}) \mid \boldsymbol{Y}\big] = \beta + \boldsymbol{r}^T(\boldsymbol{x})\boldsymbol{R}^{-1}(\boldsymbol{Y} - \mathbf{1}\beta). \tag{3.10}$$

Since only $\boldsymbol{r}$ is a function of $\boldsymbol{x}$, it is easy to show that by using the chain rule, the derivative of Y with respect to x_k is

$$\frac{\partial\,\mathrm{E}\big[Y(\boldsymbol{x}) \mid \boldsymbol{Y}\big]}{\partial x_k} = \frac{\partial\,\mathrm{E}\big[Y(\boldsymbol{x}) \mid \boldsymbol{Y}\big]}{\partial \boldsymbol{r}}\frac{\partial \boldsymbol{r}}{\partial x_k} = \dot{\boldsymbol{r}}_k^T\boldsymbol{R}^{-1}(\boldsymbol{Y} - \mathbf{1}\beta), \tag{3.11}$$

where $\dot{\boldsymbol{r}}_k$ is the derivative of $\boldsymbol{r}$ with respect to x_k.

Since the values of $\boldsymbol{x}$ are typically standardized before computing the correlations (see Section 3.3.3), this transformation must be accounted for in the computation of $\dot{\boldsymbol{r}}_k$. Consider a linear transformation for each component of x to a standardized space: $x'_i = a_i + b_i x_i$. Then by the chain rule, and for the correlation function of Eq. (3.9), the derivative of the correlation vector with respect to x_k is given by

$$\dot{\boldsymbol{r}}_k = \frac{\partial \boldsymbol{r}}{\partial x_k} = \frac{\partial \boldsymbol{r}}{\partial x'_k}\frac{\partial x'_k}{\partial x_k} = \left[\frac{\partial c(\boldsymbol{x}, \boldsymbol{x}^{(i)})}{\partial x'_k}\right]_i \frac{\partial x'_k}{\partial x_k} = \left[-2\xi_k\left(x'_k - x'^{(i)}_k\right) c(\boldsymbol{x}, \boldsymbol{x}^{(i)}) b_k\right]_i, \tag{3.12}$$

where $\boldsymbol{x}^{(i)}$ is the i^{th} training point, and $c(\boldsymbol{x}, \boldsymbol{x}^*)$ is computed based on the standardized values of $\boldsymbol{x}$, as usual.

Finally, some remarks on the dimensionality of the input are probably warranted. Many users would like to know how many training points are typically required to model a function having a certain number of inputs, and for how many inputs it may be feasible to use the Gaussian process interpolation approach. The number of training points needed may actually depend more on the complexity of the function than on the dimensionality of the input, making

it difficult to provide any general rules of thumb. However, as the number of inputs increases, there is certainly more "space" between the training points, suggesting that the number of points needed will generally increase rapidly with the dimensionality of the input. However, this effect is diminished somewhat by the fact that in practice, most models are not strongly sensitive to all of their inputs (O'Hagan, 2006).

In fact, through the process of estimating the correlation parameters (discussed below), those inputs to which the model is most sensitive are identified (with the parametrization of Eq. (3.9), larger values of ξ correspond to more important inputs). As pointed out by O'Hagan (2006), the Gaussian process approach effectively "projects points down through those smooth [unimportant] dimensions into the lower dimensional space of inputs that matter." While it is true that there has been limited experience with very high dimensions in practice, and most realistic simulations will not respond strongly to a very large number of inputs, O'Hagan (2006) asserts that "GP emulation can be implemented effectively with up to 50 inputs on modern computing platforms."

3.3 Parameter estimation

Before applying the Gaussian process model for prediction, values for the d parameters $\boldsymbol{\xi}$ and q parameters $\boldsymbol{\beta}$ must be specified, and if variance estimation is also of interest, then the process variance, λ, must be estimated as well. There are several methods used in practice to estimate good values of the parameters governing the GP. These can range from intuitive approaches like cross validation (CV), to a very complicated full Bayesian analysis that accounts for the uncertainty in the parameters being estimated (see, for example, Paulo, 2005; Rasmussen, 1996).

This work focuses on maximum likelihood estimation (MLE). Bayesian parameter esti-

mation tends to be most appropriate when the amount of training data is small and does not sufficiently characterize the function being approximated. However, when GP interpolation is used to approximate deterministic computer simulations, Bayesian parameter estimation may be prohibitively expensive (or at the very least very computationally demanding), and the additional uncertainty that is being addressed may be insignificant. Cross-validation provides an intuitive approach to parameter estimation, but it has been shown to perform worse than MLE for the approximation of deterministic computer simulations (Mardia and Marshall, 1984), and cross-validation does not allow one to estimate the process variance, which is needed to quantify the uncertainty in the GP interpolation.

3.3.1 Formulation of the MLE optimization problem

Maximum likelihood estimation provides a natural approach to estimating the parameters that govern the mean and covariance functions of the Gaussian process model. Based on the Gaussian assumption, the observed training values represent a realization of a multivariate normal distribution. The basic idea of MLE is to find the particular mean vector and covariance matrix that define the most likely multivariate normal distribution to result in the observed data. Given a particular parametrization of the mean and covariance functions, the problem is reduced to estimating the governing parameters of these functions.

The likelihood function is simply given by the joint PDF of the observed responses, as in Eq. (3.3). Recall that the p-dimensional multivariate normal PDF with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is

$$f(\boldsymbol{y}) = (2\pi)^{-p/2} \left|\boldsymbol{\Sigma}\right|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right]. \tag{3.13}$$

For computational reasons, it is easier to work with the log of the likelihood when performing maximum likelihood estimation. Taking the log gives

$$\log f(\boldsymbol{y}) = -\frac{p}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}). \tag{3.14}$$

By substituting the expressions for the mean and covariance for the Gaussian process, one obtains the following expression for the log likelihood:

$$\log f(\boldsymbol{Y} \mid \boldsymbol{x}, \lambda, \boldsymbol{\xi}, \boldsymbol{\beta}) = -\frac{m}{2}\log 2\pi - \frac{1}{2}\log\left[(\lambda)^m|\boldsymbol{R}|\right] - \frac{1}{2\lambda}(\boldsymbol{Y}-\boldsymbol{F\beta})^T\boldsymbol{R}^{-1}(\boldsymbol{Y}-\boldsymbol{F\beta}). \tag{3.15}$$

Because most optimization routines minimize the objective function, it is common to work with the negative of the log likelihood. In addition, the multiplicative constant $1/2$ and the additive constant $m\log 2\pi$ can be dropped because they do not affect the optimization. Let the resulting modified negative log likelihood function be defined as NL:

$$NL = m\log\lambda + \log|\boldsymbol{R}| + \frac{1}{\lambda}(\boldsymbol{Y}-\boldsymbol{F\beta})^T\boldsymbol{R}^{-1}(\boldsymbol{Y}-\boldsymbol{F\beta}). \tag{3.16}$$

The optimization problem can now be formulated as

$$\begin{aligned} &\min_{\lambda,\boldsymbol{\xi},\boldsymbol{\beta}} \; NL \\ &\text{s.t.} \quad \lambda > 0,\ \xi_i > 0 \ \ \forall i \end{aligned} \tag{3.17}$$

The constraints on the ξ_i ensure that the correlation matrix $\boldsymbol{R}$ remains positive semi-definite.

While one could solve the MLE problem as posed in (3.17), there is one other change that is commonly made to simplify the process. That is, in order to allow for the use of unconstrained

optimization routines, the variables $\boldsymbol{\xi}$ can be transformed into a space that is not bounded (it will be evident shortly why this does not also have to be done for the process variance). The most straightforward transformation to work with that accomplishes this is the logarithmic one. Thus, define a new set of variables $\boldsymbol{\omega} = \log \boldsymbol{\xi}$, where the log is taken element-wise. With this transformation, the correlation function of Eq. (3.9) becomes

$$c(\boldsymbol{x}, \boldsymbol{x}^*) = \exp\left[-\sum_{i=1}^{d} e^{\omega_i} (x_i - x_i^*)^2 \right]. \tag{3.18}$$

Thus, the new optimization problem to be solved is

$$\begin{aligned} &\min_{\lambda, \boldsymbol{\omega}, \boldsymbol{\beta}} \; NL \\ &\text{s.t.} \quad \lambda > 0 \end{aligned} \tag{3.19}$$

Either of the optimization problems (3.17) or (3.19) could be attacked using a constrained, multi-dimensional optimization routine that does not make use of gradients. However, for this particular problem, the gradients can be easily expressed analytically, allowing for a more powerful optimization routine to be used. Also, as will be seen below, making use of the gradients will allow the optimization routine to work with $\boldsymbol{\omega}$ only, so that no transformation of λ is needed and unconstrained optimization algorithms can be applied.

3.3.2 Gradients of the negative log likelihood

The above optimization problem can be computationally expensive to solve, particularly when the number of training points, m, is large, because each evaluation of NL requires the inversion of the $m \times m$ correlation matrix $\boldsymbol{R}$. Further, there are some cases in which the maximum likelihood problem must be solved many times (see, for example, the iterative point selection

procedure presented in Section 3.4). Fortunately, the gradients of NL with respect to the design variables are available analytically, allowing for much more efficient gradient-based optimization routines to be used.

First, consider the derivatives of NL with respect to each of the ω_i. Using linear algebra, it can be shown that

$$\frac{\partial NL}{\partial \omega_k} = \text{trace}(\boldsymbol{R}^{-1}\dot{\boldsymbol{R}}_k) - \frac{1}{\lambda}(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta})^T \boldsymbol{R}^{-1}\dot{\boldsymbol{R}}_k \boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta}), \tag{3.20}$$

where $\dot{\boldsymbol{R}}_k = \dfrac{\partial \boldsymbol{R}}{\partial \omega_k}$, which is the derivative of the correlation matrix. This derivative is found by differentiating Eq. (3.18):

$$\dot{\boldsymbol{R}}_k = \frac{\partial \boldsymbol{R}}{\partial \omega_k} = \left[\frac{\partial c\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right)}{\partial \omega_k}\right]_{i,j} = \left[-e^{\omega_k}\left(x_k^{(i)} - x_k^{(j)}\right)^2 c\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right)\right]_{i,j}. \tag{3.21}$$

Next, consider the gradient of the objective function with respect to the process variance, λ. With straightforward math, it can be shown that

$$\frac{\partial NL}{\partial \lambda} = \frac{m}{\lambda} - \frac{1}{(\lambda)^2}(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta})^T \boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta}). \tag{3.22}$$

Notice that one can solve for the minimizer $\hat{\lambda}$ in terms of $\boldsymbol{\omega}$ and $\boldsymbol{\beta}$ by setting Eq. (3.22) equal to 0. This manipulation shows that conditional on the other parameters, the optimal value of λ is

$$\hat{\lambda} = \frac{1}{m}(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta})^T \boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta}). \tag{3.23}$$

This makes it possible to remove λ from the numerical optimization algorithm. For every iteration on $\boldsymbol{\omega}$, the exact optimum $\hat{\lambda}$ can be found immediately using Eq. (3.23). Note that

Eq. (3.23) can be rearranged and substituted into Eq. (3.16) to simplify NL as follows:

$$NL = m \log \hat{\lambda} + \log |\boldsymbol{R}| + m. \tag{3.24}$$

The additive constant, m, can of course be dropped. Given that λ is being updated at each optimization iteration using Eq. (3.23), one can maximize the likelihood function by minimizing Eq. (3.24). Keep in mind though, that if Eq. (3.23) is not being used, the more general expression for NL given by Eq. (3.16) must be minimized instead.

Finally, consider the gradient of the objective function with respect to $\boldsymbol{\beta}$. It is possible to take this derivative with respect to the entire vector $\boldsymbol{\beta}$, using matrix calculus:

$$\frac{\partial NL}{\partial \boldsymbol{\beta}} = -\frac{2}{\lambda}(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta})^T \boldsymbol{R}^{-1} \boldsymbol{F}. \tag{3.25}$$

As before, it is possible to set this gradient equal to 0 and find the analytical optimum with respect to the other parameters. This yields

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{F}^T \boldsymbol{R}^{-1} \boldsymbol{F}\right)^{-1} \boldsymbol{F}^T \boldsymbol{R}^{-1} \boldsymbol{Y}, \tag{3.26}$$

which is also the generalized least squares estimate for $\boldsymbol{\beta}$. Notice that the optimum value of $\boldsymbol{\beta}$ is a function of $\boldsymbol{\omega}$ and not of λ. Thus, for a given vector $\boldsymbol{\omega}$, one first calculates $\hat{\boldsymbol{\beta}}$ using Eq. (3.26) and then uses this value to calculate $\hat{\lambda}$ using Eq. (3.23).

3.3.3 Computational considerations

The majority of the computational burden that arises when dealing with Gaussian processes is the inversion of the correlation matrix, $\boldsymbol{R}$, which is an order m^3 operation. Thus, the computa-

tion time is very sensitive to the number of training points being used. Luckily, there are very efficient methods for dealing with this inverse.

Since the correlation matrix is symmetric, the Cholesky decomposition can be employed to handle computations involving $\boldsymbol{R}^{-1}$. Cholesky decomposition is about a factor of 2 times faster than other methods that handle matrix inversion, and it can be used very efficiently when the inverse of the matrix isn't explicitly needed. Mathematically, the Cholesky decomposition of a symmetric matrix $\boldsymbol{R}$ is the lower triangular matrix $\boldsymbol{L}$ such that $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{R}$. The "matrix left division" $\boldsymbol{R}^{-1}\boldsymbol{b}$ can then be computed efficiently using back-substitution, without explicitly computing $\boldsymbol{R}^{-1}$.

While the objective function NL and its gradients require multiple matrix left divisions, most of them involve the same term: $\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta})$. In fact, for any one particular iteration of the optimization algorithm having the same value for $\boldsymbol{\omega}$, the Cholesky decomposition of $\boldsymbol{R}$ only needs to be computed once. It can then be re-used to compute the necessary matrix left divisions.

This same concept should be used when making multiple predictions based on the same Gaussian process model. If a fixed set of correlation parameters are being used for each prediction, the matrix $\boldsymbol{R}$ will not change, and its Cholesky decomposition should only be computed once. The result is that new predictions can be computed very efficiently.

The second major computational issue is the conditioning of the correlation matrix $\boldsymbol{R}$. This matrix can become ill-conditioned when there are large correlations among the training data. Such large correlations are more likely when the dimensionality of the input (d) is small, the underlying functional relationship is simple, or the number of training points (m) is large. A large correlation value between two training points means that these points are "close" together

in the parameter space (this closeness is scaled by the correlation parameters, $\boldsymbol{\xi}$), which is an indication that the points are providing the model with redundant information.

While the inclusion of redundant training data will result in an ill-conditioned correlation matrix for most "reasonable" choices of $\boldsymbol{\xi}$ (in which case one is encouraged to consider the greedy point selection algorithm presented in Section 3.4), it is also possible that the numerical optimizer will encounter values of $\boldsymbol{\xi}$ that result in especially large correlations, even for "well-conditioned" training data. In some cases, the optimizer will encounter a value for $\boldsymbol{\xi}$ that results in a correlation matrix $\boldsymbol{R}$ that is singular to working precision and can not be inverted by the computer. In such a case, one will not be able to compute an objective function value or gradient. To surmount this problem, the objective function can be programmed such that when the correlation matrix is singular to working precision (i.e. when the Cholesky decomposition fails), a "bad" (large) objective function value is returned (large negative gradients for $\boldsymbol{\xi}$ can also be returned in order to encourage larger values of $\boldsymbol{\xi}$, i.e. shorter correlation lengths).

Another good practice is to normalize all of the inputs so that each input variable has the same scale (one possibility is to normalize each component of $\boldsymbol{x}$ to have zero mean and unit variance). Such a normalization is a pre-processing step that is done before performing the parameter estimation. Then, once a prediction is desired at an untested location $\boldsymbol{x}^*$, it will be necessary to apply the same transformation to $\boldsymbol{x}^*$. There are a couple reasons for doing this normalization step. First, making sure all of the input variables are on the same scale makes the computations more stable. Applying this transformation also results in the ξ_i having approximately the same scale as well, which is useful, among other things, for simplifying the choice of starting values. The resulting maximum likelihood estimates of each ξ_i also provide an indication of the sensitivity of the response to each input: a large value of ξ_i represents a

shorter correlation length in that dimension and a sensitivity of the response to x_i.

3.4 Multivariate output: time and space coordinates

In many cases, the computer simulation may output the response quantity of interest (e.g. temperature) at a large number of time instances and/or spatial locations. Such cases are sometimes termed multivariate output, because the response at each time or space instance can be thought of as a separate output variable.

Unfortunately, though, this introduces a considerable amount of additional complexity when the Gaussian process is used to model the code output. The simplest solution is probably to use a small number of features to represent the entire output. However, in many cases one would like to take account of the entire output spectrum, in order to ensure agreement to the experimental data at all output locations.

If the dimensionality of the output spectrum is small (say, four or five outputs), one might consider building a separate, independent Gaussian process model for each output quantity. However, this approach becomes far too cumbersome when there are many time and/or space instances to consider. When a large output spectrum is of interest, one possible approach is to treat those variables that index the output spectrum (e.g. time, location) as additional inputs to the surrogate. In this way, only one surrogate is needed, and the output can be treated as a scalar quantity.

This approach, however, introduces its own difficulties. Consider a design of computer experiments based on 50 LHS samples for a computer simulation that outputs the response quantity at 1,000 time instances. When time is parametrized as an input, this gives a total of 50,000 training points for the Gaussian process model. This will make the MLE process virtually impossible, since it will require the repeated inversion of a $50,000 \times 50,000$ correlation

matrix. Further, if there is a significant degree of autocorrelation with time (which will almost certainly be the case, particularly if the code output uses small time intervals), this correlation matrix will be highly ill-conditioned, and likely singular to numerical precision.

There are several possible methods for dealing with these issues. One approach that has been used in the past is a decomposition of the correlation matrix that is applicable when the training data form a grid design (Bayarri et al., 2002; Kennedy and O'Hagan, 2000b). A grid design will occur, for example, if the simulator output is a function of time, and each simulator run reports the response at the same time instants. The inverse of the correlation matrix is then computed based on a Kronecker product, so instead of inverting a $50,000 \times 50,000$ matrix, two matrices are inverted, one of size 50×50 and one of size $1,000 \times 1,000$. However, this method is fairly complicated to implement, and it does not do anything to improve the conditioning of the full correlation matrix.

Most other solutions are based on the omission of a subset of the available points. Since the response is most likely strongly autocorrelated in time, many of the points are redundant anyway. The difficulty, though, is how to decide which points to throw away. Considering again the above example, even if the number of time instances is reduced from 1,000 to 20, there are still 1,000 training points (20 time instances $\times$ 50 LHS samples) for the Gaussian process, which may still be unnecessarily large.

In the following, an algorithm is presented that obviates the need for a subjective selection of a training point subset. The algorithm presented below is based on the "greedy algorithm" concept. The basic idea of a greedy algorithm is to follow a problem solving procedure such that the locally optimal choice is made at each step (Cormen et al., 2001). This concept is applied below to the problem of choosing among available surrogate model training points

by iteratively adding points one at a time, where the point added at each step is that point corresponding to the largest prediction error. This approach has several advantages:

1. The point selection technique is easier to implement than the Kronecker product factorization of the correlation matrix.

2. It is not restricted to maintaining the grid design. That is, the algorithm may choose a subset of points such that code run 1 may be represented at time instance 1, but code run 2 may not get represented at time instance 1. Further, a non-uniform time spacing may be selected: perhaps there is more "activity" in the early time portion, so more points are chosen in that region.

3. The amount of subjectivity associated with choosing which points to retain is strongly reduced. Instead of deciding on a new grid spacing, one can instead specify a desired total sample size or maximum error.

4. The one-at-a-time process of adding points to the model makes it readily apparent precisely when numerical matrix singularity issues begin to come into play (if at all). This is particularly useful for very large data sets containing redundant information.

The greedy point selection approach is outlined below. Denote the total number of available points by m_t, the set containing the selected points by Θ, the set containing the points not yet selected by Ω, and the size of Θ by m. Also, denote the maximum allowed number of points as m^*, the desired cross-validation prediction error by δ^*, and the current vector of cross-validation prediction errors by $\boldsymbol{\delta}$.

1. Generate a very small (~ 5) initial subset Θ. Ideally, this is chosen randomly, since the original set of points is most likely structured.

2. Use MLE to compute the Gaussian process model parameters associated with the points in Θ.

3. Repeat until $m \geq m^*$ or $\max(\boldsymbol{\delta}) \leq \delta^*$:

   (a) Use the Gaussian process model built with the points in Θ to predict the $m_t - m$ points in the set Ω. Store the absolute values of these prediction errors in the vector $\boldsymbol{\delta}$.

   (b) Transfer the point with the maximum prediction error from Ω to Θ.

   (c) For the current subset Θ, estimate the Gaussian process model parameters using MLE.

As an example, consider a Gaussian process model for the two dimensional Rosenbrock function[3],

$$f(x_1, x_2) = (1 - x_1)^2 + 100\left(x_2 - x_1^2\right)^2,$$

on the usual bounds $-2 \leq x_1 \leq 2$, $-2 \leq x_2 \leq 2$. A set of 10,000 points are randomly generated within these bounds, and the "greedy" point selection algorithm is used to choose a subset of $m = 35$. The resulting maximum prediction error is 1.58×10^{-2}, with a median prediction error of 2.87×10^{-3}. The 5 random initial points, along with the remaining selected points are plotted in Figure 3.1. The convergence of the maximum prediction error is plotted with a semi-log scale in Figure 3.2.

[3]This particular function is used here as an example primarily because it is commonly used as a test function, particularly within the optimization community, and so it should be familiar to most readers. However, some readers will note that this is not an especially suitable function for the stationary Gaussian process model: this is true. Specifically, this function contains both large- and small-scale variations, which are not necessarily appropriately modeled using a stationary covariance formulation. However, the function is used here in spite of these difficulties, in part to illustrate the robustness of the Gaussian process interpolation framework, and also to emphasize that when used appropriately, this interpolation scheme is capable of representing a wide variety of functional forms.

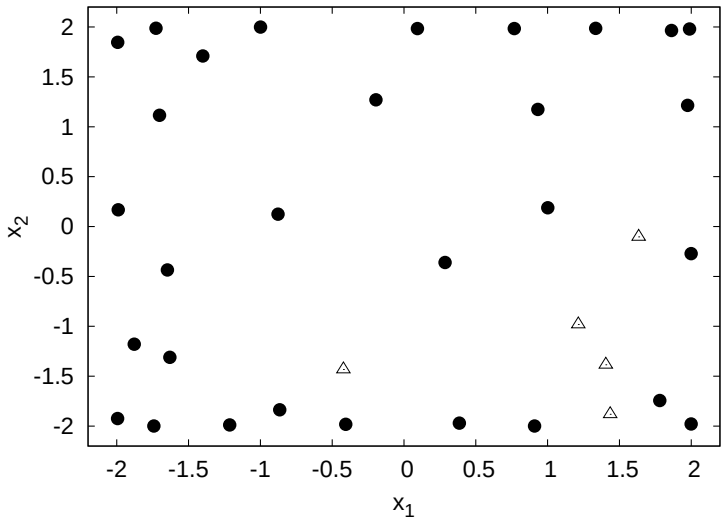


Figure 3.1: Initial (triangles) and selected (circles) points chosen by the "greedy" algorithm with Rosenbrock's function.

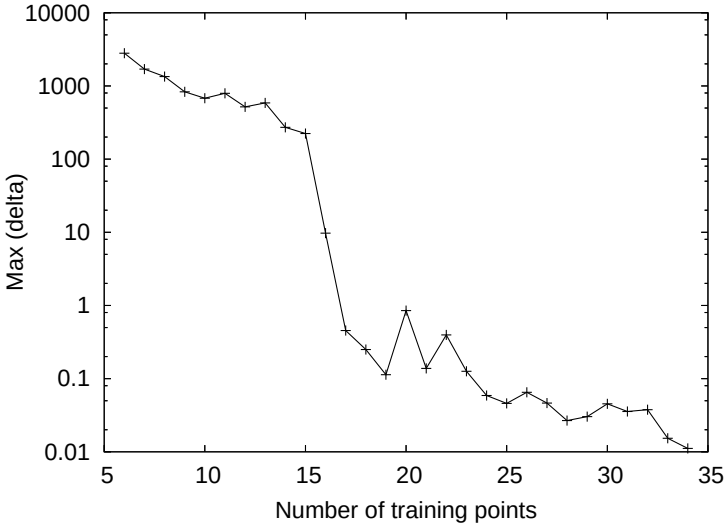


Figure 3.2: Semi-log plot of maximum prediction error versus m for Rosenbrock's function.

This example clearly shows the power of Gaussian process modeling for data interpolation. From Figure 3.1, it is obvious that the point selection algorithm tends to pick points on the boundary of the original set. This is expected, and is because the Gaussian process model needs these points in order to maintain accuracy over the entire region. Only a relatively small number of points are needed at the interior because of the interpolative accuracy of the model.

It is also interesting to note that the decrease in maximum prediction error is not strictly monotonic. Adding some points may actually worsen the predictive capability of the Gaussian process model in other regions of the parameter space. Nevertheless, until matrix ill-conditioning issues begin to take effect, the overall trend should still show a decrease in maximum prediction error.

3.5 Accounting for observation uncertainty

So far, the Gaussian process models that have been discussed assume that the response values in the training data are observed without error or uncertainty. The result is that the predictions obtained using the models discussed above are forced to exactly interpolate the training points. However, there may be cases when the training data are observed with error and/or uncertainty, and one would prefer the predictions to "smooth" the observed data as opposed to fitting them exactly. For example, one might want to use a Gaussian process to model the bias between simulator predictions and experimental observations. While the simulator predictions are deterministic, the experimental observations are undoubtedly subject to variability.

Accounting for observation errors can be done in a straightforward manner using the Gaussian process model.[4] If the m training points are noise-corrupted observations of the true

[4]The following can be viewed as a Bayesian treatment, in which one begins with a Gaussian process prior distribution for an unknown function, and the likelihood for each training observation is a normal distribution with variance $\lambda_{exp,i}$ (see Rasmussen, 1996, for a rigorous derivation).

underlying process, then the covariance matrix associated with the training points can be written as $\boldsymbol{Q} \equiv \lambda \boldsymbol{R} + \boldsymbol{\Sigma}_{exp}$, where $\boldsymbol{\Sigma}_{exp}$ is the covariance matrix that characterizes the observation noise. If the observations are independent, then $\boldsymbol{\Sigma}_{exp} = \mathrm{diag}(\lambda_{exp,1}, \ldots, \lambda_{exp,m})$, where $\lambda_{exp,i}$ is the variance associated with the observation error for the i^{th} observation. The observation variances may be assumed known, but it would also be possible to estimate them based on the training data using maximum likelihood estimation (which would only be meaningful if repeated observations are present).

The preceding sections used a special notation in which the data covariance matrix is decomposed as $\lambda \boldsymbol{R}$, the product of the process variance and the data correlation matrix. However, as is apparent from the above definition of the new data covariance matrix, this separation is no longer possible. Previously, $\boldsymbol{r}(\boldsymbol{x}^*)$ was defined as the vector of correlations between $\boldsymbol{x}^*$ and the training points, but $\boldsymbol{k}(\boldsymbol{x}^*)$ is now used instead, which is defined as the vector of *covariances* between $\boldsymbol{x}^*$ and the training points. Note that $\boldsymbol{k}(\boldsymbol{x}^*)$ is not a function of $\boldsymbol{\Sigma}_{exp}$.

Using the new notation, the conditional distribution of a point $\boldsymbol{x}^*$ can be expressed by the mean and variance (previously Eqs. (3.4) and (3.5)):

$$\mathrm{E}\big[Y(\boldsymbol{x}^*) \mid \boldsymbol{Y}\big] = \boldsymbol{f}^T(\boldsymbol{x}^*)\boldsymbol{\beta} + \boldsymbol{k}^T(\boldsymbol{x}^*)\boldsymbol{Q}^{-1}(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta}) \tag{3.27}$$

and

$$\mathrm{Var}\big[Y(\boldsymbol{x}^*) \mid \boldsymbol{Y}\big] = \lambda - \boldsymbol{k}^T\boldsymbol{Q}^{-1}\boldsymbol{k}. \tag{3.28}$$

Note that when the experimental variances are zero, these equations are equivalent to Eqs. (3.4) and (3.5). Also, even if the λ_{exp} are large, the variance predicted by Eq. (3.28) at or near one of the training points still has an upper bound of λ, even though the uncertainty associated

with that observation, $\lambda_{exp,i}$, may be greater than λ. This reinforces the importance that the parameter selection process plays for the predictions and their uncertainty estimates.

3.5.1 Parameter estimation

Maximum Likelihood Estimation (as discussed in Section3.3) can be used for this model also, but some additional considerations come into play. First, consider the negative log of the likelihood function, as before. The likelihood function is again derived from Eq. (3.14), but the multiplicative constant $1/2$ will be retained here so that the likelihood can be combined with a prior distribution for the Gaussian process parameters, if desired. The negative log likelihood is

$$NL = \frac{1}{2}\log|\boldsymbol{Q}| + \frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta}\right)^T \boldsymbol{Q}^{-1}\left(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta}\right). \tag{3.29}$$

Recall from Section 3.3 that by taking the derivative of NL with respect to λ, it was possible to find its conditional optimum value. However, now that λ can no longer be separated from the data covariance matrix, that result no longer applies. Moreover, a bigger problem has arisen: with non-zero experimental variances λ_{exp}, the optimum value of λ may tend to zero, which is obviously infeasible and of no practical use. There are two possibilities for dealing with this problem:

1. One could include a prior distribution for λ that naturally counteracts the insistence of the likelihood for λ to go to zero. Thus, instead of searching for values that maximize the likelihood function, one would search for parameter values that maximize the posterior distribution (this is sometimes referred to as maximum *a posteriori* (MAP) estimation).

2. Alternatively, one could work with a "penalized" or "restricted" likelihood function. In this case, an additional term is simply added to NL.

In practice, these two alternatives are essentially the same. However, some analysts may be reluctant to include a prior distribution for λ because of the apparent subjectivity involved with choosing an appropriate prior. Thus, restricted maximum likelihood estimation (RMLE) is presented here.

The RMLE method was first proposed by Patterson and Thompson (1971), and Harville (1974) later presented a more convenient representation. The motivation behind the development of RMLE appears to be the fact that when the covariance parameters are chosen based on regular MLE, their maximum likelihood estimates take no account of the loss in degrees of freedom that results from estimating $\boldsymbol{\beta}$. The idea is based on re-formulating the likelihood as a function of error contrasts, where an error contrast is simply any linear combination $\boldsymbol{b}^T\boldsymbol{Y}$ of the observations such that $\mathrm{E}\left[\boldsymbol{b}^T\boldsymbol{Y}\right] = 0$. The technique is thus based on maximizing the likelihood function associated with a particular set of $m - q$ linearly independent error contrasts, rather than the full likelihood function.

The resulting likelihood function, which will be denoted by NLR, is

$$NLR = \frac{1}{2}\log|\boldsymbol{Q}| + \frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{F}\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{Q}^{-1}\left(\boldsymbol{Y} - \boldsymbol{F}\hat{\boldsymbol{\beta}}\right) + \frac{1}{2}\log\left|\boldsymbol{F}^T\boldsymbol{Q}^{-1}\boldsymbol{F}\right|, \tag{3.30}$$

where $\hat{\boldsymbol{\beta}}$ is the same as Eq. (3.26), but with $\boldsymbol{R}$ replaced by $\boldsymbol{Q}$. The only differences from Eq. (3.29) are the additional term at the end and the replacement of $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$. The use of $\hat{\boldsymbol{\beta}}$ directly inside the likelihood function does not add anything new, however, since $\hat{\boldsymbol{\beta}}$ would have been chosen as the optimal value anyway. The additional term in Eq. (3.30) will effectively prevent the optimum value of λ from being zero by penalizing small values of λ. Further, the use of the restricted likelihood function takes appropriate account of the fact that q degrees of freedom are lost in the estimation of $\boldsymbol{\beta}$.

3.5.2 Gradient information

As before, the gradients of the negative log likelihood can be made available to the optimization algorithm being used to significantly improve the performance. The gradients of NLR differ from those presented in Section 3.3.2 because of the inclusion of λ_{exp} and because of the additional term in the likelihood function. The derivations are based on matrix calculus, and the relevant equations are given below.

The gradient of NLR with respect to any covariance parameter, θ, is given by

$$\frac{\partial NLR}{\partial \theta} = \frac{1}{2}\text{trace}\left(\boldsymbol{Q}^{-1}\dot{\boldsymbol{Q}}\right) - \frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{F}\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{Q}^{-1}\dot{\boldsymbol{Q}}\boldsymbol{Q}^{-1}\left(\boldsymbol{Y} - \boldsymbol{F}\hat{\boldsymbol{\beta}}\right) - \frac{1}{2}\text{trace}\left[\boldsymbol{F}\left(\boldsymbol{F}^T\boldsymbol{Q}^{-1}\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T\boldsymbol{Q}^{-1}\dot{\boldsymbol{Q}}\boldsymbol{Q}^{-1}\right], \quad (3.31)$$

where $\dot{\boldsymbol{Q}}$ is the matrix of derivatives of $\boldsymbol{Q}$ with respect to θ. For the log correlation scale parameter, ω, the matrix of derivatives is

$$\frac{\partial \boldsymbol{Q}}{\partial \omega_k} = \left[-e^{\omega_k}\left(x_k^{(i)} - x_k^{(j)}\right)^2 \lambda c\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right)\right]_{i,j}. \quad (3.32)$$

When dealing with a covariance matrix that can not be decomposed into a variance term and a correlation matrix, it makes sense to work with the log of λ, since it will need to be optimized numerically. Defining $\gamma = \log(\lambda)$ gives

$$\frac{\partial \boldsymbol{Q}}{\partial \gamma} = \left[e^{\gamma} c\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right)\right]_{i,j}. \quad (3.33)$$

3.5.3 Computational considerations

As mentioned above, one result of the inclusion of measurement uncertainties in the Gaussian process model is that an analytical optimum value of the process variance, λ, is no longer available. Thus, unlike before, it becomes necessary to choose a starting value for λ (or preferably the log of λ). The importance of the starting value is that a bad choice can lead to numerical problems with the likelihood computations, which will in turn cause trouble for the numerical optimization algorithm. This may happen if one attempts to compute NLR with a value of λ that is largely inconsistent with the scale of the observed response values.

One possibility is to set the initial value of λ equal to the variance of the observed response values, which will generally be on the same scale as the process variance. An alternative approach is to first scale the data in $\boldsymbol{Y}$ to have unit variance, in which case unity is an appropriate starting value for λ.

Several steps must be taken if the response values are to be scaled. Consider that one simply wants to transform the original response values $\boldsymbol{Y}$ to have unit variance. Denote the sample variance of the observed response values as s_Y^2. The transformation is effected by dividing each value in $\boldsymbol{Y}$ by s_Y. The observation variances, $\lambda_{exp,i}$, must also be scaled accordingly, by dividing each by s_Y^2. Finally, it will be necessary to rescale the conditional mean and variance after applying Eqs. (3.27) and (3.28). This is done by simply multiplying the conditional expected value by s_Y and the conditional variance by s_Y^2.

3.6 Summary

Gaussian process interpolation is a powerful tool that can be used to develop inexpensive approximations to computer simulations. This chapter provides a comprehensive overview from

an engineering standpoint of the capabilities and computational implementation considerations associated with using Gaussian process interpolation as a surrogate for expensive computer simulations.

The majority of the concepts discussed in this chapter do not constitute original research. The exception is the iterative point selection algorithm presented in Section 3.4. This point selection approach is proposed here as a tool to streamline the GP model by removing redundant data, and it is envisioned to be especially useful when the GP surrogate is used to model a response that is a function of time. The point selection approach is applied for the case study of Section 6.4.

CHAPTER IV

MODEL VALIDATION AND UNCERTAINTY PROPAGATION

How much confidence do I have in my model? To what degree do the outcomes predicted by my model agree with outcomes observed in the laboratory or in the field? Is my model suitable for its purpose? These are questions that are addressed via the process known as *model validation*. Model validation involves comparing model predictions with observed outcomes in order to establish some amount of confidence in the model's predictive capability, with respect to its intended use.

Generally speaking, model assessment is not a new concept at all, and in fact, statisticians have long been concerned with developing quantitative assessments of the accuracy of their models (consider the coefficient of determination, the fraction of variance explained, significance testing of regression coefficients, etc.) However, the very nature of statistical models is that they are based on *data*; they are intended to model the relationship that is empirically observed among physical quantities, and without data, there is no model. On the other hand, computational simulations are usually grounded in physics. Simulations are computational implementations of the conceptual models that we construct to model the behavior of the physical world. As such, there is no requirement that the simulation be based on, or even informed by actual data. Even having a sparse set of observed system response values with which to compare against simulation predictions is not a guarantee. Thus, the nature of model assessment for *computer simulations* is not nearly as straightforward as that for statistical models, and is in fact a much more difficult problem, requiring substantial consideration: it is this particular

type of model assessment that is most commonly referred to as *model validation*.

Significant research efforts have been undertaken to develop unified and universally applicable procedures for model validation, but as should be apparent after considering the case studies presented in Chapter VI, the appropriate approach to model validation assessment seems to be highly situation-dependent. In particular, the intended use of the model should naturally play a large role in guiding its validation assessment. In addition, an important question to ask is: What factors are driving the uncertainty in the model's validity? This question encourages the analyst to consider, among other things, the nature of the experimental data, such as whether or not experiments are conducted for multiple different system configurations, whether or not instrumentation error is significant, how accurately the experimental conditions (such as boundary conditions) are known, and the degree of variability in the observed system response.

Some background and theory regarding model validation is presented in Section 4.1, but model validation seems to be best studied from an applied perspective. It is hoped that the variety of case studies given in Chapter VI will shed light on the different types of model validation scenarios that one might encounter, and approaches that are appropriate for each. The case study of Section 6.1 provides perhaps the most rigorous validation assessment, via the use of statistical significance testing and power analysis. The corresponding testing theory is presented here in Section 4.1.2. Of the available validation approaches, the most effort from a theoretical development perspective is devoted here to the significance testing approach. This is not a suggestion that significance testing constitutes a superior means of quantitative validation assessment; instead, it is felt that while significance testing itself is well-established, the tool is often applied incorrectly or haphazardly for validation assessment. As such, Section 4.1.2

will emphasize the appropriate use of significance testing for model validation assessment.

The second section of this chapter, Section 4.2, discusses uncertainty propagation. Uncertainty propagation is included here because it is often a pre-requisite to validation assessment: when model inputs are uncertain or random variables, model predictions should be compared to observations in light of the model output uncertainty that is implied by the input uncertainty. Section 4.2 discusses appropriate approaches for estimating the probability distribution of the computer simulation output that is implied by the probability distributions associated with the simulation inputs. In particular, the non-parametric probability density estimation tool known as kernel density estimation is discussed in Section 4.2.2, and one approach to dimensionality reduction, principal component analysis, is presented in Section 4.2.3.

4.1 Model validation

4.1.1 Background

Model validation makes up one part of a type of quality assurance process for computational models referred to as Verification and Validation, or simply "V&V." The field itself is broad, and deals with such topics as mesh convergence and discretization error, the design of experiments, and methods for comparing predictions and experimental observations. Validation explores the degree to which the predictive capability of the model is suitable for a particular purpose, whereas verification addresses whether or not the conceptual model is implemented correctly in the form of a computer program. Additional information on model verification can be found in Roache (1998); Knupp (2002); Rebba et al. (2006); Oberkampf and Trucano (2002); AIAA (1998); ASME (2006); Balci (1997); Sargent (2004).

Professional societies and standards committees have played an important role in guiding development in the field of Verification and Validation; see, for example, guides published by

AIAA (1998); ASME (2006); ANS (1987); ISO (1991). One of the first challenges in V&V was to pinpoint precisely what is meant by verification and validation. Although several formalisms have since been put forth, some of the earliest and most widely regarded definitions are those originally published by Schlesinger (1979) in connection with the Society for Computer Simulation. For example, model validation is defined as

> Substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model.

Several aspects of this definition are important. First, validation is concerned only with whether or not the model has a satisfactory range of accuracy with regards to an intended application. Thus, the process of validation only has meaning once the intended application is specified. This entails the characterization both of a domain of applicability for the model, as well as its intended use. For instance, if the intended use of a model is solely to predict acceleration response, then validation is not concerned with the accuracy of stress predictions. Further, the phrase "domain of applicability" in the definition reminds that validation is not concerned with the predictive capability of the model over all possible boundary, loading, initial conditions, etc., but that the predictive capability should only be assessed for a particular domain of interest.

Most of the previous work in the V&V field has dealt with outlining general frameworks and methodologies. Consequently, validation can be divided into several sub-fields, such as the design of the validation experiments, the design of the computer experiments, uncertainty quantification, and validation (or comparison) metrics. An in-depth V&V overview is given by Oberkampf and Trucano (2002), which discusses, in addition to code verification, all of

the validation steps mentioned above. Other, more conceptual, reviews are given by Sargent (2004) and Balci (1997).

Work dealing with the development and application of generally applicable quantitative validation metrics has been limited. In fact there has been significant evolution over time in terms of what are viewed as important features of a validation metric. For example, Oberkampf and Trucano (2002) state that "a useful validation metric should only measure the agreement between the computational results and the experimental data." The underlying philosophy that gave rise to this viewpoint was the emphasis that the *accuracy* and *adequacy* of a particular model are strictly separate issues. That is, accuracy is a measure of agreement between predictions and observations. The purpose of the accuracy metric is to support a decision based on the more important *adequacy* consideration, which reflects whether or not the model is suitable (i.e., adequate) for its intended use. Note that in later work, Oberkampf and Barone (2006) describe several desired features of a validation metric, arguing that such a metric should, among other things, depend on the number of experimental replications of a measurement quantity, in order to reflect a "level of confidence". Such a metric clearly does not measure just the accuracy of the model, as suggested in the previous work (Oberkampf and Trucano, 2002). This is merely one example of how the philosophy of model validation has evolved over time.

Some have studied model validation from the perspective of statistical hypothesis testing. Although it is not necessarily a popular approach within the validation community, hypothesis testing nevertheless provides a well-established foundation for quantifying considerations such as sample size, inherent variability, and type I/II errors. Recent work dealing with the use of hypothesis testing for validation assessment includes Paez and Urbina (2002); Hills and Leslie (2003); Dowding et al. (2004); Chen et al. (2004). The Bayesian perspective on hypothesis

testing has also been explored by Mahadevan and Rebba (2005); Rebba et al. (2006); Rebba and Mahadevan (2007). Jiang and Mahadevan (2007) even discuss a method for incorporating Bayesian hypothesis testing with risk considerations for the purpose of decision making.

A good overview of quantitative metrics is given by Rebba (2005), which discusses classical and Bayesian hypothesis testing, as well as other alternatives such as decision-theoretic approaches and model reliability.

An additional challenge when developing validation metrics arises when the response quantity of interest from the simulation is multivariate. This can occur if multiple different responses are of interest, such as stress and temperature, or if one response varies over time or space (although in such cases, the analyst should carefully consider whether or not a scalar "summary statistic" might capture all of the relevant information contained in a high-dimensional response).

It is well understood that statistical metrics that compare multiple responses simultaneously must be carefully developed so that dependencies among the various response measures are accounted for. It appears that Balci and Sargent (1982) were the first to apply multivariate statistical methods for model validation. Rebba and Mahadevan (2006) provide a detailed discussion of multivariate methods, including classical and Bayesian hypothesis testing for both distance and covariance similarity, as well as computational issues such as data transformations.

4.1.2 Significance testing

The use of significance testing (or hypothesis testing) is not widespread in the validation community, but the method nevertheless provides several features that might be deemed useful for a validation metric. Significance testing addresses in a rigorous sense whether or not a set of

observed data offer significant evidence against a certain hypothesis, in particular taking account of sample size and variability. Concepts such as the probabilities of type I and II errors also provide constructive ways of thinking about how the decision maker might approach the validation problem (see, for example, Jiang and Mahadevan, 2007; Balci and Sargent, 1981). Alternatively, the use of confidence intervals or regions is structurally analogous to that of significance testing, and is sometimes used instead for ease of interpretation (see Balci and Sargent, 1984, for an example of the use of confidence regions for multivariate validation inference).

When using a significance testing approach within the context of model validation or assessment, the analyst must take care to formulate the problem in an appropriate manner. First, consider the one-sample Student's t-test, which is one of the most straightforward and widely used significance tests available.

Consider that n independent observations of the random variable x are available, where $x \sim N(\mu, \sigma^2)$. If neither the mean or variance of x is known (the usual case), then the t-test can be employed to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$, where μ_0 is a hypothesized value (the manner in which these hypotheses may relate to model validation will be considered shortly). The test is based on the observed value of the so-called t-statistic, which is given by

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \tag{4.1}$$

where $\bar{x}$ is the sample mean of the observed data and s is the sample standard deviation of the data. The assumptions associated with this test are

1. The n observations $x_1, \ldots, x_n$ are independent of each other

2. The underlying population from which the observations are taken has a normal distribu-

tion

When the analyst suspects that the assumption of normality may not hold, distribution-free procedures are available as well (see Devore, 2000, for examples of distribution-free procedures and discussions of the trade-offs involved).

Based on the assumptions listed above, it is easy to show that the t-statistic of Eq. (4.1) has a Student's t-distribution, with $n-1$ degrees of freedom. Thus, H_0 is rejected in favor of H_1 at the significance level α when $|t| \geq t_{\alpha/2,n-1}$. At the usual significance level of $\alpha = 0.05$, there is a 5% probability of rejecting H_0 when it is in fact true (type I error).

Now return to the objective of model validation. In order to apply significance testing for validation, one must first formulate the problem in terms of a null and alternative hypothesis. A formulation to test the hypothesis that "the model is valid" against the alternative that "the model is not valid" would be desirable, but such hypotheses do not directly lend themselves to quantitative significance testing. In fact, whether or not the model is "valid" can only be measured in terms of subjective requirements in view of the model's intended use, and the significance testing results should be viewed only as quantitative guidance towards the decision problem of assessing model validity.

That said, the first step to formulate a significance test in support of validation assessment is to select the response quantity of interest. Presently scalar response quantities are considered, but multivariate response quantities are also discussed below. Because the results of the significance test will be driven exclusively by the response quantity of interest, its selection is important. Care should be taken to select a response quantity that is most representative of the model's intended use. For example, if the model is to be used to predict the maximum acceleration over time on a particular location of a structural dynamics component, then if pos-

sible, the significance test should be formulated in terms of maximum acceleration. In some cases (particularly in structural dynamics), one must resort to constructing low-dimensional summary *features* of a complicated response (typically an acceleration time history; for details on feature extraction, see Guratzsch, 2007; Farrar and Sohn, 2000; Sammon, 1969; Koontz and Fukunaga, 1972; Nigam, 1983).

Next, the nature of the validation data must be considered before the appropriate type of significance test can be determined. Significance testing is based on repeated, independent observations that come from a partially characterized probability distribution. In the context of model validation, the data on which the significance test is based may consist of:

1. Repeated experimental observations from nominally identical systems
2. Multiple realizations of the model output, in which model inputs describing the system configuration are held constant, but uncertain inputs are varied, perhaps according to a prescribed probability distribution.
3. Both experimental observations and model output realizations

For this discussion, the focus will be on Case 1. While in practice the simulation may often be expensive, there are a variety of tools that facilitate the characterization of the model output distribution when only a finite number of model evaluations are available. Such tools include Polynomial chaos expansion (c.f. Ghanem and Spanos, 1991) and other response surface approximation techniques, such as Gaussian process interpolation (Chapter III). Thus, for the sake of this discussion, it is assumed that the statistics of the model output distribution (in particular, the mean value) are fully characterized.

For case 1, the appropriate hypotheses for a significance test about the equality of location are $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$, where μ represents the mean of the response quantity of

interest corresponding to the population of experiments. Recall that μ is not known, because only a finite number of repeated experimental observation are available. On the other hand, μ_0 represents the mean of the population of model outputs. It is important to understand that μ_0 is a known constant in this formulation, because it is assumed that the model output distribution is fully characterized. As such, the variance of the model output distribution does not directly factor in for this case. (However, additional significance tests can also be done to compare the variance of the model output distribution to that of the experimental population, if such agreement is of interest to model validity).

The experimental data can now be used to construct the t-statistic of Eq. (4.1), and determine whether or not there is sufficient evidence to reject H_0 at a particular significance level. Note that the significance testing procedure is removed from model validation on several levels. First, the hypothesis H_0 does not necessarily correspond to ultimate model validity. Second, failure to reject H_0 is not equivalent to a confirmation that H_0 is true: in fact, failure to reject H_0 might be due entirely to having an insufficient number of observations (small n). For these reasons, significance testing results should be interpreted with care.

An additional note is that it may be the case that there is an insufficient number of repeated observations with which to establish a meaningful estimate of the variance, σ^2 (in particular, the sample estimate, s^2, can not be computed when $n = 1$). When this is the case, a value for σ^2 may be assumed, or the computer model being validated may be used to estimate σ^2. In particular, if model inputs that represent physical quantities that are random variables are characterized as such, then it may be perfectly reasonable to assume that the model output variance is equal to the variance of the experimental observations. In this case, one would set σ^2 equal to the variance of the model output distribution. Keep in mind that even when this is

done, σ^2 still represents the variability in the distribution of the experimental observations x.

However, when a value for σ^2 is assumed as opposed to being estimated from the n observations $x_1, \ldots, x_n$, the t-test on statistic (4.1) is no longer correct. When σ^2 is "known," as opposed to being estimated from the data, the z-test is used instead. This test is based on the statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}, \tag{4.2}$$

where $z \sim N(0, 1)$ under H_0. The difference between the t-test and the z-test is that the t-test accounts for the degrees of freedom that are lost in the estimation of σ with s, whereas these degrees of freedom are not lost when σ is not estimated from the data.

An additional step that can be taken is the quantification of the "power" of the test. Statistical power is the probability of rejecting H_0 when H_1 is true. Thus, computing the power can help determine whether or not the failure of rejection of hypothesis was due primarily to an insufficiently small sample size.

Power is computed by considering the distribution of the test statistic under the alternative hypothesis. The hypothesis $H_1 : \mu \neq \mu_0$ alone does not provide enough information to derive the distribution of the test statistic. In order to do so, the amount of deviation of μ from the hypothesized value μ_0 must be specified. Further, the true value of σ^2 must also be known to derive the exact distribution of the test statistic. This is unfortunate because σ^2 is not known in practice when applying the t-test (although it is known for the z-test). Thus, some value must be assumed, and a reasonable choice is the sample estimate, but a conservative (large) guess for σ^2 will yield a conservative (small) estimate of the power.

The distribution of the test statistic under the alternative hypothesis is a non-central t distribution with noncentrality parameter $\delta = \sqrt{n}(\mu - \mu_0)/\sigma$ (Srivastava, 2002). Thus, the power

of the significance test in rejecting H_0 can be found by computing the appropriate integral of this distribution:

$$P(|t| > t_{\alpha/2,n-1} \mid \delta). \tag{4.3}$$

An additional note is that model validity will generally not require that the model output mean is *exactly* equal to the experimental population mean, as specified by H_0. Typically, there will be an acceptable amount of error between the predictions and observations, such that H_0 might be formulated as $H_0 : |\mu - \mu_0| < \varepsilon$. The critical values for this significance test are based on the non-central t-distribution. However, such a test should be considered only if the data warrant a rejection of the corresponding point null hypothesis. The interval hypothesis formulation is discussed in more detail by Rebba and Mahadevan (2007) for both classical and Bayesian testing. From a practical standpoint, the interval formulation may be most useful when the sample size n is large, and one is concerned that rejection of a point null hypothesis may be due to a difference between μ and μ_0 that is not of practical concern for the validity of the model.

Multivariate testing

When the validity of the model depends on predicting multiple response quantities, then additional considerations come into play. In most cases the response quantities will have dependencies, and it is thus incorrect to apply univariate validation metrics (like Student's t-test) separately to each of the response measures. Incorrect conclusions could be reached because the univariate tests do not account for dependencies between the variables. In such cases, appropriate multivariate methods should be considered.

It appears that Balci and Sargent (1982) were the first to illustrate the use of multivari-

ate significance tests for the validation of models with multiple responses. They employed Hotelling's T^2 statistic (c.f. Srivastava, 2002), which is the multivariate analogue of Student's t-statistic. In analogy to the univariate case, the multivariate test has the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, where $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_0$ are now vectors of dimension p, which is the number of response quantities being compared. If a sample of n multivariate experimental observations is available, then the test is based on the statistic

$$\left(\frac{f-p-1}{fp}\right) T^2, \tag{4.4}$$

where Hotelling's T^2 statistic is given by

$$T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0), \tag{4.5}$$

$f = n - 1$, and $\boldsymbol{S}$ is the sample covariance matrix of the data. In analogy to the univariate t-test, the assumptions associated with this test are

1. The n observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent of each other

2. The underlying population has a multivariate normal distribution

Under the null hypothesis and the assumptions listed above, the test statistic of Eq. (4.4) has an F-distribution with $(p,\ f - p + 1)$ degrees of freedom. Note that T^2 can be viewed as the sample analogue of the Mahalanobis squared distance of the sample mean from the hypothesized value.

It was mentioned above that with the univariate test, σ^2 is sometimes taken to be equal to the model output distribution, as opposed to being estimated from the data, when there is an insufficient number of experimental observations n. The same concept applies here, except that

more observations are needed to estimate the covariance as the dimension of the data increases. From Eq. (4.4), it is apparent that the test statistic is only meaningful when $n \geq p + 2$. If this requirement is not met, one might take $\boldsymbol{\Sigma}$ to be equal to the covariance of the model output, given that the variability associated with the model inputs is characterized in a way that is meaningful in terms of the actual system behavior (e.g., the input distributions do not represent model parameter uncertainty). In analogy to the univariate case, when the covariance is not estimated from the observed data, the appropriate test statistic is

$$n\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0\right)^T \boldsymbol{\Sigma}^{-1} \left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0\right), \tag{4.6}$$

which has a chi-square distribution with p degrees of freedom under the null hypothesis.

As with the univariate t test, the power of rejecting the null hypothesis can also be computed for the multivariate T^2 test. The power is of course based on the amount of deviation of $\boldsymbol{\mu}$ from $\boldsymbol{\mu}_0$, but it is also based on the true covariance matrix, $\boldsymbol{\Sigma}$. As before, the true covariance will not be known, and the most logical guess is the sample covariance of the data. Under the alternative hypothesis, the test statistic of Eq. (4.1) has a noncentral F distribution, with noncentrality parameter $\delta^2 = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)$ (Srivastava, 2002). Thus, the power of the test for rejecting H_0 at significance level α is

$$P(F > F_{p,f-p+1,\alpha} \mid \delta^2). \tag{4.7}$$

When the covariance matrix is singular (not of full rank), it is not possible to compute the test statistics of Eqs. (4.4) and (4.5). In this case, one might consider a test based on the first k principal components (see Section 4.2.3), where k is the rank of the covariance

matrix (Srivastava, 2002). If the principal components are given by $\boldsymbol{A}_{(k)}^T\boldsymbol{x}$, then the test of $H_0:\ \boldsymbol{A}_{(k)}^T\boldsymbol{\mu} = \boldsymbol{A}_{(k)}^T\boldsymbol{\mu}_0$ is based on the statistic

$$\frac{f-k+1}{fk} n \left[\boldsymbol{A}_{(k)}^T(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\right]^T \left[\boldsymbol{A}_{(k)}^T\boldsymbol{S}\boldsymbol{A}_{(k)}\right]^{-1} \left[\boldsymbol{A}_{(k)}^T(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\right], \tag{4.8}$$

which has an F-distribution with k and $f-k+1$ degrees of freedom under H_0. If the covariance is not estimated from the data, then the test statistic becomes

$$n \left[\boldsymbol{A}_{(k)}^T(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\right]^T \left[\boldsymbol{A}_{(k)}^T\boldsymbol{\Sigma}\boldsymbol{A}_{(k)}\right]^{-1} \left[\boldsymbol{A}_{(k)}^T(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\right], \tag{4.9}$$

which has a chi-square distribution on k degrees of freedom under the null hypothesis.

4.2 Uncertainty propagation

4.2.1 Background

Uncertainty propagation is one of the most fundamental means of quantifying uncertainty in model predictions. Only the case in which the model input uncertainty is quantified through the use of probability density functions is considered here; other treatments include the Dempster-Shafer theory of evidence (Shafer, 1976) and possibility theory (Dubois and Prade, 2001; Zadeh, 1978). As considered here, the objective of uncertainty propagation is simple: estimate the probability density function for the model output(s) that is implied by the probability density functions for the model inputs.

Explicit, closed-form relationships between the inputs and outputs are rarely available when dealing with computer simulations (although see Haldar and Mahadevan, 2000, for appropriate approaches when the functional relationship is known). As such, this section is

geared towards uncertainty propagation when this relationship is only observable by running the simulation for a particular set of inputs.

The most straightforward approach for this case is known as Monte Carlo simulation. Monte Carlo simulation is a sampling based approach, in which a large number (typically tens of thousands) of random realizations of the input parameters are generated, and the simulator is run for each sample. The output samples are then used to make inference about the model output distribution, for example estimating the mean, estimating a reliability level, or plotting a histogram.

The problem with basic Monte Carlo sampling, however, is that it requires a very large number of evaluations of the computer simulation in order to accurately characterize the output distribution, and in practical applications, obtaining this number of evaluations is often not feasible. For this reason, there are several techniques available for reducing the variance in sampling-based estimators. Latin hypercube sampling (McKay et al., 1979) is a common approach that has the goal of attaining a more even distribution of the sample points in the parameter space. When reliability estimation is of interest, importance sampling (c.f. Haldar and Mahadevan, 2000) is popular. Importance sampling uses a sampling density that is concentrated in the failure region, so that samples aren't needlessly wasted in the other regions of the parameter space.

An alternative to efficient sampling techniques is to use an inexpensive approximation to the input/output relationship in lieu of the expensive computer simulation. Such approximations are often called surrogate models, or response surface approximations. A variety of methods are available for developing response surface approximations, including the development of models with reduced degrees of freedom, polynomial regression, multivariate adaptive

regression splines (Friedman, 1991), neural networks, non-intrusive polynomial chaos (Isukapalli et al., 1998), and Gaussian process interpolation. The use of Gaussian process interpolation for surrogate modeling has been of particular interest within the scientific community for studies involving both uncertainty quantification and optimization (examples include Bichon et al., 2008; Jones et al., 1998; Kennedy and O'Hagan, 2001; Bayarri et al., 2002; Simpson et al., 2001; Kaymaz, 2005; Kennedy et al., 2006; Oakley and O'Hagan, 2002).

Giunta et al. (2006) compare response surface based methods for uncertainty propagation to efficient sampling techniques (including Latin hypercube sampling). Samples sizes ranging from 10 to 121 were considered, and the methods were used to estimate the output mean and variance, as well as a 5% probability level. It was found that the use of a kriging (Gaussian process) response surface approximation performed significantly better than LHS, particularly for failure probability estimation, and particularly for moderate to high sample sizes (> 25). Although it may not be appropriate to generalize these results to other functional forms and dimensions, the results may be taken as suggestive of the power of the Gaussian process model for uncertainty quantification.

In any case, before the model output distribution can be estimated, the joint distribution of the model inputs must first be characterized. This often entails using a set of samples, or observations, of the model inputs to estimate an associated probability density function.

One of the most widely used and straightforward methods for characterizing randomness is the use of the normal distribution, which is defined in terms of two parameters, a mean and a variance. The ubiquity of the normal distribution may be in part due to the fact that its use is often justified by the central limit theorem and that it is often an appropriate model for randomness that is observed in the physical world. Algorithms for generating random realiza-

tions of normal random variables are also well established, making the normal distribution a convenient choice for probabilistic simulations.

The multivariate extension of the normal distribution is also a widely used representation for multivariate data. First, the multivariate normal distribution is simple to specify because it is fully defined by a mean vector and covariance matrix (equivalently, a set of marginal normal distributions and the pairwise correlation coefficients among the variables). And as with the univariate normal distribution, algorithms are readily available for generating random samples from the multivariate normal model.

In some cases, however, the normal or multivariate normal model will not be appropriate. A variety of procedures are available for assessing the suitability of the normal model for univariate data, including the normal probability plot (c.f. Haldar and Mahadevan, 2000), the Lilliefors test (Lilliefors, 1967), and the Shapiro-Wilk test (Shapiro and Wilk, 1965). Srivastava (2002) also discusses a few procedures available for assessing the suitability of the multivariate normal model for a set of multivariate data, which is not as straightforward.

In the univariate case, or when dealing with multiple independent variables, non-normality is not typically a problem for uncertainty propagation. Simulation techniques exist for a variety of parametric non-normal probability distributions, and several transformations are also available to help achieve normality (see Rebba, 2005, for an overview of transformations). However, when a group of dependent variables do not fit the multivariate normal model, the situation is not as simple. Transformations that achieve joint normality are more difficult to find, and simulation techniques are not as widely available. One lesser known difficulty is that when the variables are non-normal, specifying the marginal distributions and correlation coefficients does not necessarily result in a fully specified joint distribution. Some specialized

techniques are available to sample from fully-specified parametric joint distributions (Johnson, 1981), but they will not be applicable to most practical problems. Other sampling techniques that have been developed include approximations based on partially specified joint distributions (Lurie and Goldberg, 1998; Iman and Conover, 1982) and the use of bivariate copulas (Haas, 1999).

As an alternative to parametric techniques, which are only applicable when the data conform to the specified probability model, several non-parametric techniques are available that are more generally applicable. Two such non-parametric techniques are the polynomial chaos expansion (c.f. Ghanem and Spanos, 1991; Ghanem and Dham, 1998; Debusschere et al., 2004; Ghanem et al., 2008; Ghanem, 1999) and kernel density estimation. Polynomial chaos expansion is a method whereby a random variable is expanded as a set of orthogonal basis functions on independent, standardized random variables. Kernel density estimation is discussed below in Section 4.2.2.

When dealing with a large number of dependent random variables, uncertainty propagation can often be simplified by finding a more compact representation of the high-dimensional random vector. If the original high-dimensional set can be well-approximated by a lower-dimensional set, then the problem of density estimation need only be concerned with the lower-dimensional set. In probability and statistics, the canonical approach to finding a compact representation of a high-dimensional random vector is what is known as principal component analysis; this approach is outlined below in Section 4.2.3.

4.2.2 Kernel density estimation

Kernel density estimation is a non-parametric method for constructing an estimate of the probability density of a random variable based on a set of observations. The idea is very simple:

the estimate is formed by adding up a set of "kernel densities" centered at each of the observations. The kernel densities function to smooth out the density so that the overall trends of the underlying density can be captured.

Formally, the kernel density estimate $\hat{f}(x)$ is expressed as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{4.10}$$

where n is the number of observations, X_i represents the i^{th} observation, $K(\cdot)$ is the kernel function, and h is a smoothing parameter known as the window width, or bandwidth. The value of h determines the width of the kernel estimate that is placed on each observation. Large values of h will smooth out the features of the density estimate, while small values of h will capture the fine structure of the observations.

Before constructing a kernel density estimate, the analyst must choose both a form for the kernel function, $K(\cdot)$, and a value for the window width. The estimator tends not to be very sensitive to the choice of the kernel function (Simonoff, 1996), so the Gaussian kernel is commonly chosen for ease of computation and smoothness. In this case, $K(\cdot)$ is simply the standard normal PDF, and h is sometimes referred to as the standard deviation of the kernel.

The density estimate is known to be much more sensitive to the choice of the window width. Ideally, one would like to estimate h by minimizing an error measure against the true underlying density, such as the mean integrated squared error (MISE), but doing so would require knowledge of the true density. Some rules of thumb are available that provide quick estimates to h based on the observed data. For example, the "Gaussian reference rule," which

is based on a Gaussian assumption for the underlying density, is given by the simple formula

$$h = 1.059\sigma n^{-1/5}. \tag{4.11}$$

A more rigorous approach is to estimate h by finding the value that minimizes the cross-validation score function (Silverman, 1986)

$$M_1(h) = \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} K^* \left(\frac{X_i - X_j}{h} \right) + \frac{2}{nh} K(0), \tag{4.12}$$

where the function $K^*(\cdot)$ is defined as

$$K^*(t) = K^{(2)}(t) - 2K(t), \tag{4.13}$$

and $K^{(2)}(t)$ is the convolution of $K(\cdot)$ with itself, which in the case of a standard Gaussian kernel is a Gaussian density with variance 2.

Now consider the multivariate case, in which one wishes to construct a joint density estimate based on an $n \times d$ random sample $\boldsymbol{X}$. Several extensions of the density estimate given by Eq. (4.10) are available for multivariate estimation. One of the most straightforward is given by

$$\hat{f}(\boldsymbol{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K \left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h} \right), \tag{4.14}$$

in which the window width is the same in each dimension. Note that $K(\boldsymbol{x})$ is now a multivariate kernel, and one common choice is the standard multivariate normal density (refer to Silverman, 1986; Simonoff, 1996, for more possibilities).

Because the density estimate of Eq. (4.14) uses the same window width for each dimension,

it is usually necessary to first standardize the data in some manner so that the units of measure do not have an adverse effect. One possibility is to first "pre-whiten" or "sphere" the data to have unit covariance. This can be accomplished using the linear transformation given by

$$\boldsymbol{x}' = \boldsymbol{S}^{-1/2}\boldsymbol{x}, \tag{4.15}$$

where $\boldsymbol{S}$ is the sample covariance matrix of the observations. The resulting kernel density estimate of the sphered data is given by

$$\hat{f}(\boldsymbol{x}) = \frac{|\boldsymbol{S}|^{-1/2}}{nh^d} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{S}^{-1/2}(\boldsymbol{x} - \boldsymbol{X}_i)}{h}\right). \tag{4.16}$$

In analogy to Eq. (4.12), h can be chosen to minimize the least-squares cross-validation score

$$M_1(h) = \frac{1}{n^2h^d} \sum_{i=1}^{n}\sum_{j=1}^{n} K^*\left(\frac{\boldsymbol{X}'_i - \boldsymbol{X}'_j}{h}\right) + \frac{2}{nh^d}K(\boldsymbol{0}), \tag{4.17}$$

where $K^*(\boldsymbol{t}) = K^{(2)}(\boldsymbol{t}) - 2K(\boldsymbol{t})$; and the convolution of the standard multivariate normal kernel with itself, $K^{(2)}(\boldsymbol{t})$, is the multivariate normal kernel with covariance $2\boldsymbol{I}$.

One important note regarding multivariate kernel density estimation is that the estimators suffer from the "curse of dimensionality," meaning that as the dimension increases, progressively larger sample sizes will be needed to achieve comparable accuracy. This is because in high dimensions there will almost surely be large regions of the parameter space that do not have any data in them. Simonoff (1996) suggests that "the 'empty space phenomenon' in higher dimensions ... argues against very effective direct density estimation in more than four or five dimensions." One possible solution to this problem is to construct the density estimate based on a lower-dimensional representation of the data; principal component analysis

provides one framework for obtaining such a representation, and it is discussed next.

4.2.3 Principal component analysis

Principal Component Analysis (PCA) is a method whereby a random vector is transformed to a new set of random variables, the principal components, which are uncorrelated. The central idea is that PCA can be used to reduce the dimensionality of the data set by providing a more compact representation of the original random vector.

The PCA is often viewed as a variance maximization technique. This is because the first principal component maximizes the variance of all possible linear combinations of the original variables, the second principal component provides the next maximum variance linear combination, and so on (these variance maximizations are subject to the normalization constraints that the norm of each weight vector is one, as well as the requirement that the weight vectors be orthogonal to each other). As such, a reduced set of principal components can often be used to capture the majority of the variance of the original variables, resulting in a lower-dimensional representation of the original random vector.

The PCA transformation is given by

$$\boldsymbol{y} = \boldsymbol{A}^T \boldsymbol{x}, \tag{4.18}$$

where $\boldsymbol{x}$ is the original random vector, $\boldsymbol{y}$ contains the principal components, and $\boldsymbol{A}$ is a matrix containing as columns the eigenvectors of the covariance matrix of the random vector $\boldsymbol{x}$, ordered according to descending eigenvalues. Thus, the first principal component is the inner product of the first eigenvector with $\boldsymbol{x}$, and so on. In practice, the true covariance matrix is rarely known, and the eigenvectors are typically computed based on the sample covariance of

a set of observations of $\boldsymbol{x}$.

Let the eigenvalues of the covariance matrix be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, where p is the dimensionality of $\boldsymbol{x}$. The variance of the i^{th} principal component is then given by the value of the i^{th} eigenvalue, λ_i. It also follows that if the first k principal components are used to approximate the original random vector, the fraction of the total variance in $\boldsymbol{x}$ explained by $\boldsymbol{y}_{(k)}$ is

$$\frac{\sum_{i=1}^{k} \text{Var}[y_i]}{\sum_{i=1}^{p} \text{Var}[x_i]} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}, \tag{4.19}$$

and the mean-square error associated with the approximation is

$$\varepsilon^2(k) = \text{E}\left[(\boldsymbol{x} - \hat{\boldsymbol{x}}(k))^T (\boldsymbol{x} - \hat{\boldsymbol{x}}(k))\right] = \sum_{i=k+1}^{p} \lambda_i, \tag{4.20}$$

where the original variables are approximated through the reverse transformation as

$$\hat{\boldsymbol{x}}(k) = \boldsymbol{A}_{(k)} \boldsymbol{y}_{(k)}. \tag{4.21}$$

Clearly, the eigenvalues provide guidance as to an appropriate number of components to be retained for the representation of $\boldsymbol{x}$.

When the variables in $\boldsymbol{x}$ are not measured in comparable units, or their variances are dissimilar, it is preferable to calculate the principal components based on the correlation matrix. In this case, the components are expressed as

$$\boldsymbol{y} = \boldsymbol{A}^T \boldsymbol{D}_s^{-1} \boldsymbol{x}, \tag{4.22}$$

where $\boldsymbol{D}_s^{-1}$ is the diagonal matrix of standard deviations and $\boldsymbol{A}$ contains the eigenvectors of

the correlation matrix.

The PCA decomposition of a random vector is also generally equivalent to the Karhunen-Loeve Decomposition (KLD; refer to Liang et al., 2002, for a detailed comparison of PCA, KLD, and singular value decomposition). While the Karhunen-Loeve expansion was originally developed as a means of representing a continuous-time stochastic process, the discrete version of the KLD is equivalent to PCA. Note that the KLD is usually performed after centering the data to have zero mean; this is also useful with PCA as well.

4.3 Summary

While model validation is an important step in the modeling and simulation process, the community has found it difficult to move towards universally accepted practices for model assessment using experimental data, and validation remains an active research area. This chapter highlights the use of statistical significance testing as a means for comparing experimental data and model predictions. However, this emphasis is not intended as an endorsement that significance testing is preferred over other quantitative alternatives. Instead, the aim of Section 4.1.2 is to outline the proper formulation of significance tests for model validation assessment, and to clarify the appropriate interpretation of the test results. The use of significance testing for validation is illustrated in Section 6.1.2.

As discussed in this chapter, model validation analysis is often accompanied by uncertainty propagation, which is the process of estimating the probability distribution of the model output that is implied by probability distributions associated with model inputs. Thus, common practices for uncertainty propagation are discussed in this chapter as well. While none of the topics presented in Section 4.2 constitute new research, the novel approach proposed in Section 6.2.2 for modeling high-dimensional probability distributions makes use of both kernel density es-

timation and principal component analysis, which are discussed in Sections 4.2.2 and 4.2.3, respectively.

CHAPTER V

CALIBRATION OF COMPUTER SIMULATIONS

Many mathematical models that describe the behavior of physical systems can be broken down into two parts: an assumed form for the relationship among a set of quantities, and a set of parameters that are needed to fully determine this relationship. Consider, for example, the model which describes the law of universal gravitation. Mathematically, this model is expressed as

$$F = G\frac{m_1 m_2}{r^2},$$

where F is the attractive force between two bodies having masses m_1 and m_2, r is the distance between the bodies, and G is commonly known as the gravitational constant. With this example, the two components of the model are clear. The assumed form simply states that the attractive force is proportional to the product of the masses and inversely proportional to the square of the distance between them. However, knowing the form of this relationship alone would not allow one to compute F, given m_1, m_2, and r. This is where the second piece of the model comes into play. In this case, the model is specified in terms of one parameter, which is G, the gravitational constant.

In fact, it was 111 years after Newton formally postulated the law of universal gravitation (and 71 years after Newton's death) that a value for G was determined. Without an estimate to the value of G, Newton was limited to calculating the ratio of various gravitational forces. The value for G was first accurately determined by Henry Cavendish in 1798 using a torsion beam and lead spheres. The estimation of such a parameter via experimental observations might be

termed *model calibration*.

This type of model calibration is quite widespread in quantitative analysis. One common example is that of linear regression analysis. A linear relationship is postulated between a dependent variable and one or more independent variables, and then a set of observed values are used to estimate the unknown parameters that determine the model. While parameter estimation is generally well understood for those simple cases in which the postulated model is a closed-form mathematical expression, the calibration of complex computer simulations is not nearly as straightforward.

First, the meaning of the term *calibration*, when used in reference to computer simulations, must be defined more explicitly, because in the modeling and simulation fields, the word calibration can have several interpretations. For this work, *model calibration* refers to:

> The process of adjusting model input parameters in order to improve the agreement between the model output and observed data.

From here on, the term *parameters* will refer to the set of all numerical constants inside the computational simulation that must be specified in order to use the simulation for prediction.

Thus, in this context, the calibration of computational simulations poses several challenges that do not arise in the simpler model calibration exercises:

1. For computational simulations, in virtually no case will the relationship between the inputs and outputs be of a form which allows the model to be manipulated analytically or inverted so that the unknown parameters can be solved for analytically.

2. The class of computational simulations considered here are characterized by long runtimes, such that an exhaustive exploration of the parameter space (for the purpose of

finding those parameters which yield model outputs that agree with the observations) is prohibitive.

3. There may be a wide range of model parameters that provide comparable fits to the observed data. While this scenario may also present itself in the simpler calibration analyses, this lack of uniqueness is particularly common when dealing with computational simulations.

4. As opposed to being a scalar, the output of the simulator may consist of a variety of quantities, and the output may be a function of temporal and/or spatial coordinates. As such it is not necessarily straightforward to develop a metric of comparison between model predictions and observations.

5. In some cases, the analyst may not have the ability to evaluate the simulation, but may be given only a database of previous simulator "runs," specifying the corresponding inputs and outputs for each run.

6. The parameters governing a computational simulation may have little or no physical meaning, often rendering it difficult to know what correspond to "feasible" values or ranges of the parameters.

The rest of this chapter describes a variety of techniques for the calibration of computational simulations. The emphasis is placed on taking a rigorous account of the amount of uncertainty in the resulting parameter estimates. The importance of this uncertainty is clear, because uncertainty associated with the model inputs implies uncertainty associated with the model predictions. If only a single, point estimate is obtained for the model parameters, then the analyst may be discounting a substantial source of uncertainty in the predictions. While

uncertainty in estimated model parameters is certainly not the only source of modeling uncertainty, it is one source that can be quantified using several approaches.

Section 5.2 introduces one of the most straightforward approaches to model calibration, which is based on nonlinear regression analysis. This approach generally boils down to using numerical optimization schemes to solve a "least-squares" problem, but nonlinear regression also provides a variety of approaches for quantifying the uncertainty in the resulting estimates.

While understanding the fundamentals of the "classical" nonlinear regression approach is important, this dissertation places an emphasis on Bayesian inference for model calibration, which is discussed in detail in Section 5.3. The underlying concepts of Bayesian inference are presented in Chapter II. The Bayesian approach is particularly well suited for uncertainty analysis in the calibration of computer simulations, and the framework can be used to take account of a variety of uncertainty sources. Illustrations of the Bayesian approach for model calibration are given in Sections 6.3, 6.4, and 6.5.

Finally, Section 5.4 discusses a particular Bayesian formulation of the calibration problem which is commonly known as the "Kennedy and O'Hagan" framework. This approach is quite ambitious, and is probably the most comprehensive framework for accounting for uncertainty in the calibration process that has been reported in the literature. Kennedy and O'Hagan directly address the issue of simulator expense, and they also introduce an approach to modeling the simulator bias as a function of observable quantities (such as time). While the Kennedy and O'Hagan framework is extremely comprehensive, it can be very difficult, if not impossible, to implement for large-scale calibration analyses. A detailed discussion is provided in Section 5.4 to compare the Kennedy and O'Hagan framework to the more straightforward Bayesian framework presented in Section 5.3.

5.1 Background

Previous work dealing with the calibration and uncertainty quantification of expensive simulations is limited. Campbell (2006) gives an overview of various statistical methods that have been proposed for the calibration of computer simulations. One of the most straightforward approaches is to pose the calibration problem in terms of nonlinear regression analysis (Trucano et al., 2006). The problem is then attacked using standard optimization techniques to minimize, for example, the sum of the squared errors between the predictions and observations. Vecchia and Cooley (1987); Vugrin et al. (2007) illustrate the use of such a method to obtain point estimates and various types of confidence intervals for groundwater flow models.

Other methods which have been proposed include the Generalized Likelihood Uncertainty Estimation (GLUE) procedure (Beven and Binley, 1992), which is somewhat Bayesian in that it attempts to characterize a predictive response distribution by weighting random parameter samples by their likelihoods. However, the GLUE method does not assume a particular distributional form for the errors, which prevents the application of rigorous probabilistic approaches, including maximum likelihood estimation. Methods having their foundation in system identification and being related to the Kalman filter have also been proposed for model calibration, and are particularly suited for situations in which new data become available over time (Stigter and Beck, 1994; Banks, 2001). However, these methods tend to be limited in their applicability to dynamic systems with particular relationships between the unknowns and the observables.

In order to enable exhaustive exploration of the parameter space, even when the simulation being calibrated is expensive, inexpensive approximations, or "surrogate" models are often employed. With the increasing complexity of computer simulations, there has been substantial

interest in techniques for the design and analysis of computer experiments. The use of Gaussian process interpolation has been particularly popular (see, for example, Sacks et al., 1989b; Santner et al., 2003; Martin and Simpson, 2005; Kennedy and O'Hagan, 2001; Currin et al., 1991; Morris et al., 1993). Other approaches that have been considered include techniques for combining simulations with different levels of complexity (Kennedy and O'Hagan, 2000a).

One of the milestone papers for model calibration is Kennedy and O'Hagan (2001). Not only does their formulation treat the computational simulation as a black-box, replacing it by a Gaussian process surrogate, but it also purports to account for all of the uncertainties and variabilities that may be present. Towards this end, they formulate the calibration problem using a Bayesian framework, and both multiplicative and additive "discrepancy" terms are included to account for any deviations of the predictions from the experimental data that are not taken up in the simulation parameters. Further, the additive discrepancy term is formulated as a Gaussian process indexed by the scenario variables (boundary conditions, initial conditions, etc.) which describe the system being modeled. In this regard, their formulation is particularly powerful for cases in which experimental data are available at a relatively large number of different scenarios, and predictions of interest are characterized by extrapolations (or interpolations) in this scenario space. Implementation of their complete framework is quite demanding and requires extensive use of numerical integration techniques such as quadrature or Markov Chain Monte Carlo integration.

Other work that has employed the Bayesian framework has focused more on general inverse problem analysis than on model calibration (although model calibration can be viewed as an inverse problem), and several different approaches have been proposed for dealing with the computational expense that Bayesian inference entails. Wang and Zabaras (2005) make

use of proper orthogonal decomposition and Galerkin projection. Balakrishnan et al. (2003) make use of the non-intrusive polynomial chaos representation introduced by Isukapalli et al. (1998). Marzouk et al. (2007) proposes an intrusive polynomial chaos formulation for efficient Bayesian inference for inverse problems.

Although the Bayesian approach to calibration and uncertainty quantification appears quite promising, there have been few attempts in the literature to illustrate how calibration methodologies providing uncertainty representations should be applied to "large-scale" problems, in which simulation time is long, the number of parameters to be estimated may be large, the amount of experimental data is small, and the response quantity is multivariate. The example reported by Kennedy and O'Hagan (2001) deals with a relatively large amount of experimental data, a small parameter space, and a scalar response quantity.

Furthermore, part of the power of the Bayesian approach is its flexibility, but there has been little previous work which shows how the Bayesian model calibration approach can be extended to account for additional forms of uncertainty that are common to real-world modeling and simulation applications. Such extensions include the ability to handle measurement uncertainty characterized with bounds (as opposed to a Gaussian distribution) and model input parameters with prescribed uncertainty distributions.

5.2 Nonlinear regression

Regression analysis is the study of the relationship between a dependent response variable and a set of independent, explanatory, variables. Linear regression analysis is restricted to the case in which the dependent variable is expressed as a linear function of the unknown parameters. Nonlinear regression analysis is the extension for which the relationship between the dependent variable and the unknowns may have any functional form. As such, nonlinear

regression analysis provides one way of thinking about the calibration of computer simulations, because it allows the simulation to be viewed as a general, black-box function.

When nonlinear regression analysis is applied to the calibration of computer simulations, the dependent variable is the simulator output, the independent variables are typically observable experimental conditions, and the unknowns are those internal simulation parameters that are to be estimated. Thus, the relationship between the dependent and independent variables is expressed as

$$y = G(\boldsymbol{\theta}, \boldsymbol{s}), \tag{5.1}$$

where y is the dependent variable, $\boldsymbol{\theta}$ is a p-dimensional vector of calibration parameters, $\boldsymbol{s}$ is the vector of dependent variables, and $G(\cdot, \cdot)$ represents the computer simulation.

Consider that the calibration parameters are to be estimated using n experimental observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ of the dependent variable(s) that correspond to the values of the independent variables $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$. The nonlinear regression model, which relates the predicted and observed values, will be written as

$$y_i = G(\boldsymbol{\theta}, \boldsymbol{s}_i) + \varepsilon_i. \tag{5.2}$$

In the simplest case, the random errors, ε_i, are taken to be independently and identically distributed as $\varepsilon_i \sim N(0, \sigma^2)$. This formulation may be appropriate if the assumption of independence is reasonable and all of the y_i have the same units. However, when some of the y_i represent different quantities (possibly different features of the same response), then the i.i.d. model is no longer appropriate. To take account of observations with different units, as well as dependencies among the observations, an $n \times n$ weighting matrix $\boldsymbol{\omega}$ can be incorporated into

the model, so that the ε_i have a joint distribution $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{\omega}^{-1})$. In the case of independent observations, $\boldsymbol{\omega}$ is diagonal, and the diagonal elements represent the weights given to each observation.

When the experimental data consist of repeated observations of multiple different response features, then the observed variance of each feature can be used to estimate an appropriate weight for that feature. In this case, the value of all diagonal elements of $\boldsymbol{\omega}$ corresponding to a particular feature could be set to the inverse of the sample variance observed for that feature.

The weighted least-squares estimator, $\hat{\boldsymbol{\theta}}$, is that which minimizes the weighted sum of squared errors function[1]:

$$S(\boldsymbol{\theta}) = \left[\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{\theta})\right]^T \boldsymbol{\omega} \left[\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{\theta})\right], \tag{5.3}$$

where $\boldsymbol{G}(\boldsymbol{\theta}) = \big(G(\boldsymbol{\theta}, \boldsymbol{s}_1), \ldots, G(\boldsymbol{\theta}, \boldsymbol{s}_n)\big)^T$. It is easy to see that when the weighting matrix is the identity matrix (equivalently, when the errors are independently and identically distributed), then $S(\boldsymbol{\theta})$ can be written as

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[y_i - G(\boldsymbol{\theta}, \boldsymbol{s}_i)\right]^2. \tag{5.4}$$

Unlike the case in which the relationship between the outputs and the unknowns is linear, there is no general analytical solution for the value of $\boldsymbol{\theta}$ that minimizes the sum of squared errors function for nonlinear relationships. Nonlinear regression analysis thus typically relies on numerical optimization procedures for finding the minimum. In fact, there are a variety of specialized techniques for solving nonlinear least-squares problems (Levenberg-Marquardt

[1]It is shown in Section 5.5 that the least-squares estimator is in this case also the maximum likelihood estimator.

methods are particularly widespread; c.f. Seber and Wild, 2003).

In regression analysis, the amount of uncertainty associated with a particular estimate of the unknown parameters is expressed through confidence intervals. For example, a 95% confidence interval for a parameter θ_0 captures the notion that one has a certain amount of confidence that the true, unknown, value of θ_0 lies within that interval. As the confidence level increases, the size of the interval must increase as well. From a frequentist standpoint, the confidence level has a specific interpretation as a probability: in the long run, 95% of all confidence intervals constructed at the 95% confidence level will contain the true value of the parameter.

When one wants to make inference about multiple parameters, multi-dimensional confidence regions come in to play, as opposed to simple intervals. In nonlinear regression analysis, there are a couple of common approaches for constructing approximate confidence regions. One of the simplest is the "linear approximation confidence region." First, define the $n \times p$ matrix of derivatives of the model outputs with respect to the calibration inputs as

$$\boldsymbol{V} = \left[\frac{\partial G(\boldsymbol{\theta}, \boldsymbol{s}_i)}{\partial \boldsymbol{\theta}_j} \right]_{i,j}. \tag{5.5}$$

Since $\boldsymbol{V}$ will most likely be a function of $\boldsymbol{\theta}$, let $\hat{\boldsymbol{V}}$ denote $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$. Then for large n, the $100(1-\alpha)\%$ linear approximation confidence region for $\boldsymbol{\theta}$ consists of all values of $\boldsymbol{\theta}$ which satisfy the inequality

$$\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T \hat{\boldsymbol{V}}^T \boldsymbol{\omega} \hat{\boldsymbol{V}} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) \leq p s^2 F_{\alpha,p,n-p}, \tag{5.6}$$

where $F_{\alpha,p,n-p}$ is the upper α probability point of the F-distribution with $(p, n-p)$ degrees of freedom, and s^2 is the sample estimate of σ^2, given by $s^2 = S(\hat{\boldsymbol{\theta}})/(n-p)$. Notice that (5.6) is a quadratic form in $\boldsymbol{\theta}$, and the resulting confidence region will be an ellipsoid.

Confidence regions constructed based on the linear approximation theory can be particularly inaccurate for small sample sizes and strongly nonlinear models. An alternative approach is to consider contours of $S(\boldsymbol{\theta})$. Since $S(\boldsymbol{\theta})$ measures the "closeness" of the data to the predictions, it seems intuitive to base the confidence region for $\boldsymbol{\theta}$ on the contours of this function. Such a region might have the form (Seber and Wild, 2003)

$$S(\boldsymbol{\theta}) \leq cS(\hat{\boldsymbol{\theta}}), \tag{5.7}$$

for some constant $c > 1$. Regions of this form are often called "exact" confidence regions because they are not based on any approximations. The difficulty, however, is that the particular coverage probabilities corresponding to the various contour levels are not generally known. A common approach is to employ asymptotic theory, which can be used to show that for large enough n, the following confidence region based on the contours of the sum of squares function holds (Seber and Wild, 2003):

$$S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}}) \leq ps^2 F_{\alpha,p,n-p}. \tag{5.8}$$

Following Seber and Wild (2003), confidence regions constructed using the above inequality will be referred to as "exact" regions because they are based on contours of the likelihood function (see Section 5.5), even though the particular confidence levels are based on an asymptotic approximation.

The primary disadvantage of confidence regions of the form (5.8) is the computational expense. Each evaluation of the inequality requires computing $S(\boldsymbol{\theta})$ for a particular $\boldsymbol{\theta}$, which in turn may require up to n evaluations of the computer simulation $G(\cdot, \cdot)$. As such, it may be

very expensive to find or plot a particular confidence region. Donaldson and Schnabel (1987) present a comparison of several approaches to constructing confidence regions that includes the linear approximation and "exact" methods, as well as three different derivative estimation schemes. Using a Monte Carlo coverage probability study, they found the "exact" method to be very reliable, whereas the linear approximation region may be quite inaccurate when the sample size is small or $G(\cdot, \cdot)$ is highly nonlinear.

One of the limitations of the above confidence regions is that they do not lend themselves to graphical visualization when the dimension, p, is greater than two or three. A couple of options are available for visualization when $p > 2$. The first option amounts to plotting two-dimensional "slices" of the full confidence region for specific values of the "nuisance" variables (those parameters not represented in the two-dimensional plot). Rawlings et al. (1998) provide several examples of this approach for linear regression. While such an approach is simple, the two-dimensional slices are only applicable to particular values of the nuisance variables, and are particularly difficult to interpret for $n \geq 4$.

An alternative is to construct a two-dimensional joint confidence region for two parameters of interest, which *ignores* the other $(p - 2)$ parameters. This concept is analogous to well-known methods for constructing univariate confidence intervals in linear regression. It is also similar to the idea of marginal distributions, which comes into play in Bayesian inference. Consider that $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, where $\boldsymbol{\theta}_2$ contains the p_2 parameters of interest (for a two-dimensional confidence region, $p_2 = 2$). The linear approximation confidence region of (5.6) becomes (Seber and Wild, 2003):

$$\left(\boldsymbol{\theta}_2 - \hat{\boldsymbol{\theta}}_2\right)^T \hat{\boldsymbol{C}}_2^{-1} \left(\boldsymbol{\theta}_2 - \hat{\boldsymbol{\theta}}_2\right) \leq p_2 F_{\alpha, p_2, n-p}, \tag{5.9}$$

where $\hat{\boldsymbol{C}}_2$ is the $p_2 \times p_2$ matrix containing the corresponding elements of the complete covariance matrix estimated by $s^2(\boldsymbol{V}^T\boldsymbol{V})^{-1}$.

Seber and Wild (2003) also discuss an adaptation of the exact regions that can be used for plotting confidence regions for two-dimensional parameter subsets. Let $\boldsymbol{\theta}_2$ denote a p_2-dimensional parameter subset of interest, and let $\boldsymbol{\theta}_1$ denote the remaining (nuisance) parameters. Eq. (5.8) is then adapted as

$$S\left[\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2), \boldsymbol{\theta}_2\right] - S(\hat{\boldsymbol{\theta}}) \leq p_2 s^2 F_{\alpha,p_2,n-p}, \tag{5.10}$$

where $\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)$ contains those values of $\boldsymbol{\theta}_1$ that minimize $S(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ for a given $\boldsymbol{\theta}_2$. Thus, $\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)$ must be computed for each value of $\boldsymbol{\theta}_2$.

Each evaluation of inequality (5.10) involves a least squares optimization problem over the $(p-2)$ parameters in $\boldsymbol{\theta}_1$. This becomes extremely expensive, because to plot a two-dimensional confidence region may require hundreds of evaluations of the inequality itself, for various values of $\boldsymbol{\theta}_2$.

5.3 Bayesian analysis for model calibration

The goal of the Bayesian approach for model calibration is twofold: (1) use observed data to estimate the calibration parameters, $\boldsymbol{\theta}$, and (2) develop a quantitative representation of the resulting estimation uncertainty. While these are essentially the same objectives that were addressed above via nonlinear regression analysis, the Bayesian framework provides a more elegant, meaningful, and extensible approach to quantifying the uncertainty in the parameter estimates.

5.3.1 Formulation

As with the previous section, n observed response values $\boldsymbol{y} = (y_1, \ldots, y_n)$ are to be used to make inference about the value of a set of simulation inputs $\boldsymbol{\theta}$. The simulation is again represented by the operator $G(\boldsymbol{\theta}, \boldsymbol{s})$, where the vector of inputs $\boldsymbol{s}$ contains the "scenario-descriptor" inputs, which may typically represent boundary conditions, initial conditions, geometry, etc. Kennedy and O'Hagan (2001) term these inputs "variable inputs," because they take on different values for different realizations of the system; in classical analysis, they are often referred to as independent variables or covariates.

A probabilistic relationship between the model output, $G(\boldsymbol{\theta}, \boldsymbol{s})$, and the observed data, $\boldsymbol{y}$, is next postulated. A simple but powerful model is the same relationship that was introduced in Eq. (5.2), namely

$$y_i = G(\boldsymbol{\theta}, \boldsymbol{s}_i) + \varepsilon_i, \tag{5.11}$$

where ε_i is a random variable that can encompass both measurement errors on y_i and modeling errors associated with the simulation $G(\boldsymbol{\theta}, \boldsymbol{s})$. The most frequently used assumption for the ε_i is that they are i.i.d $N(0, \sigma^2)$, which means that the ε_i are independent, zero-mean Gaussian random variables, with variance σ^2. Of course, more complex models may be applied, for instance enforcing a parametric dependence structure among the errors.

Once the probabilistic model is specified, a likelihood function for the unknowns may be developed. As discussed in Chapter II, the likelihood function plays a central role in Bayesian inference. Based on the model defined by Eq. (5.11), the likelihood function for $\boldsymbol{\theta}$ is the product of n normal probability density functions:

$$L(\boldsymbol{\theta}) = f(\boldsymbol{d} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\left(y_i - G(\boldsymbol{\theta}, \boldsymbol{s}_i)\right)^2}{2\sigma^2} \right]. \tag{5.12}$$

where $\boldsymbol{d}$ is used generically to represent the observed data and in this case simply contains the experimental observations $\boldsymbol{y}$. Bayes' theorem (Eq. (2.1)) is now applied using the likelihood function of Eq. (5.12) along with a prior distribution for $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$, to come up with a posterior distribution, $f(\boldsymbol{\theta} \mid \boldsymbol{d})$, which represents the belief about $\boldsymbol{\theta}$ in light of the data $\boldsymbol{d}$:

$$f(\boldsymbol{\theta} \mid \boldsymbol{d}) \propto \pi(\boldsymbol{\theta}) L(\boldsymbol{\theta}). \tag{5.13}$$

The posterior distribution for $\boldsymbol{\theta}$ represents the complete state of knowledge, and may even include effects such as multiple modes, which would represent multiple competing hypotheses about the true (best-fitting) value of $\boldsymbol{\theta}$. Summary information can be extracted from the posterior, including the mean (which is typically taken to be the the "best guess" point estimate) and standard deviation (a representation of the amount of residual uncertainty). It is also possible to extract one or two-dimensional marginal distributions, which simplify visualization of the features of the posterior.

However, as discussed in Chapter II, the posterior distribution can not usually be constructed analytically, and this will almost certainly not be possible when a complex simulation model appears inside the likelihood function. One of the more popular numerical techniques for constructing the posterior distribution is Markov Chain Monte Carlo (MCMC) simulation, which is discussed in Section 2.3. Unfortunately, though, MCMC simulation requires hundreds of thousands of evaluations of the likelihood function, which in the case of model calibration equates to hundreds of thousands of evaluations of the computer model $G(\cdot, \cdot)$. For most realistic models, this number of evaluations will not be feasible. In such situations, the analyst must usually resort to the use of a more inexpensive surrogate (a.k.a response surface approximation) model. Such a surrogate might involve reduced order modeling (e.g., a coarser mesh)

or data-fit techniques such as Gaussian process (a.k.a kriging) modeling.

This work adopts the approach of using a Gaussian process surrogate to the true simulation. This particular surrogate modeling approach is attractive here for several reasons:

1. The Gaussian process model is incredibly flexible, and can be used to fit data associated with virtually any functional form.

2. The Gaussian process model is stochastic, thus providing both an estimated response value and an uncertainty associated with that estimate. Conveniently, the Bayesian framework makes it possible to take account of this uncertainty.

3. With regards to fit accuracy, the Gaussian process model has been shown to be competitive with most other modern data fit methods, including Bayesian neural networks and Multiple Adaptive Regression Splines (Rasmussen, 1996; Giunta et al., 2006), and it can represent functions with multiple inputs.

For Bayesian model calibration with an expensive simulation, the uncertainty associated with the use of a Gaussian process surrogate can be accounted for through the likelihood function by incorporating the direct variance estimates from the GP. Although a complete Bayesian approach would even treat the parameters governing the GP as objects of Bayesian inference, this approach is rather complicated and is not believed to contribute much to the overall uncertainty analysis (Kennedy and O'Hagan, 2001). The approach recommended here is to estimate the parameters governing the GP *a priori* using the observed simulator outputs, and to treat them as known constants for the remainder of the calibration analysis.

Through the assumptions used for Gaussian process modeling, the response conditional on a set of observed "training points" follows a multivariate normal distribution. For a discrete set of new inputs, this response is characterized by a mean vector and a covariance matrix (see

Eqs. (3.4) through (3.6)). Denote the mean vector and covariance matrix corresponding to the inputs $(\boldsymbol{\theta}, \boldsymbol{s}_1), \ldots, (\boldsymbol{\theta}, \boldsymbol{s}_n)$ as $\boldsymbol{\mu}_{GP}$ and $\boldsymbol{\Sigma}_{GP}$, respectively. It is easy to show that the likelihood function for $\boldsymbol{\theta}$ is then given by a multivariate normal probability density function (note that the likelihood function of Eq. (5.12) can also be expressed as a multivariate normal probability density, with $\boldsymbol{\Sigma}$ diagonal):

$$L(\boldsymbol{\theta}) = (2\pi)^{-n/2} \left|\boldsymbol{\Sigma}\right|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}_{GP})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_{GP})\right], \tag{5.14}$$

where $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I} + \boldsymbol{\Sigma}_{GP}$, so that both $\boldsymbol{\mu}_{GP}$ and $\boldsymbol{\Sigma}$ depend on $\boldsymbol{\theta}$.

Simply put, since the uncertainty associated with the surrogate model is independent of the modeling and observation uncertainty captured by the ε_i, the covariance of the Gaussian process predictions ($\boldsymbol{\Sigma}_{GP}$) simply adds to the covariance of the error terms ($\sigma^2 \boldsymbol{I}$). (An excellent illustration of the effect on the calibration results of acknowledging the Gaussian process surrogate uncertainty is Figure 6.24 and the accompanying discussion.) As mentioned before, if a more complicated error model is desired (i.e. one in which the errors are not independent of each other), $\sigma^2 \boldsymbol{I}$ is replaced by a full covariance matrix.

5.3.2 Prescribed input uncertainties

In some cases it may be of interest to study how the results of a calibration analysis are affected by additional modeling uncertainties. In the Bayesian setting, the most obvious approach would be to augment the set of calibration parameters $\boldsymbol{\theta}$ with the additional uncertain model inputs. If the data $\boldsymbol{d}$ do not provide any information about these additional uncertain inputs, then they will essentially be sampled over their prior distribution, potentially resulting in an increase in the uncertainty in the original calibration parameters. On the other hand, if the

data $\boldsymbol{d}$ do provide information about the additional inputs, then their posterior distribution will most likely reflect less uncertainty than their prior. However, if one is strictly interested in the effect of additional prescribed input uncertainties, such inputs can not be treated as calibration inputs, because their posterior may not match the prescribed distribution of interest. Thus, this section presents a method which allows the Bayesian calibration analysis to take account of prescribed uncertainties for additional model inputs.

Denote those inputs to the simulation $G(\cdot)$ having prescribed probability distributions by $\boldsymbol{\xi}$. By explicitly writing these additional inputs, the simulation model is now a function of the calibration inputs, the scenario-descriptor inputs, and the inputs with prescribed distributions: $y = G(\boldsymbol{\theta}, \boldsymbol{s}, \boldsymbol{\xi})$. Denote the probability density function associated with $\boldsymbol{\xi}$ by $f(\boldsymbol{\xi})$. In order to develop the posterior distribution for $(\boldsymbol{\theta}, \boldsymbol{\xi})$ in which the distribution of $\boldsymbol{\xi}$ is not refined by $\boldsymbol{d}$, it is necessary to artificially assume that the data $\boldsymbol{d}$ are statistically independent of $\boldsymbol{\xi}$. Whether or not this is true in reality can be checked by treating $\boldsymbol{\xi}$ as a calibration parameter in $\boldsymbol{\theta}$, but by artificially enforcing the assumption, the parameters $\boldsymbol{\xi}$ are held to the prescribed distribution $f(\boldsymbol{\xi})$.

Assuming that $\boldsymbol{\xi}$ is independent of $\boldsymbol{d}$ yields

$$f(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \boldsymbol{d}) \propto \pi(\boldsymbol{\theta}) L(\boldsymbol{\theta}) f(\boldsymbol{\xi}).$$

Since the simulation output is a function of $\boldsymbol{\xi}$, $L(\boldsymbol{\theta})$ is as well, so for clarity the likelihood

function will be written as $L(\boldsymbol{\theta};\boldsymbol{\xi})$[2], which yields:

$$f(\boldsymbol{\theta},\boldsymbol{\xi} \mid \boldsymbol{d}) \propto \pi(\boldsymbol{\theta})L(\boldsymbol{\theta};\boldsymbol{\xi})f(\boldsymbol{\xi}). \tag{5.15}$$

Ultimately, though, one is interested in the posterior of $\boldsymbol{\theta}$ after marginalizing over the "nuisance" variable $\boldsymbol{\xi}$, which is

$$f(\boldsymbol{\theta} \mid \boldsymbol{d}) \propto \int \pi(\boldsymbol{\theta})L(\boldsymbol{\theta};\boldsymbol{\xi})f(\boldsymbol{\xi})\, d\boldsymbol{\xi}. \tag{5.16}$$

This marginalization is trivial if $f(\boldsymbol{\theta},\boldsymbol{\xi} \mid \boldsymbol{d})$ is constructed using Markov Chain Monte Carlo sampling. One possibility for constructing $f(\boldsymbol{\theta},\boldsymbol{\xi} \mid \boldsymbol{d})$ is to use a component-wise scheme to sequentially sample each component of $(\boldsymbol{\theta},\boldsymbol{\xi})$ from its respective full conditional distribution. Each component of $\boldsymbol{\theta}$ can be sampled using the Metropolis algorithm, by sampling the ith component from its full conditional:

$$f(\theta_i \mid \boldsymbol{\theta}_{-i},\boldsymbol{\xi},\boldsymbol{d}) \propto \pi(\boldsymbol{\theta})L(\boldsymbol{\theta};\boldsymbol{\xi}), \tag{5.17}$$

where $\boldsymbol{\theta}_{-i}$ contains all components of $\boldsymbol{\theta}$ except for θ_i. Notice that $f(\boldsymbol{\xi})$ does not appear in Eq. (5.17) because it does not depend on $\boldsymbol{\theta}$.

Further, if the joint distribution of $\boldsymbol{\xi}$ is such that it can be sampled (in particular, if the components of $\boldsymbol{\xi}$ are independent and can be sampled), the vector $\boldsymbol{\xi}$ can be directly sampled at each iteration. This is because the full conditional of $\boldsymbol{\xi}$ is equal to $f(\boldsymbol{\xi})$: at each iteration a sample of $\boldsymbol{\xi}$ is drawn from $f(\boldsymbol{\xi})$, which is its full conditional.

[2]Although it is tempting to write $L(\boldsymbol{\theta},\boldsymbol{\xi})$, I avoid doing so because this is really $f(\boldsymbol{d} \mid \boldsymbol{\theta},\boldsymbol{\xi})$; since $\boldsymbol{\xi}$ is (assumed to be) statistically independent of $\boldsymbol{d}$, this would reduce to $f(\boldsymbol{d} \mid \boldsymbol{\theta}) = L(\boldsymbol{\theta})$. Thus, I write $L(\boldsymbol{\theta};\boldsymbol{\xi})$ to emphasize that it is a function of $\boldsymbol{\xi}$, but there is no statistical relationship between $\boldsymbol{\xi}$ and $\boldsymbol{d}$.

In short, the process for accounting for prescribed input uncertainties within the Bayesian calibration analysis is very simple, given that Markov Chain Monte Carlo is used to construct the posterior for $\boldsymbol{\theta}$. To account for the additional total uncertainty introduced by the inputs, $\boldsymbol{\xi}$, having prescribed uncertainties, one simply samples a random realization of $\boldsymbol{\xi}$ at each iteration of the MCMC sampler.

5.3.3 Characterized observation and modeling uncertainty

In the probabilistic error model of Eq. (5.11), ε is a random variable that encompasses both observation uncertainty in the data $\boldsymbol{d}$ and modeling error: together, these effects result in a difference between the observations and the predictions. In most cases, the overall magnitude of this net effect (represented by the variance of ε, σ^2) is not known, and σ^2 is treated as an object of Bayesian inference along with the calibration inputs, $\boldsymbol{\theta}$. However, in some cases, the experimental instrumentation may be understood well enough that the error associated with the observed data $\boldsymbol{d}$ can be characterized using a parametric probability distribution. For example, the experimenter might claim that the errors in the measurements $\boldsymbol{d}$ follow a Gaussian distribution with zero mean and standard deviation equal to 10% of the measured value.

Similarly, the error associated with the analysis code $G(\cdot)$ may also be characterized in some sense. For example, based on a mesh convergence study, an analyst may be able to quantitatively characterize the magnitude of the error associated with the output of $G(\cdot)$.

When the error/uncertainty associated with the observations and/or analysis code can be characterized, it would be nice to include it in the probabilistic model. In most cases, one would still want to retain a separate ε term, which would represent all other sources of error that lead to a difference between the predictions and observations. Thus, a new probabilistic

model might be formulated as:

$$y_i = G(\boldsymbol{\theta}, \boldsymbol{s}_i) + \varepsilon_i + u_i, \tag{5.18}$$

where the random variable u_i represents the characterized uncertainty associated with either the observation y_i or the analysis code output $G(\boldsymbol{\theta}, \boldsymbol{s}_i)$.

For simple cases in which both ε and u are Gaussian random variables, one can simply replace the two of them with one random variable which is their sum, and it will be Gaussian as well. However, while ε is most often taken to be Gaussian, other distributions might be chosen for u. For example, the experimentalist might characterize the measurement uncertainty with bounds, in which case it would be most appropriate to use a uniform probability distribution for u. In such cases, it may be very difficult to analytically express the probability distribution of the sum $\varepsilon + u$, and alternative methods may be more prudent.

One possibility is to use the same approach that was taken in Section 5.3.2 and sample $\boldsymbol{u}$ along with $\boldsymbol{\theta}$. First, denote the joint probability density function for $\boldsymbol{u} = (u_1, \ldots, u_n)$ by $f(\boldsymbol{u})$. Then, analogously to Eq. (5.15), this results in

$$f(\boldsymbol{\theta}, \boldsymbol{u} \mid \boldsymbol{d}) \propto \pi(\boldsymbol{\theta}) L(\boldsymbol{\theta}; \boldsymbol{u}) f(\boldsymbol{u}). \tag{5.19}$$

Considering the case in which $\boldsymbol{u}$ represents characterized observation uncertainty, it is clear that the likelihood function for $\boldsymbol{\theta}$ depends on $\boldsymbol{u}$ in the sense that after subtracting the effect of u_i, the observation is actually given by $y_i - u_i$. That is, the likelihood function of Eq. (5.12) would become

$$L(\boldsymbol{\theta}; \boldsymbol{u}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\left(y_i - u_i - G(\boldsymbol{\theta}, \boldsymbol{s}_i)\right)^2}{2\sigma^2} \right]. \tag{5.20}$$

Thus, as outlined in Section 5.3.2, the approach is to sample a random realization of $\boldsymbol{u}$ from $f(\boldsymbol{u})$ at each iteration of the MCMC sampler. Then, before computing $L(\boldsymbol{\theta})$, the observed data are artificially perturbed as $\boldsymbol{y} - \boldsymbol{u}$.

5.4 The Kennedy and O'Hagan framework

The work of Kennedy and O'Hagan (2001) is what initially brought to light the power of the Bayesian approach for the calibration of computer simulations. However, their formulation contains several technical and philosophical differences from the approach presented in Section 5.3. Perhaps of most interest is their incorporation of a scenario-dependent "bias" or "model inadequacy" function, which is intended to capture systematic discrepancies between the predictions and observations. This section will outline those factors which differentiate the Kennedy and O'Hagan approach, and discuss why such extensions are often not necessary or not applicable.

Kennedy and O'Hagan (hereafter KOH) employ a probabilistic relationship between the predictions and observations that is only a slight extension of Eq. (5.11):

$$y_i = \rho\eta(\boldsymbol{\theta}, \boldsymbol{s}_i) + \delta(\boldsymbol{s}_i) + \varepsilon_i. \tag{5.21}$$

$\eta(\cdot,\cdot)$ is analogous to $G(\cdot,\cdot)$ in Eq. (5.11), but a different notation is used here because KOH treat it explicitly as an unknown function, as discussed in more detail below. ρ is an unknown "regression parameter" and $\delta(\cdot)$ is the model inadequacy function. The addition of constant regression parameters, such as ρ above, is a simple matter, and such extensions could easily be incorporated into the framework discussed in Section 5.3 at the analyst's discretion.

The addition of the model inadequacy function, however, does add a significant degree

of complexity to the analysis. The reason for this is that the model inadequacy function is treated as a Gaussian process indexed by the scenario inputs, $\boldsymbol{s}$. This adds a fundamental difficulty, because the observed simulator outputs must now be used to estimate the parameters governing a Gaussian process for $\eta(\cdot, \cdot)$ *and* a Gaussian process for $\delta(\cdot)$. Further, the Gaussian process for $\delta(\cdot)$ is a function of the scenario inputs but not the calibration inputs, which further complicates its estimation because the code outputs are functions of both the calibration inputs and the scenario inputs. The approach discussed by Kennedy and O'Hagan (2000b) involves a complicated approximation scheme for integrating out $\boldsymbol{\theta}$ (after using the observed code outputs to estimate the Gaussian process for $\eta(\cdot, \cdot)$), but $\delta(\cdot)$ is often estimated in practice by holding $\boldsymbol{\theta}$ fixed at a nominal value.

Aside from the difficulties associated with the estimation of the governing parameters of the Gaussian process for the model inadequacy function, there are many applications in which its use is simply not relevant. The purpose of the model inadequacy function is to estimate the relationship between the model bias and the scenario variables, in the hope that this bias can be extrapolated to an "application configuration" in order to improve the predictive capability of the calibrated model. For example, the scenario variables may represent location on a two-dimensional grid, and the model inadequacy function may capture the notion that the model bias is an increasing function of the x-coordinate, say. However, there are several prerequisites for this type of bias estimation analysis:

1. $\boldsymbol{s}$ contains one or more continuous variables that represent a smooth mapping among various system configurations, *including the configuration corresponding to the intended use of the simulation*.

2. The value of $\boldsymbol{s}$ is known and quantifiable for each experimental observation, as is the

value of $\boldsymbol{s}$ for the configuration corresponding to the intended use of the simulation.

3. The experimental data contain enough information (at the very least representing several different values of $\boldsymbol{s}$) to quantify a systematic relationship between the model bias and $\boldsymbol{s}$.

It happens that for many real calibration problems, the first requirement is not even met (see, for example, the applications of Sections 6.2, 6.3, 6.4, and 6.5) because the relationship among the configurations of interest can not be represented in terms of a set of scenario variables. Input excitations may change from sinusoidal to shock, components may come together to create new systems, and often the ultimate system configuration of interest is so different from those for which experimental data are available that there is simply no hope of achieving a parametric mapping between the two regimes.

Further, even when it may be possible to quantify the mapping in terms of a set of scenario variables $\boldsymbol{s}$, the experimental data simply may not provide enough information about the relationship between $\boldsymbol{s}$ and δ to construct a meaningful model inadequacy function. One final caution is that Gaussian process modeling is generally intended for data interpolation, and using such models for extrapolation (which is in most cases the purpose of the model inadequacy function) should be done with caution (refer to Chapter III: in most cases, extrapolative predictions made using a Gaussian process will be strongly influenced by the unconditional mean function of the process).

The next difference between the KOH approach and that discussed in Section 5.3 is the manner in which the Gaussian process representation of the computer simulation is treated. The approach presented in Section 5.3 is to consider this GP as a surrogate model to the true computer simulation, which is (understandably) statistically independent of the experimental

observations. The approach provided by KOH, however, is different from both a theoretical and a philosophical standpoint.

The KOH methodology is developed from the standpoint that $\eta(\cdot, \cdot)$ is an unknown function, given a Gaussian process prior distribution. Interestingly, Kennedy and O'Hagan do not explicitly derive the equations for the posterior distribution of $\eta(\cdot, \cdot)$ conditional on the observed simulation outputs, focusing only on the posterior distribution of the "real process." As a result, their posterior distribution is developed in one Bayesian updating step based on *both* the experimental observations and the observed simulator outputs. This is fundamentally different from the approach of Section 5.3, in which the Bayesian updating is viewed in terms of the experimental data only, and the observed simulator outputs are only used beforehand to construct the surrogate to the simulator.

As a result of their consideration of $\eta(\cdot, \cdot)$ as an unknown function and not a surrogate to the simulator, the experimental observations and the observed code outputs are *not* statistically independent, and the covariance between these pieces of data appears explicitly in the likelihood, and is based on the covariance function for the GP $\eta(\cdot, \cdot)$. While this approach seemingly provides a more comprehensive treatment, it results in a formulation that will often be computationally intractable in practice. The complete covariance matrix under this approach is of size $m + n$, where m is the number of observed simulator runs and n is the number of experimental data points. The use of Markov Chain Monte Carlo simulation to construct the posterior distribution will require hundreds of thousands of evaluations of the inverse of this $(m + n) \times (m + n)$ matrix.

Computationally, there is a very large difference between the two approaches. If the Gaussian process model of the simulator is constructed *a priori* and considered as a surrogate (as

discussed in Section 5.3), the $m \times m$ covariance matrix corresponding to the observed simulator outputs only needs to be inverted once, and the data covariance matrix that gets inverted at each MCMC iteration is then only of size $n \times n$. This discrepancy is further amplified by the fact that in most cases m will be much larger than n (more simulator runs than experiments). Further, the $n \times n$ experimental data covariance matrix will often be diagonal (or at least well-conditioned), whereas the $(m + n) \times (m + n)$ combined covariance matrix will tend to be ill-conditioned, and may turn out to be singular to working precision in many cases, making construction of the posterior distribution difficult, if not impossible.

One final difference between the KOH formulation and the straightforward formulation developed in Section 5.3 is the treatment of the unknowns. KOH attempt to develop their framework under a "full Bayesian" philosophy, in which all unknowns are treated as objects of Bayesian inference. In short, the difference is that much of the theoretical development given in KOH treats the governing parameters of the Gaussian process models as objects of Bayesian inference, whereas a more straightforward approach is to simply estimate these parameters *a priori* and treat them as known constants for the remainder of the analysis. Not only does the treatment of the GP parameters as objects of Bayesian inference add a significant amount of complexity, but KOH eventually acknowledge that accounting for the uncertainty in the estimates of the GP parameters does not add to the overall calibration uncertainty analysis in any meaningful way.

5.5 Equivalencies to least-squares estimation

The least-squares estimator in nonlinear regression analysis discussed in Section 5.2 is often attractive because of its interpretation as an estimator that minimizes a simple and understandable error metric. This section provides a brief discussion to show that under certain fairly

general conditions, the least-squares estimator is also a maximum likelihood estimator, as well as a Bayesian maximum posterior estimator.

First, recall the nonlinear regression model of Eq. (5.2),

$$y_i = G(\boldsymbol{\theta}, \boldsymbol{s}_i) + \varepsilon_i,$$

and the joint error model that was assumed for $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$:

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\omega}^{-1}),$$

where $\boldsymbol{\omega}$ is the $n \times n$ weighting matrix (akin here to an inverse correlation matrix). Based on this error model, the likelihood function for the unknowns comes from the multivariate normal probability distribution:

$$L(\boldsymbol{\theta}, \sigma^2) = (2\pi)^{-n/2} \left|\sigma^2 \boldsymbol{\omega}^{-1}\right|^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{\theta}))^T \boldsymbol{\omega} (\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{\theta}))\right]. \tag{5.22}$$

Substituting in the weighted sum of squares function of Eq. (5.3) yields

$$L(\boldsymbol{\theta}, \sigma^2) = (2\pi)^{-n/2} \left|\sigma^2 \boldsymbol{\omega}^{-1}\right|^{-1/2} \exp\left[-\frac{1}{2\sigma^2} S(\boldsymbol{\theta})\right], \tag{5.23}$$

and taking the logarithm gives

$$\log L(\boldsymbol{\theta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2} n \log \sigma^2 - \frac{1}{2}\log\left|\boldsymbol{\omega}^{-1}\right| - \frac{1}{2\sigma^2} S(\boldsymbol{\theta}). \tag{5.24}$$

Finally, the expression can be simplified by dropping those terms that do not depend on $\boldsymbol{\theta}$ or

σ^2:

$$\log L(\boldsymbol{\theta}, \sigma^2) \propto -\frac{1}{2} n \log \sigma^2 - \frac{1}{2\sigma^2} S(\boldsymbol{\theta}). \tag{5.25}$$

From Eq. (5.25), it is clear that the value of $\boldsymbol{\theta}$ that maximizes the likelihood function is the value that minimizes $S(\boldsymbol{\theta})$, the weighted sum of squares function. Thus, when a multivariate normal model is used for the errors, the least-squares estimator is a maximum likelihood estimator.

Now consider the Bayesian posterior distribution for $\boldsymbol{\theta}$, which has the form

$$\log f(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{d}) = c + \log \pi(\boldsymbol{\theta}, \sigma^2) + \log L(\boldsymbol{\theta}, \sigma^2), \tag{5.26}$$

where c is a constant that does not depend on $\boldsymbol{\theta}$ or σ^2, and $\pi(\boldsymbol{\theta}, \sigma^2)$ is the prior distribution for the unknowns. The prior distribution for $\boldsymbol{\theta}$ and σ^2 will almost certainly be separable (meaning $\boldsymbol{\theta}$ and σ^2 are *a priori* independent), which results in

$$\log f(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{d}) = c + \log \pi(\boldsymbol{\theta}) + \log \pi(\sigma^2) + \log L(\boldsymbol{\theta}, \sigma^2), \tag{5.27}$$

where $\pi(\boldsymbol{\theta}, \sigma^2) = \pi(\boldsymbol{\theta})\pi(\sigma^2)$, and for notational simplicity the function $\pi(\cdot)$ is indicated by its arguments (i.e., $\pi(\sigma^2)$ is the prior distribution for σ^2, but $\pi(\boldsymbol{\theta})$ is a different function altogether). Given the same error model as above, the expression for the log likelihood, given by Eq. (5.25), can be substituted to obtain

$$\log f(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{d}) = c + \log \pi(\boldsymbol{\theta}) + \log \pi(\sigma^2) - \frac{1}{2} n \log \sigma^2 - \frac{1}{2\sigma^2} S(\boldsymbol{\theta}). \tag{5.28}$$

Now consider the "constrained uniform" prior distribution for $\boldsymbol{\theta}$, given by Eq. (2.5), which

results in

$$\log f(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{d}) = \begin{cases} c + \log \pi(\sigma^2) - \dfrac{1}{2} n \log \sigma^2 - \dfrac{1}{2\sigma^2} S(\boldsymbol{\theta}) & \text{if } \boldsymbol{\theta} \in \Omega, \\ \log 0 & \text{if } \boldsymbol{\theta} \notin \Omega, \end{cases} \tag{5.29}$$

where $\log 0$ indicates here that the posterior distribution does not have support for $\boldsymbol{\theta} \notin \Omega$. From Eq. (5.29), it should be evident that the value of $\boldsymbol{\theta}$ that maximizes the posterior distribution is $\hat{\boldsymbol{\theta}}$, the least-squares estimator, as long as $\hat{\boldsymbol{\theta}} \in \Omega$ (as long as the prior constraints do not exclude $\hat{\boldsymbol{\theta}}$). This means that under the Gaussian error model and as long as the prior distribution for $\boldsymbol{\theta}$ is either uniform or constrained uniform, the maximum posterior estimate will be the same as the least-squares estimate.

5.6 Summary

This chapter discusses a variety of generally applicable approaches that allow one to address the problem of model calibration and parameter estimation uncertainty for complex computer simulations. The classical nonlinear regression approach presented in Section 5.2 is well-established, and this section does not discuss any new research.

On the other hand, the Bayesian approach to model calibration is a relatively new and active research area. While the idea of using Bayesian inference for model calibration analysis is not new, the particular formulation of Section 5.3.1, which explicitly includes a Gaussian process surrogate, has not been presented in the literature. In addition, the extensions proposed in Sections 5.3.2 and 5.3.3 are intended to address the second research objective. The framework discussed in Section 5.4 is that proposed by Kennedy and O'Hagan (2001). Part of the purpose of Section 5.4 is to analyze the differences between Kennedy and O'Hagan's

formulation and that presented in Section 5.3.1, and this discussion directly addresses the third research objective (see Section 1.2).

CHAPTER VI

APPLICATIONS AND CASE STUDIES

This chapter provides five case studies for the exploration of the uncertainty analysis methods presented in this dissertation. Sections 6.1 and 6.2 present solutions to two hypothetical model validation "challenge" problems developed at Sandia National Laboratories (Dowding et al., 2008; Red-Horse and Paez, 2008). In each case, the challenge problem provides the analyst with one or more mathematical models, as well as a set of corresponding experimental data for validation. The objectives are to first use the experimental data to assess the validity of the given models, and to then use the given models to predict whether or not a critical system will meet a specified probabilistic reliability requirement.

The first challenge problem, which is addressed in Section 6.1, pertains to the validation of a simple model for one-dimensional heat transfer. The solution to this challenge problem illustrates the use of the statistical significance testing procedures discussed in Section 4.1.2. Kennedy and O'Hagan's model calibration framework, discussed in Section 5.4, is also compared to the conventional Bayesian calibration framework, discussed in Section 5.3.1, for the calibration and corresponding uncertainty quantification associated with the heat transfer model. Emphasis is also placed on developing a comprehensive quantification of the uncertainty present in the final assessment of the system performance.

The second challenge problem (Section 6.2) pertains to the validation of a linear model for predicting the dynamic response of a three-degree-of-freedom subsystem which contains a "weak" nonlinearity. The analysis of this challenge problem illustrates two cases of a possible

classification of validation inference: fully characterized and partially characterized experiments. This analysis also makes extensive use of Gaussian process surrogate models, as well as a random sampling approach for correlated, non-normal variables (Section 6.2.2).

Sections 6.3 and 6.4 both present applications of the Bayesian calibration methodology discussed in Section 5.3 to real-world modeling and simulation projects at Sandia National Laboratories. In both cases, Gaussian process surrogate models are used to enable Bayesian calibration inference. The application of Section 6.3 deals with a small amount of data and a relatively large number of parameters (12), while the application of Section 6.4 deals with the case in which the system response is highly multivariate. Section 6.4 also illustrates the use of the Bayesian calibration extensions proposed in Sections 5.3.2 and 5.3.3.

Section 6.5 proposes a "top-down" approach to the calibration of hierarchical simulations. To illustrate and assess the viability of the approach, it is applied to data from Sandia National Laboratories pertaining to the behavior of a system of nonlinear bolted joints.

6.1 Model validation challenge problems: thermal application

This section addresses the thermal validation challenge problem (Dowding et al., 2008) developed at Sandia National Laboratories, which is a hypothetical problem that presents the analyst with several pieces of validation data and a corresponding mathematical model. The first objective put forth by the challenge problem is to use material characterization data to estimate a probabilistic model for the physical properties that are inputs to the mathematical model. The second and third objectives involve assessing the model's accuracy based on available experimental data (model validation). The final objective is to use the model to predict whether or not a specified regulatory requirement will be met.

The physical process under consideration is one-dimensional transient heat conduction

through a slab of material, as illustrated in Figure 6.1. The mathematical model that is to be used to model the process is the truncated infinite series solution given by

$$T(x,t) = T_i + \frac{qL}{k}\left\{\frac{(k/\rho C)t}{L^2} + \frac{1}{3} - \frac{x}{L} + \frac{1}{2}\left(\frac{x}{L}\right)^2 - \frac{2}{\pi^2}\sum_{i=1}^{6}\frac{1}{i^2}\exp\left[-i^2\pi^2\frac{(k/\rho C)t}{L^2}\right]\cos\left(i\pi\frac{x}{L}\right)\right\}, \quad (6.1)$$

where T_i is the initial temperature (here $25°$C), k is the thermal conductivity of the material, ρC is the volumetric heat capacity, t is time, L is the length of the slab, and q is the applied heat flux. One of the assumptions of this particular model is that the material properties, k and ρC, are treated as constants, when in reality they are not necessarily independent of temperature. The empirical temperature-dependence of the material properties is discussed in Section 6.1.1.

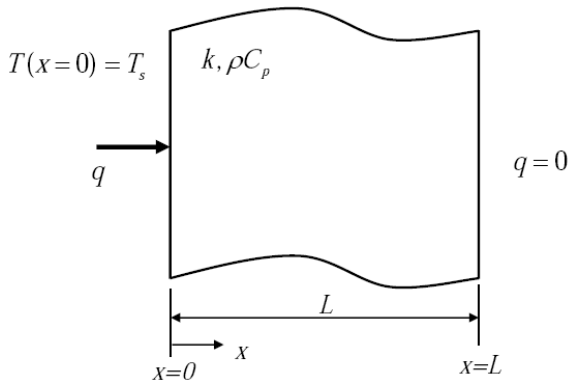


Figure 6.1: Schematic of the heat conduction problem (from Dowding et al., 2008)

6.1.1 Use of material data and mathematical model

The first challenge problem objective is to use data from the "material characterization experiments" to characterize a probabilistic model for the material properties k and ρC, which are inputs to the heat-transfer model given by Eq. (6.1). Several challenges arise at this stage, as it is quickly apparent that there is a relationship between the thermal conductivity, k, and temper-

ature. However, the inclusion of a temperature-dependent material model requires modification of the given analytical heat transfer solution (e.g., implementation of an iterative solution scheme), and doing so is decidedly inconsistent with the purpose of the challenge problem's validation activities, which are to assess the accuracy of and make predictions with an inherently "flawed" model (one that ignores the temperature-dependence of the material properties).

This analysis refrains completely from adjusting the heat-transfer model to account for temperature dependency of the thermal conductivity. Each evaluation of the heat transfer model is thus performed by simply evaluating Eq. (6.1) for a particular (constant) realization of k and ρC. Not only is this consistent with the "code verification" discussed in the challenge problem description (Dowding et al., 2008), but by treating the model as a "black box," the methods presented are more generally applicable.

The procedure used here is to characterize each of the material properties using (independent) probability distributions. The variance of ρC is estimated directly from the data, whereas the variance of k is estimated using a simple linear regression model of the material characterization data as a function of temperature (see Figure 6.2), which allows for the isolation of the variance related to specimen-to-specimen variability. In each case, the mean values are estimated directly from the data. The normal probability distribution model is employed for both properties. The normal distribution is used partly for parsimony: a more complex probabilistic model may be less interpretable and also may be at risk of over-characterizing the actual property variation. But the normal distribution is also chosen because the data do not strongly suggest otherwise. This is confirmed by two well-known tests for normality, the Lilliefors test (Lilliefors, 1967) and the Shapiro-Wilk test (Shapiro and Wilk, 1965). Each test is applied to the data for ρC and the regression residuals for k. The results are included in Ta-

ble 6.1, suggesting that in no case do the data provide strong evidence against the assumption of normality.

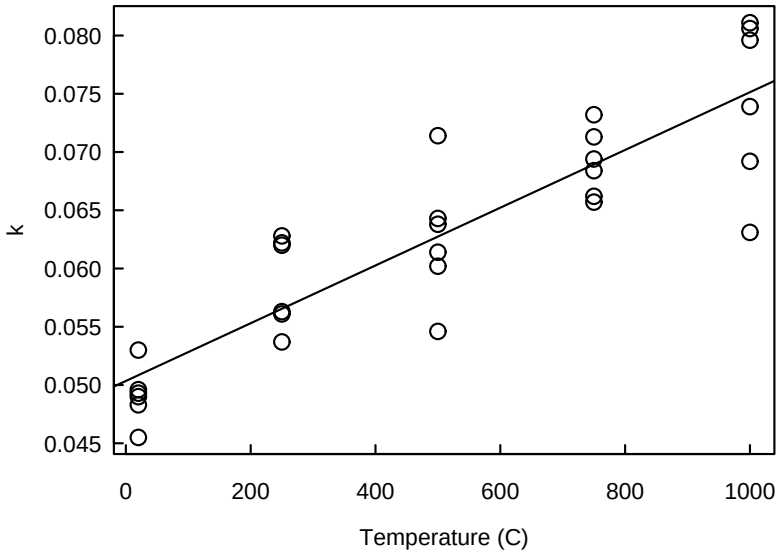


Figure 6.2: Linear regression of k on temperature

Additionally, for the validation exercises, it is important to obtain accurate estimates of the material property means that will allow for relevant predictions at each experimental configuration. This consideration suggests the use of a censoring scheme when estimating the property distributions for the validation exercises. Only property measurements corresponding to temperatures less than or equal to $500°$C are used for analysis of the "ensemble validation" data, and only property measurements made at temperatures less than or equal to $750°$C are used for analysis of the "accreditation" data.

The resulting statistics of the material properties are given in Table 6.1 (for the high data level).

6.1.2 Model validation

One of the primary objectives for this challenge problem is to illustrate an approach whereby the provided hypothetical experimental observations are used to develop a quantitative assess-

Table 6.1: Statistics of material property data (high data level) and p-values for normality tests

Data set	n	Property	Mean	Stand. Dev.	Shapiro Test (p-value)	Lilliefors Test (p-value)
$T \leq 500$°C	18	k (W/m°C)	0.0569	0.00435	0.48	0.24
		ρC (J/m^3°C)	3.92E5	3.69E4	0.79	0.92
$T \leq 750$°C	24	k	0.0599	0.00398	0.70	0.31
		ρC	3.94E5	3.76E4	0.45	0.84
All T	30	k	0.0628	0.00470	0.51	0.49
		ρC	3.94E5	3.63E4	0.44	0.58

ment of the validity of the heat transfer model given by Eq. (6.1). Because multiple repeated experiments are available and part of the objective is to quantify the amount of confidence that is developed using different amounts of data, a statistical significance testing approach is presented here. The significance testing results will be used to assess whether or not the observed differences between the predictions and observations might be attributable to inherent variation or lack of data; the "power" of the tests in reaching a conclusion will also be discussed.

Following Section 4.1.2, the first step when applying significance testing for model validation is to select the response feature or features of interest. For this application, the system response is the temperature of the device over time, and for each of the ensemble validation experiments, this response is measured at ten time instants. While it would be possible to treat the response as a ten-dimensional multivariate quantity and apply multivariate testing methods (see the second part of Section 4.1.2; in fact McFarland and Mahadevan, 2008, present a multivariate significance testing approach for this very challenge problem), only a scalar response quantity is considered here.

The primary reason for considering a scalar response quantity is that such an approach is more relevant to the intended use of the model. This is because the purpose of the given heat transfer model is to predict whether or not the specified regulatory requirement will be met, and this regulatory requirement is defined in terms of the device temperature at $t = 1000$ seconds

only. By constructing the model validation assessment in terms of the device temperature at $t = 1000$ seconds, the validation assessment is focused on the model's intended use, and unnecessary requirements are not placed on the model.

It should also be noted that for this particular problem, very little information is lost by considering a scalar response feature. By analyzing the covariance matrix of the full ten-dimensional response, it is clear that the device temperatures at each of the ten time instants are almost linearly dependent. This observation suggests that even if model validity were defined in terms of the model's accuracy at all ten time instants, very little would be gained by formulating the problem in terms of a multivariate response quantity.

The actual heat-transfer model provided in the thermal challenge problem, Eq. (6.1), is computationally trivial to evaluate, making it straightforward to obtain the model's output distribution using simple Monte Carlo simulation (recall that two of the model's input parameters are random variables). However, in many cases, the model may be costly and/or time-consuming to evaluate, making it necessary to use approximation methods to estimate the model's output distribution. In such cases, efficient approximation methods such as the Stochastic Response Surface Method (Isukapalli et al., 1998) or Gaussian process modeling (Chapter III) can be used to estimate the model's output distribution, in which case the validation analysis described in the remainder of this section would be carried out in the same manner.

The thermal challenge problem provides hypothetical validation data for two domains, the "ensemble validation" domain and the "accreditation" domain. The experiments conducted inside the accreditation domain correspond to larger values of applied heat flux, q. Additionally, the "ensemble validation" domain consists of experiments conducted at four different configu-

rations (different combinations of q and L; see Figure 6.3). Since each of these configurations corresponds to a different statistical population, the data corresponding to each system configuration are considered separately.

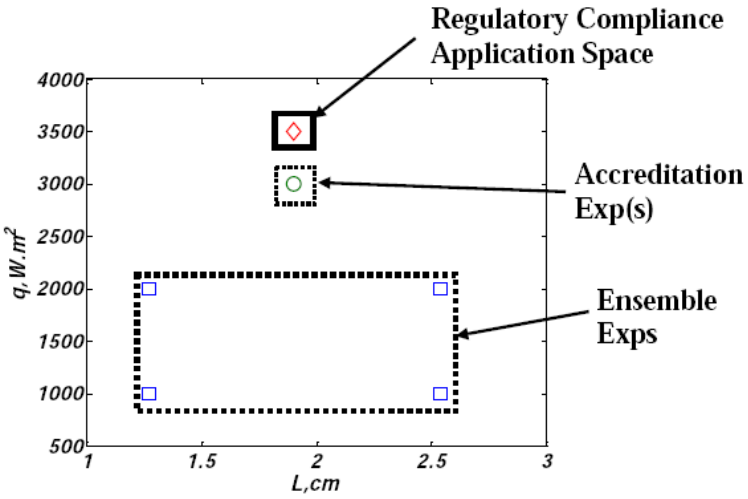


Figure 6.3: Parameter space describing the ensemble, accreditation, and application domains (from Dowding et al., 2008)

For this challenge problem, the analyst is also asked to address the effect of data quantity by reporting validation assessments for low, medium, and high amounts of data. These correspond to (in addition to varying amounts of material characterization data) one, two, and four repeated experiments being available at each ensemble configuration, and one, one, and two experiments being available at the accreditation configuration.

The agreement between the model output and the validation data is illustrated graphically in Figure 6.4 (for the ensemble validation domain only). The value of the response feature is plotted on the x-axis, and the y-axis is used to distinguish the four ensemble validation configurations. For each case, the model output mean is compared to four repeated experimental observations available at the high data level. In addition, 95% confidence intervals for the mean of the experimental populations (computed assuming the variance is known and equal to the model output variance) are also plotted.

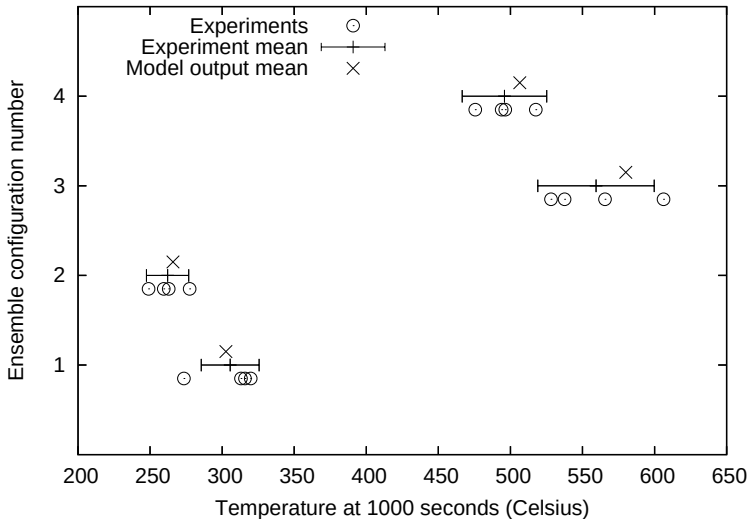


Figure 6.4: Comparison of model output and experimental observations for each of the four ensemble validation configurations

The most commonly used significance test is the test for equality of means, and this is the test outlined in Section 4.1.2. Some readers may be concerned that in this case, such a test is only partially relevant to the intended use of model. Since the heat transfer model is to be used to compute a probability level of the response, the entire model output distribution is of interest, not just the mean. The problem is that validating higher order moments (such as the variance) or entire distributions will require much larger sample sizes than $n = 4$. If the number of experimental replicates were large enough, it would certainly be possible to construct more detailed tests; for example, a Kolmogorov-Smirnov test (c.f. Papoulis and Pillai, 2002) could be used to test the experimental data against the entire model output distribution. However, given the present limitation on sample size, a test of agreement between the *location* of the model output and the experiments is the best that can reasonably be expected (in fact, as shown shortly, due to the large variance of the system response, the given validation data do not even support strong conclusions regarding the model output mean).

As discussed in Section 4.1.2, univariate tests for the equality of means may be based on

either the t-statistic or the z-statistic, depending on whether or not the variance is estimated using the observed samples. The z-statistic is used here for three reasons: first, the low data level ($n = 1$) does not provide enough samples to estimate a variance. Second, the z-test is more powerful than the t-test (meaning it has a better chance of rejecting a false hypothesis) because there is no uncertainty in the variance. Finally, the model output distribution provides a more direct estimate of the variance. It is appropriate to equate the variance of the experiments to the variance of the model output in this case because the dominant mechanisms governing the variability within the two population are the same. Specifically, the variation among the experimental measurements is due to specimen-to-specimen material property variability, and it is this same material property variation that is modeled probabilistically and propagated through the simulation model to obtain its output distribution.

Thus, the z-test for equality of means is applied, and for each case the variance, σ^2, is set equal to the variance of the corresponding model output. The test is conducted at each of the five validation configurations shown in Figure 6.3, and the test is repeated for the low, medium, and high data levels. The achieved significance level (p-value) of each test is tabulated in Table 6.2. The power of each test in detecting a difference $|\mu - \mu_0|$ equal to one standard deviation is computed as discussed in Section 4.1.2 and tabulated in Table 6.3.

Table 6.2: Achieved significance levels (p-values) for model validation significance tests for equality of means

Config. / Data Level	Low	Medium	High
Ensemble 1	0.52	0.41	0.77
Ensemble 2	0.67	0.81	0.62
Ensemble 3	0.72	0.33	0.31
Ensemble 4	0.30	0.30	0.48
Accreditation	0.84	0.84	0.33

Note that the p-value is the lowest level of significance at which the data suggest the rejection of the null hypothesis. Typically, a p-value less than 0.05 is considered significant evidence

Table 6.3: Power of significance tests in detecting a difference in means equal to one standard deviation

Config. / Data Level	Low	Medium	High
Ensemble 1	0.17	0.29	0.52
Ensemble 2	0.17	0.29	0.52
Ensemble 3	0.17	0.29	0.52
Ensemble 4	0.17	0.29	0.52
Accreditation	0.17	0.17	0.29

against the null. The results of the significance tests indicate that for each case the experimental data offer insufficient evidence to suggest that the locations of the model predictions are different than those of the observations.

The power values indicate that for the ensemble scenarios, the tests have less than a 50% chance of detecting a difference of means equal to one standard deviation when one and two experimental observations are available, and approximately a 50% chance of detecting the difference with four observations. Given that detecting a difference of one standard deviation is of interest, these results indicate that the low and medium data levels are insufficient to make a strong statement about model validity, and even the high data level is only marginally useful for such inference.

Taken together though, the results based on the high data level might be used to infer that within the domain of the ensemble experiments, the accuracy of the model has been established with a moderate amount of confidence to be better than one standard deviation. This conclusion might be reached since four independent validation tests, each with a power of 52%, were incapable of rejecting H_0 (in fact, the probability that at least one of these tests would detect the specified difference is 0.93). However, if different physical behavior is thought to be important in the accreditation domain, then more experiments are probably needed to make a statement about the model's predictive capability in that domain (even for the high data level, with 2 repeated experiments, the power of the significance test is only 0.29).

6.1.3 Calibration of the heat transfer model

Even though the above validation assessment did not suggest that the given mathematical model is not suitable for its intended use, it is still possible to take account of the observed system response data for the purpose of model calibration, and doing so is only expected to improve the predictive capability of the model. This particular case study provides a unique opportunity to illustrate some of the model calibration approaches outlined in Chapter V. First, the various configurations at which experimental data are available are quantified by two continuous scenario-descriptor variables, $\boldsymbol{s} = (q, L)$, which means that there exists the opportunity to employ the bias-correction formulation proposed by Kennedy and O'Hagan (2001) and discussed in Section 5.4. Second, the given mathematical model is not capable of accounting for the known temperature-dependency present in the thermal conductivity, k, which provides an opportunity for some creativity in the calibration process.

In order to exploit the interesting properties of this particular case study, two different calibration and uncertainty analysis approaches will be illustrated. For the first approach, the Bayesian parameter estimation formulation described in Section 5.3.1 is implemented, and the linear temperature-dependence model for k is used in a simple manner to account for the fact that the various configurations correspond to different temperature levels (although the given mathematical model for the temperature response is still not modified to take account of the temperature dependence). The second approach will illustrate the use of the bias-correction formulation proposed by Kennedy and O'Hagan (2001).

Approach 1

For this approach, the Bayesian calibration framework described in Section 5.3.1 is adopted. First, consider the classification of the heat transfer model's input parameters into the calibration inputs and the variable inputs. Clearly, there are two variable inputs, q and L, which describe the scenario and vary from one experiment to the next; this gives $\boldsymbol{s} = (q, L)$.

For this model, the number of possible adjustable "calibration parameters" is small, but several options still exist. In particular, the material properties, k and ρC, although subject to parametric variability, may still be useful as calibration parameters. One possibility is to consider the location parameters governing the probability distributions of k and ρC as calibration parameters. Since ρC does not depend on temperature and is treated using a normal distribution, the mean of this distribution, $\mu_{\rho C}$, will be taken as one calibration parameter.

It would also be possible to take the μ_k as the second calibration parameter, and this is in fact what is done in approach 2. However, for this approach, which does not include a scenario-dependent bias-correction factor, it would be nice to acknowledge the fact that the temperature-dependence for k suggests that different values of k might be appropriate as model inputs for the five different configurations at which the calibration data are available. To allow for this, one possibility is to consider again the linear model for k on temperature introduced in Section 6.1.1, which is written here as

$$k_i = \beta_0 + \beta_1 T_i + \varepsilon_{i,k}, \tag{6.2}$$

where T is temperature, in Celsius[1]. From the model of Eq. (6.2), the randomness in k is

[1]The special notation $\varepsilon_{i,k}$ is used here to emphasize that this is not the same ε that appears in the probabilistic model defining the calibration analysis, Eq. (5.11)

captured by the variance of ε_k, the "location" of k by β_0, and the relationship to temperature by β_1. This model suggests that the intercept β_0 might be useful as the locational calibration parameter for k. Thus, the calibration parameters are taken to be $\boldsymbol{\theta} = (\beta_0, \mu_{\rho C})$. The slope, β_1, is held fixed at the nominal value suggested by the material property data, which is 2.48×10^{-5}.

One difficulty is to find a way to incorporate the temperature-dependent model of Eq. (6.2) into the calibration analysis without modifying the given mathematical model, which takes as input only one fixed value of k. A simple approach is adopted here in which a "representative" temperature value, T, is determined for each configuration of the system for which calibration data are available. The representative temperature for each configuration is taken to be the average experimentally observed temperature for that configuration, at time $t = 500$ seconds (this nominal time is chosen because it is the midpoint). Thus, the model predictions $G(\cdot, \cdot)$ for a given realization of the calibration parameter β_0 will involve different values of k for each configuration, depending on the representative temperatures, T.

It is also necessary to decide what particular system response quantity of the heat-transfer model will be used for calibration. Since the objective is to enhance the predictive capability of the model, it is important to choose a quantity that is relevant to the intended use of the model for predictions in the application domain. Although it would be possible to calibrate using the measured temperature responses at each time instance, doing so would significantly increase the complexity of the analysis. In addition, the temperature responses show such strong linear dependency that it is unlikely that much additional information would be obtained. Since the intended use of the model is to make predictions at $t = 1000$ seconds only, the calibration analysis will consider only the scalar response $T(t = 1000)$ (recall that the validation analysis of Section 6.1.2 considered the same scalar response quantity).

The usual assumption of normally distributed errors is taken: $\varepsilon_i \sim N(0, \sigma_i^2)$. For this analysis, these errors are envisioned to be dominated by observation variability, and the variance σ_i^2 corresponding to each configuration is treated as a known constant and set equal to the sample variance of the experimental observations for that particular configuration. Note that there are four repeated experimental observations at each of the four ensemble configurations, and two repeated observations for the accreditation configuration, for a total of $n = 18$ data points for the calibration.

Finally, a prior distribution is needed for the calibration parameters, $\boldsymbol{\theta}$. The vague prior distribution of Eq. (2.2) is used, $\pi(\boldsymbol{\theta}) \propto 1$, so that the calibration results will be dominated by the data. The posterior distribution for $\boldsymbol{\theta}$ is constructed using Markov Chain Monte Carlo simulation, as discussed in Section 2.3. What is ultimately of interest, however, is the results when the calibrated thermal model is used to make predictions for the untested application domain. The implementation of the calibrated thermal model in the application domain is discussed in Section 6.1.4, along with a comprehensive uncertainty analysis.

Approach 2

For the second approach, the Bayesian calibration framework developed by Kennedy and O'Hagan (2001) and discussed in Section 5.4 is illustrated. Kennedy and O'Hagan's framework provides for a scenario dependent "model inadequacy function," which describes the systematic bias between the model predictions and observations. The thermal challenge case study provides an excellent opportunity to explore this approach because the calibration data are available five different configurations, each defined in terms of the variable inputs $\boldsymbol{s} = (q, L)$. This model inadequacy function is formulated with the intention that it can be used to extrapolate the value of the simulator bias to the untested, application domain. As mentioned above,

since the given heat transfer model is computationally inexpensive, no surrogate is needed, and the actual heat transfer model is used in Eq. (5.21) in lieu of the unknown function $\eta(\cdot,\cdot)$.

The first step is to specify the calibration inputs. As was done for approach one, the locational distribution parameters for the material properties will be taken as calibration parameters. However, for this approach, the value of the thermal conductivity, k, will not allowed to be temperature dependent: the distribution for k will simply be taken as $k \sim N(\mu_k, \sigma_k^2)$, where σ_k^2 is still the variance after regressing k on temperature, but the expected value is no longer considered as a function of temperature. This simplification is taken in the hopes that the model inadequacy function $\delta(\cdot)$ can capture the same effect that is achieved by allowing k to depend on temperature. Thus, for this approach, the calibration parameters are taken to be $\boldsymbol{\theta} = (\mu_k, \mu_{\rho C})$.

The Gaussian process for the model inadequacy, $\delta(\cdot)$, is constructed using training points which are observation of the bias $d_i - G(\boldsymbol{\theta}, \boldsymbol{s}_i)$ for various configurations $\boldsymbol{s}$. First, note that this bias depends on the calibration parameters, $\boldsymbol{\theta}$, which are in fact unknowns. However, the purpose of the model inadequacy function as developed by Kennedy and O'Hagan (2001) is to capture the model bias as a function of the scenario inputs, $\boldsymbol{s}$, not the calibration inputs, $\boldsymbol{\theta}$. The approach taken here to eliminate the dependency on $\boldsymbol{\theta}$ is to develop $\delta(\cdot)$ based on a nominal value of the calibration inputs, which is taken to be the mean values of the material properties, based on the material characterization data (a more complex alternative is given by Kennedy and O'Hagan, 2000b).

It is also apparent that since the observed values of the system response d_i contain variability, the training data for $\delta(\cdot)$ also contain variability (or observation uncertainty). Thus, the Gaussian process representation must be developed accordingly, using the methods presented

in Section 3.5. The observation errors, $\lambda_{exp,i}$, are treated as known values and set equal to the corresponding sample variances of $\boldsymbol{d}$ for each configuration, $\boldsymbol{s}_i$.

A mean function must also be chosen for $\delta(\cdot)$. The role of the mean function in Gaussian process modeling can be trivial in some cases, but it can also become important when the Gaussian process is used to predict the value of the process beyond the range of observed data, particularly when systematic trends are present in the data (see Section 3.2). Based on observed trends for the heat-transfer model, and also to avoid over-parametrization, a linear mean function is adopted for $\delta(\cdot)$:

$$E[\delta(\boldsymbol{x})] = \beta_0 + \beta_1 q + \beta_2 L, \tag{6.3}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ are coefficients to be estimated (not to be confused with the coefficients introduced in approach one via Eq. (6.2)).

The parameters governing the Gaussian process $\delta(\cdot)$ are treated as known constants, and before proceeding with the calibration, their values must be estimated based on the data. As discussed in Section 3.5.1, traditional Maximum Likelihood Estimation can be problematic when observation errors are present. To overcome this problem, the restricted maximum likelihood approach is adopted, in which case the parameters governing the mean and covariance functions are estimated via the minimization of the function given by Eq. (3.30).

The model inadequacy function is illustrated in Figure 6.5, conditional on the observed data. While $\delta(\cdot)$ is a function of both q and L, it is plotted here as a function of q for L fixed at 1.9 cm. The training points are also included for illustration (the average bias is shown for each, which is based on four repeated observations for the ensemble configurations and two for the accreditation configuration), and they are each labeled with the corresponding value of

L to emphasize that these training points do not all share the same value of L for which $\delta(\cdot)$ has been plotted.

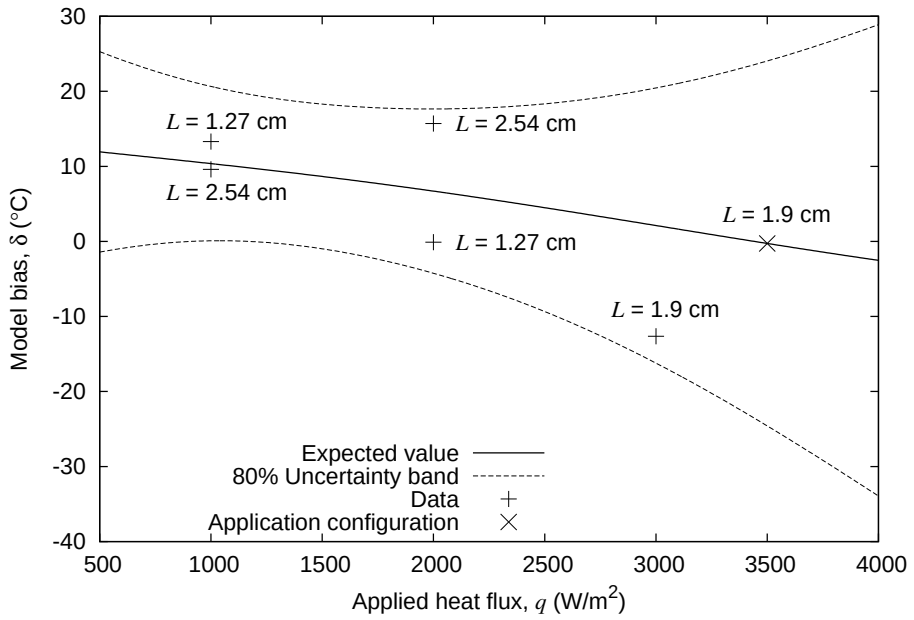


Figure 6.5: Conditional expected value and uncertainty bands for the model inadequacy function, plotted as a function of applied heat flux for $L = 1.9$ cm

Otherwise, the prior distribution for $\boldsymbol{\theta}$, the response quantity, and the distribution of the errors ε_i are all the same as in approach one. The only differences between the two approaches are the inclusion of the model inadequacy function for approach two and the use of a temperature-dependent model for k in approach one. As with the first approach, Markov Chain Monte Carlo sampling is used to construct the posterior distribution for $\boldsymbol{\theta}$.

Section 6.1.4 discusses how these results can be used to estimate the probability of failure of a device in the application domain, along with confidence bounds for the assessment.

6.1.4 Assessment of regulatory compliance

The fourth and final objective for the thermal challenge problem asks the analyst to provide both an assessment as to whether or not the device will meet regulatory requirements and a

statement of confidence in this assessment. The probability of failure for the device in the application configuration is defined as

$$p_f = P[T(t = 1000\,\mathrm{s}) > 900^\circ\,\mathrm{C}], \tag{6.4}$$

and regulatory compliance is said to be achieved if the probability of failure is less than 0.01. The analysts are also asked to provide a (preferably quantitative) "level of confidence" about whether or not the regulatory condition will be met.

This objective is addressed below separately for each of the two calibration approaches. The calibrated models are used to predict the probability of failure, and in each case a variety of uncertainty sources are taken into account to construct a representation of the uncertainty in this prediction.

Approach 1

When using the results of Bayesian calibration for probability of failure prediction it is important to maintain the distinction between aleatory variability (in this case characterized by the random variables k and ρC) and the residual uncertainty in the calibration parameters represented by the posterior distribution $f(\boldsymbol{\theta} \mid \boldsymbol{d})$. In this case, each of the two calibration parameters represent locational distribution parameters governing the variability in k and ρC.

It is easy to see that the probability of failure, *conditional* on the distribution parameters governing the random variables k and ρC can be expressed as

$$p_f \mid \beta_0, \beta_1, \sigma_k^2, \mu_{\rho C}, \sigma_{\rho C}^2 = \int_\Omega f(k, \rho C)\, dk\, d\rho C, \tag{6.5}$$

where Ω is the failure region, which is given by

$$G(k, \rho C, \boldsymbol{s}^*) > 900^\circ \, \mathrm{C}, \tag{6.6}$$

$f(k, \rho C)$ is the joint probability density function for k and ρC, which are treated as independent random variables with probability distributions $k \sim N(\beta_0 + \beta_1 T^*, \sigma_k^2)$ and $\rho C \sim N(\mu_{\rho C}, \sigma_{\rho C}^2)$; $\boldsymbol{s}^* = (q = 3500, L = 0.019)$, which defines the application configuration; and T^* is the representative temperature for the application region.

There is a small problem in determining T^*, because no experimental observations are available for the application configuration. The procedure used here is to apply the thermal model with a nominal value of k (the mean of the material characterization data) and the given value of $\mu_{\rho C}$ to predict the temperature at $t = 500$ seconds and take this prediction as T^*. Note that with this procedure T^* depends on $\mu_{\rho C}$.

With the expression of Eq. (6.5), it is possible to define the posterior distribution of p_f, which represents the uncertainty in the failure probability based on the residual uncertainty after calibration of the calibration parameters $\boldsymbol{\theta} = (\beta_0, \mu_{\rho C})$. However, this notion can be taken a step further: note that the variances σ_k^2 and $\sigma_{\rho C}^2$ are not known exactly, but are instead estimated based on the finite samples provided via the material characterization data. Using Bayesian inference, it is also possible to incorporate this uncertainty into the uncertainty representation for p_f.

Since ρC is independent of temperature, the probability model $\rho C \sim N(\mu_{\rho C}, \sigma_{\rho C}^2)$ has been used. If $\sigma_{\rho C}^2$ and $\mu_{\rho C}$ are given the standard reference prior $\pi(\mu_{\rho C}, \sigma_{\rho C}^2) \propto 1/\sigma_{\rho C}^2$, then the marginal posterior distribution for the variance in light of the material characterization data

$\boldsymbol{d}_{\rho C} = (\rho C_1, \ldots, \rho C_{30})$ is given by (Lee, 2004):

$$\sigma^2_{\rho C} \mid \boldsymbol{d}_{\rho C} \sim S\chi^{-2}_{n-1}, \tag{6.7}$$

which is a multiple of what is known as an inverse chi-squared distribution, where $S = \sum_{i=1}^{n}(\rho C_i - \bar{\rho C})^2$.

Recall that the variance for k derives from the linear model of Eq. (6.2), where the $\varepsilon_{i,k}$ are taken to be i.i.d. normal with zero mean and variance σ^2_k. Given the usual reference prior $\pi(\beta_0, \beta_1, \sigma^2_k) \propto 1/\sigma^2_k$, the marginal posterior distribution for the variance is (Lee, 2004)

$$\sigma^2_k \mid \boldsymbol{d}_{k,T} \sim S_{ee}\chi^{-2}_{n-2}, \tag{6.8}$$

where $S_{ee} = S_{yy} - S^2_{xy}/S_{xx}$, $S_{yy} = \sum(k_i - \bar{k})^2$, $S_{xx} = \sum(T_i - \bar{T})^2$, and $S_{xy} = \sum(T_i - \bar{T})(k_i - \bar{k})$.

Now, the posterior distribution for p_f can be constructed such that it accounts for the residual uncertainty in the calibration parameters, as well as the uncertainty in the material property variances due to their being estimated based on finite data. This posterior can be constructed using a two-loop sampling scheme in which the outer loop generates samples of β_0, $\mu_{\rho C}$, σ^2_k, and $\sigma^2_{\rho C}$ according to their posterior distributions (this sampling is achieved here via MCMC). For each such realization, the inner loop estimates the corresponding conditional failure probability, defined by Eq. (6.5). The result is a list of samples of p_f that constitute the posterior uncertainty distribution for the failure probability. The resulting uncertainty distribution for p_f is illustrated in Figure 6.6 below.

The expected value of p_f, which is taken here to be the mean of its posterior distribution, is

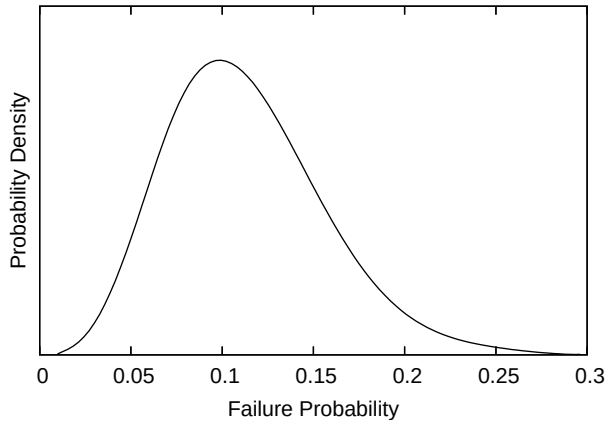


Figure 6.6: Uncertainty distribution for p_f based on calibration approach number one

0.11, which is significantly higher than the regulatory requirement specification of 0.01. One possible quantification of the level of confidence that the regulator condition will be met is given by the fraction of the uncertainty distribution for p_f that is greater than 0.01:

$$\int_{p_f > 0.01} f(p_f \mid \boldsymbol{d})\, dp_f, \tag{6.9}$$

which is found to be 0.9999 (based on 20,000) samples. Thus, one interpretation of this result is that there is a 99.99% level of confidence that the regulatory condition specified by Eq. (6.4) will not be met.

Approach 2

As mentioned above, when using Bayesian calibration results for probability of failure prediction, it is important to differentiate between aleatory (true variability) and epistemic (lack of knowledge) uncertainties. The probability of failure condition defined by Eq. (6.4) is the result of specimen-to specimen variability manifested through the treatment of the material properties k and ρC as random variables (aleatory uncertainty). However, in the Bayesian calibration analysis, the calibration parameters $\boldsymbol{\theta} = (\mu_k, \mu_{\rho C})$ are treated as random variables, but this is

an epistemic uncertainty, and must be considered separately because it does not contribute to actual variability of the response.

As with approach one, a conditional failure probability is first defined. In this case, this failure probability is conditional on the material property distribution parameters, as well as the model bias:

$$p_f \mid \mu_k, \sigma_k^2, \mu_{\rho C}, \sigma_{\rho C}^2, \delta = \int_\Omega f(k, \rho C)\, dk\, d\rho C, \tag{6.10}$$

where Ω is the failure region, given by

$$G(k, \rho C, \boldsymbol{s}^*) + \delta > 900^\circ\,\text{C}, \tag{6.11}$$

and $f(k, \rho C)$ is the joint probability density function for k and ρC, which are treated as independent random variables with probability distributions $k \sim N(\mu_k, \sigma_k^2)$ and $\rho C \sim N(\mu_{\rho C}, \sigma_{\rho C}^2)$. This conditional failure probability can be computed with simple Monte Carlo simulation.

As with approach one, an uncertainty distribution for p_f will be developed that accounts not only for the residual uncertainty in the calibration parameters, but also for the residual uncertainty in the material property variances. Although the temperature-dependent model for k is not used in this approach for obtaining model predictions, such a model should still be acknowledged when estimating the variance in k. As such, the posterior distribution for σ_k^2 given by Eq. (6.8) is used, as with approach one. The previously used posterior distribution for $\sigma_{\rho C}^2$, given by Eq. (6.7), is also employed.

The residual uncertainty in the model bias, δ is also accounted for. At each iteration of the outer loop, δ is sampled from its posterior distribution, which in this case is given by

$$\delta(\boldsymbol{s}^*) \mid \boldsymbol{d} \sim N(-0.27, 19.0)^\circ\,\text{C}. \tag{6.12}$$

As described in the corresponding discussion for approach one, the posterior uncertainty distribution for p_f is constructed using a two-loop sampling scheme. The resulting distribution for p_f is illustrated in Figure 6.7. The expected value of p_f is 0.19, and the confidence level that the regulatory requirement will not be met, given by Eq. (6.9), is 99.29%.

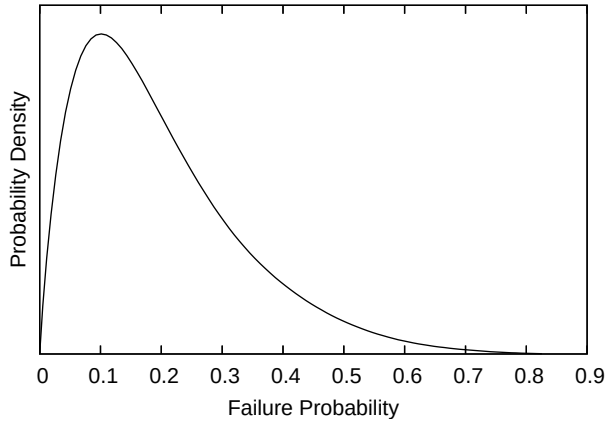


Figure 6.7: Uncertainty distribution for p_f based on calibration approach number two

There is clearly much more uncertainty in this estimate of p_f than there was with approach one. This is most likely attributable to the introduction of the model inadequacy function, $\delta(\cdot)$. Not only does the model bias at the application configuration contain a significant amount of uncertainty (see Eq. (6.12)) that contributes to uncertainty in model predictions, but the presence of $\delta(\cdot)$ within the calibration analysis as an uncertain term contributes additional uncertainty to the inference about the calibration parameters, $\boldsymbol{\theta} = (\mu_k, \mu_{\rho C})$.

In a sense, the uncertainty in the model inadequacy function manifests itself twice in the calibrated model predictions. Further, Kennedy and O'Hagan (2001) proposed that $\delta(\cdot)$ is not considered as a function of the calibration inputs, $\boldsymbol{\theta}$. However, in reality, it is highly unlikely that the model bias is independent of $\boldsymbol{\theta}$. Since the calibration procedure considers various different values of $\boldsymbol{\theta}$, it might make sense to attempt to account for the relationship between δ and $\boldsymbol{\theta}$. Such an approach might also help to reduce the "double-counting" of the uncertainty

associated with the model bias.

6.1.5 Probabilistic sensitivity analysis

As a final consideration, a simple variance decomposition is considered in order to explore the contributions of the various uncertainty sources. This will provide insight as to which uncertain variables (the calibration parameters, material property variances, or in the second approach, the model bias) contribute the most to the uncertainty in the resulting failure probability inference (i.e., the failure probability uncertainty distributions shown in Figures 6.6 and 6.7).

It is also important to mention here that the analysis discussed in this section is not analogous to the computation of "sensitivity factors" via the First Order Reliability Method (FORM; c.f. Haldar and Mahadevan, 2000). The sensitivity factors computed using FORM represent the sensitivity of the response to the random variables themselves (in this case k and ρC), as opposed to uncertainty that is attributable to lack of knowledge about, for example, the *parameters* governing the distributions of k and ρC.

The contributions of a set of factors $\boldsymbol{X}$ to the uncertainty in a response Y (in this case p_f, as defined by Eq. (6.10)) can be quantified by ranking the factors according to $\mathrm{Var}(Y \mid X_i = x_i^*)$, which is the variance obtained by fixing X_i to its true value x_i^*. However, since the true value of X_i is not known, one possibility (Saltelli et al., 2004) is to compute the expectation over all possible values of X_i, which gives

$$\int \mathrm{Var}(Y \mid x_i) f_{X_i}(x_i)\, dx_i. \tag{6.13}$$

Unfortunately, for the present problem, the computation of this expectation requires a three-

loop sampling scheme (the innermost loop to compute p_f, the next loop to compute the variance of p_f, and the final loop to average this variance over possible values of X_i), which is overly expensive. Thus, the approximation used here is to compute $\mathrm{Var}(Y \mid X_i = x_i)$ for a nominal value (in this case the mean value) x_i.

Taking this approximation and following Saltelli et al. (2004), the sensitivity index for each factor is computed as

$$S_i = \frac{\mathrm{Var}(Y) - \mathrm{Var}(Y \mid x_i)}{\mathrm{Var}(Y)}. \tag{6.14}$$

(Note that the S_i do not need to sum to one, and that it is possible to have $\mathrm{Var}(Y \mid x_i) > \mathrm{Var}(Y)$, which results in a negative sensitivity index.) One problem with the above expression is that it is only appropriate when X_i is independent of the other factors. If a group of factors are dependent, then it is more appropriate to gauge their combined effect (not to be confused with an interaction effect), which can be estimated for two parameters as

$$S_{ij}^c = \frac{\mathrm{Var}(Y) - \mathrm{Var}(Y \mid x_i, x_j)}{\mathrm{Var}(Y)}, \tag{6.15}$$

where here the factors X_i and X_j are dependent on each other, but independent of the remaining factors.

For the calibration analyses presented in Section 6.1.3, the calibration parameters $\boldsymbol{\theta}$ are dependent on each other, but the remaining factors are independent. Thus, for approach one, the calibration parameters $\boldsymbol{\theta} = (\beta_0, \mu_{\rho C})$ are dependent on each other, so their effect will be computed using Eq. (6.15). Similarly, for approach two, $\boldsymbol{\theta} = (\mu_k, \mu_{\rho C})$ are dependent, and their effect will be computed in the same way (recall that for approach two, the model bias, δ, is independent of the calibration parameters).

The resulting sensitivity indices for each approach are tabulated in Tables 6.4 and 6.5. It is clear that for both approaches, the uncertainty in the material property variances, σ_k^2 and $\sigma_{\rho C}^2$, do not contribute significantly to the uncertainty in p_f. This is probably because the number of samples available in the material characterization experiments (30) is sufficient to characterize these variances.

Table 6.4: Sensitivity of p_f to uncertain variables for approach one

Factor(s)	Sensitivity index
$\beta_0, \mu_{\rho C}$	0.63
σ_k^2	0.039
$\sigma_{\rho C}^2$	-0.054

Table 6.5: Sensitivity of p_f to uncertain variables for approach two

Factor(s)	Sensitivity index
$\mu_k, \mu_{\rho C}$	0.77
σ_k^2	8.67×10^{-4}
$\sigma_{\rho C}^2$	-0.027
δ	0.38

For both approaches, the majority of the uncertainty is attributable to the uncertainty in the calibration parameters, and this is expected, since the amount of experimental data used to estimate the calibration parameters is relatively small. An interesting observation, though, is that for approach two, while it appears that the uncertainty in μ_k and $\mu_{\rho C}$ have a significantly larger effect than the uncertainty in δ, it is important to realize that the presence of the uncertain term δ in the calibration process results in additional uncertainty in the estimation of the calibration parameters themselves. Thus, the fact that the model bias is treated as an uncertain variable results in both a direct and an indirect increase in the resulting overall uncertainty.

The primary message of the probabilistic sensitivity analysis is that in order to reduce the uncertainty present in the estimation of p_f, more data are needed with which to estimate the calibration parameters. On the other hand, additional material characterization data with which

to refine the estimates of the material property variances would not significantly reduce the uncertainty in the failure probability estimation. It is also worth noting that one of the primary reasons that the calibration analysis leaves so much residual uncertainty is that the variation in the system response (as observed via simulation or the validation data) is very large relative to the number of repeated experiments at each configuration (which is at most four). This is the same reason that the validation analysis of Section 6.1.2 is not capable of drawing a strong conclusion about the validity of the given heat transfer model.

6.1.6 Conclusions

The complexity of the thermal challenge problem case study has afforded the opportunity to illustrate a variety of uncertainty quantification techniques. Statistical significance testing was first applied to develop quantitative statements about the agreement between the model predictions and observations. Subsequently, the hypothetical experimental observations were used to calibrate the given model using two different approaches. The calibrated models were then used along with comprehensive Bayesian uncertainty quantification techniques to address whether or not the probabilistic regulatory requirement condition would be met.

The results of the significance tests do not provide evidence to suggest that the given thermal model is inconsistent with the hypothetical experimental data. However, the power calculations suggest that the inability of these tests to provide significant evidence against the null hypothesis is due in large part to having a small number of repeated experimental observations. It turns out that the significance tests conducted at each of the ensemble configurations would have only about a 50% chance of rejecting a model whose predictions differed from the experiments by one standard deviation.

The ineffectiveness of the given experimental data in providing strong validation evidence

for the thermal model is also confirmed by two Bayesian calibration analyses. Using two calibration approaches, uncertainty distributions for the probability of failure in the application domain (defined in terms of Eq. (6.4)) are developed; these uncertainty distributions represent the amount of residual uncertainty, after calibration, associated with the thermal model's prediction of the probability of failure. The corresponding uncertainty distributions are given in Figures 6.6 and 6.7, and both indicate that there is a large amount of uncertainty in the calibrated predictions. This large uncertainty is consistent with the small power of the validation tests that was found in Section 6.1.2.

This case study affords an interesting opportunity to compare two Bayesian calibration approaches, specifically approaches that do and do not make use of the model inadequacy function proposed by Kennedy and O'Hagan (2001). It is interesting to note that the inclusion of the model inadequacy function (which itself is treated as an uncertain quantity) results in significantly more uncertainty in the estimation of p_f (compare Figures 6.6 and 6.7). This is easily explainable, because not only does the model inadequacy function itself contain uncertainty, but its inclusion results in additional uncertainty in the estimation of the calibration parameters, $\boldsymbol{\theta}$.

In fact, when the model inadequacy function is included, the total resulting uncertainty in the analysis will most likely depend strongly on the conditional variance of $\delta(\boldsymbol{x})$. This is somewhat worrisome, because this conditional variance is sensitive to the estimates of the corresponding Gaussian process parameters, and such estimates are not especially robust when there is a small amount of training data and the observations contain uncertainty. In particular, the process variance can be especially sensitive to the estimation technique when there is observation uncertainty, and this is unsettling because the process variance plays a large role in

determining the magnitude of the uncertainty in the calibrated predictions.

For these reasons, my recommendation is that the Gaussian process model inadequacy function be used with care. When using this function, one should make an effort to consider how robust the function is to the estimation of the governing Gaussian process parameters (Bayesian inference about the GP parameters might provide some insight in this regard). Even when the use of the model inadequacy function is appropriate in terms of having enough available data, I would only recommend its use when the magnitude of the model bias is significantly greater than zero, in terms of its uncertainty. For example, for the model bias function used in approach two and illustrated in Figure 6.5, the 80% uncertainty bounds for δ include zero at all of the configurations of interest. This suggests that its use may only add unnecessarily to the resulting prediction uncertainty, and as such I would recommend approach one over approach two for this case study, because approach one does not include the use of the model inadequacy function.

6.2 Model validation challenge problems: structural dynamics application

This section addresses the structural dynamics validation challenge problem (Red-Horse and Paez, 2008), which is another a hypothetical problem developed at Sandia National Laboratories to gain insight into the model validation process. The problem deals with the behavior of a simple three degree of freedom "subsystem" (Figure 6.8) and its response when attached to a simple beam to form the "system" configuration (Figure 6.9). The analyst is provided with several mathematical models to predict the response of various subsystem and system configurations, as well as hypothetical experimental data which can be compared against the model predictions for the purpose of validation.

The objectives of the problem are thus two-fold:

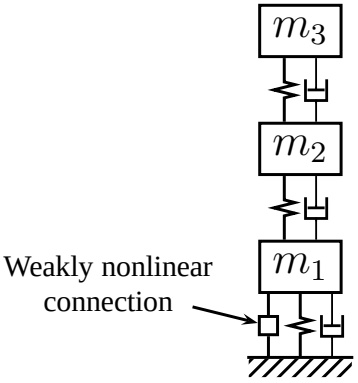


Figure 6.8: Schematic of the three degree-of-freedom subsystem for the structural dynamics challenge problem

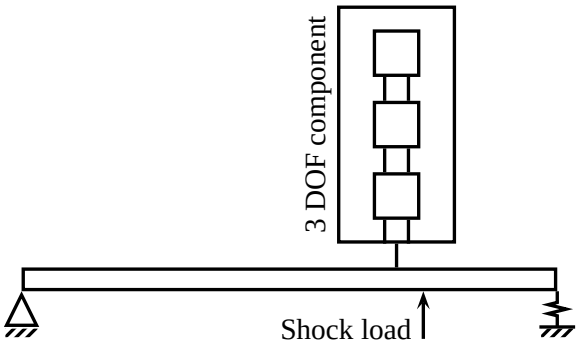


Figure 6.9: Schematic of the "accreditation" system configuration for the structural dynamics challenge problem

1. With respect to their intended use, assess the predictive capability of the given dynamics models based on available experimental data.

2. Apply the given system-level model for the target configuration, at which no experimental data are available, to predict whether or not a population of devices will meet the regulatory requirement.

6.2.1 Approach

The approach taken to address the above objectives consists of the following steps:

1. Assess the predictive capability of the subsystem model.

2. Characterize the joint distribution of the inputs based on available data.

3. Assess the predictive capability of the system model.

4. Predict the probability of failure (application configuration) using Monte Carlo simulation.

It will be made clear in Section 6.2.3 why the characterization of the input distribution is necessary for assessment of the system model, but not the subsystem model. Two different approaches to model assessment are discussed, whose applicability depends on whether the experiments are fully characterized. For the validation of the subsystem model, the corresponding experiments are fully characterized, whereas for the system model validation case, the modal parameters, which are model inputs, are unknowns for the experiments. For each case, the focus is on model assessment which is meaningful and understandable, and which takes account of the intended use of the model.

Although model assessment is an important aspect of the challenge problem, the ultimate objective is to make use of the given models for predicting the failure probability associated with a population of devices in the system configuration. The approach taken here is to specify a distribution for the random variables that characterize the device variability, and to use Monte Carlo simulation to estimate the failure probability condition. Unfortunately, though, the given mathematical model for the target application is not trivial to evaluate, so only a limited number of evaluations are available. Two possible approaches are thus:

1. Use an efficient sampling method such as LHS to estimate the failure probability condition .

2. Use the available evaluations of the true mathematical model to develop a fast response surface approximation. Then use the response surface approximation within a Monte Carlo simulation with a very large sample size.

These two approaches, among others, are discussed in Section 4.2. The study conducted by Giunta et al. (2006) suggests that the second approach (in particular the use of Gaussian process response surface approximations) has the potential to be significantly more accurate, particularly when dealing failure probability estimation. For this reason, the response surface approximation approach is taken here.

Section 6.2.2 proposes a novel approach for the characterization of and random sampling from the probability distribution for the random modal inputs. This step is necessary both for the validation assessment of the system model (second part of Section 6.2.3) and for the probability of failure prediction (Section 6.2.4). The proposed approach makes use of multivariate kernel density estimation, principal component analysis, and Markov Chain Monte

Carlo sampling to deal with the complexity of non-Gaussian density characterization for high-dimensional random vectors.

6.2.2 Input distribution characterization

Since the behavior of the subsystem is characterized using a linear model, the response is fully specified by the modal parameters: three natural frequencies, damping coefficients, and mode shapes, for a total of fifteen parameters. Thus, the natural way to characterize the variability associated with the subsystems is to treat the modal parameters as random variables. Once a joint probability distribution for the modal parameters has been specified, it is possible to propagate the randomness through the given models using Monte Carlo simulation (this also applies to the system models, which contain as a component the three-degree-of-freedom subsystem: see Figure 6.9).

The probability distribution for the modal parameters can be estimated based on available data. In this case, the statistical data for the modal parameters are available for both the random vibration (calibration) and shock (validation) force inputs. For each experiment, 20 nominally identical components were tested at three different input levels: low, medium, and high. Thus, there are 60 data points each for the random vibration and shock inputs, for a total of 120 data points. However, the distributions of the random vibration and shock data differ significantly (perhaps due to the non-linearity in the components). For this reason, and since the intended use of the model pertains to shock excitation, the input distributions will be characterized based only on the shock data.

One of the most widely used and straightforward methods for characterizing randomness is through the use of the normal distribution. Unfortunately, however, the assumption of normality is violated for several of the parameters. For example, based on the well-known Shapiro-

Wilk test for normality (Shapiro and Wilk, 1965) with $n = 60$ observations, univariate normality is rejected at the usual 0.05 significance level for the distributions of all three natural frequencies, along with the distributions of the first two damping coefficients.

In addition to the violations of normality, many of the modal parameters are highly correlated. For example, all three modal frequencies are nearly perfectly correlated with each other, and the frequencies show high negative correlations with the second and third elements of the first mode shape.

Thus, the distribution for the modal parameters is not only multivariate, but also highly correlated and non-normal, meaning that the use of the multivariate normal distribution is not appropriate. For the case of non-normal multivariate distributions, some specialized techniques are available to deal with fully-specified parametric joint distributions (see Section 4.2), but none are appropriate for the present case. This work makes use of the non-parametric density characterization approach known as kernel density estimation (KDE), which is described in Section 4.2.2.

Two challenges arise when attempting to use KDE. First, KDE does not tend to work well for high dimensions, because of the "curse of dimensionality'. That is, in high dimensions, large regions in the parameter space will have virtually no data in them. Second, there is no natural way of generating random samples from such a density representation, and random sampling will be needed in order to estimate the distribution of the system response.

The first challenge will be handled with the use of principal component analysis (PCA), which is outlined in Section 4.2.3. PCA is appropriate here because it allows a high-dimensional random vector to be re-expressed in terms of a lower-dimensional space. Finally, Markov Chain Monte Carlo sampling (Section 2.3) is adopted to overcome the second challenge, which

is to generate random samples from such an arbitrary distribution.

The first step is to apply PCA to find a reduced set of variables with which to represent the total variation, which involves an eigen-decomposition of the covariance matrix (the correlation matrix is used here because the variables are not measured in the same units). The analysis conducted here is based on the sixty realizations from the subsystem "validation" experiments (which correspond to the shock excitation). The corresponding eigenvalues are plotted in Figure 6.10.

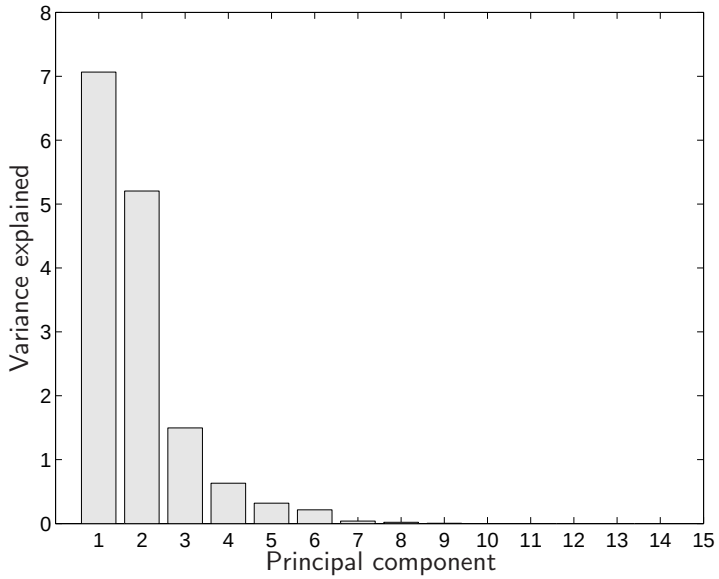


Figure 6.10: Eigenvalues corresponding to the correlation matrix of the modal parameters. Each eigenvalue represents the amount of variation explained by the corresponding principal component.

It is apparent from Fig. 6.10 that the first principal component explains approximately half of the total variation of all fifteen modal parameters (see Eq. (4.19)). By taking the first four principal components, approximately 96% of the total variation is explained. In order to balance efficiency and tractability against information loss, the first four principal components are retained to construct the probability density estimate.

The density estimate is now constructed using multivariate KDE based on the four (un-

correlated, but not necessarily independent) principal components of the modal parameters. Specifically, the density estimate given by Eq. (4.14) is employed, where the bandwidth is estimated by minimizing the cross-validation score function given by Eq. (4.17), as discussed in Section 4.2.2. Finally, MCMC sampling, which is discussed in Section 2.3, can be used to generate random samples from the resulting density estimate. The resulting samples are of course samples of the principal components, but the original variables can be obtained via the simple reverse transformation of Eq. (4.21). A simple verification of this density estimation and sampling procedure is discussed next.

Verification of density characterization and sampling

The performance of the above procedure for maintaining the distribution structure of the original variables is considered here. The idea is to compare a set of randomly generated modal parameters to that of the original data. The difficulty is that there is no straightforward way of making such a comparison, and it can become cumbersome because of the high dimensionality of the data.

To simplify the comparison process, two comparisons are presented here, in which 1,000 randomly generated realizations are compared against the original 60 data points:

1. **Correlations among the modal parameters**: A 15×15 sample correlation matrix can be computed for both the original and simulated data. To provide a compact representation of the comparison, the elements of the two matrices are plotted against each other. For perfect agreement, the points would fall on the line $y = x$.

2. **Marginal distributions**: The agreement of the marginal distributions can be assessed by comparing the empirical cumulative distribution functions (CDF's). Further, to avoid

redundant comparisons, the marginal distributions of the first four principal components are compared as opposed to those of the 15 original variables.

The agreement of the pairwise correlations is plotted in Fig. 6.11. Most of the points fall near the line $y = x$, indicating that there is very good agreement between the observed and simulated correlation coefficients.

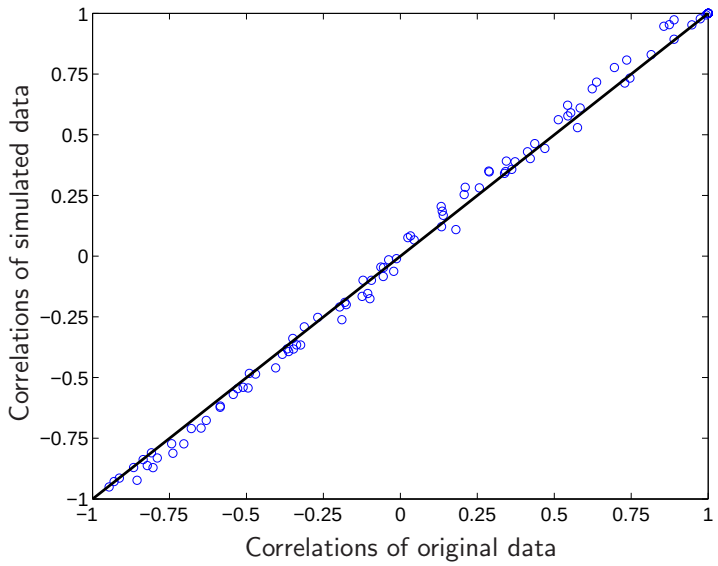
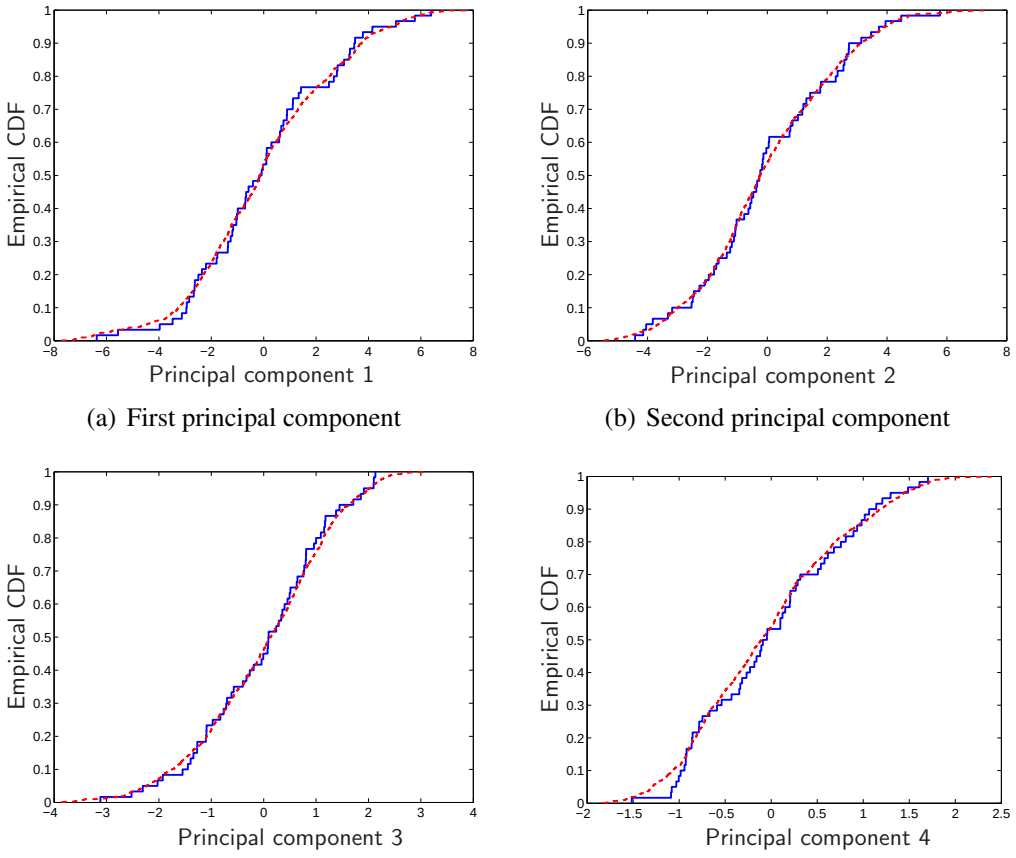


Figure 6.11: Sample correlation coefficients of original modal parameter data versus those of simulated data. The solid line represents perfect agreement.

The empirical CDF's for the first four principal components are compared in Figure 6.12, which indicates excellent agreement, particularly for the first three principal components. There is some discrepancy for the fourth component, but this can be expected, considering the erratic nature of the empirical CDF for the original data of this component.

Finally, the agreement between the marginal distributions and the correlations is a good check, but it does not necessarily indicate that the full joint distributions are matching. The difficulty is that with non-normal data, a joint distribution is not fully specified simply by the marginals and pairwise correlations. The above results are presented as a sanity check, but it is acknowledged that they are not a complete verification of the proposed sampling methodology.



(a) First principal component

(b) Second principal component

(c) Third principal component

(d) Fourth principal component

Figure 6.12: Comparison of observed empirical CDF's (solid lines) and empirical CDF's of simulated data (dashed lines).

6.2.3 Model assessment

This section addresses the validation assessment of the given structural models based on the corresponding experimental data. For the challenge problem, experimental data are available at both the subsystem and system levels, so that these models can be assessed individually. However, the method of model assessment must differ in each case because of the nature of the uncertainty associated with the predictions and observations. Thus, for this study, model assessment is thus classified into two categories:

1. **Fully characterized experiments**: for each experiment, all of the parameters corresponding to model inputs are known.

2. **Partially characterized experiments**: for each experiment, some of the parameters corresponding to model inputs are unknown.

In the following sections it is shown that the experiments at the subsystem level are fully characterized, while the experiments at the system level are only partially characterized (in reality, there will always be a certain amount of uncertainty associated with experimental conditions; nevertheless, it may still be useful in some situations to assume that this uncertainty is negligible). First, it is discussed how the quality of the subsystem model can be assessed by considering pairwise prediction/observation comparisons to derive inference about prediction error. Second, a different method is adopted, in which model assessment claims are not as strong. For this second case, each individual observation is compared to an entire model output distribution.

Subsystem model assessment

The confidence assessment of the subsystem model is an important step because all of the modeling error and uncertainty, even for the system configurations, can be attributed to the subsystem model (Red-Horse and Paez, 2008). In order to assess the subsystem model, the analyst is given both the linear model with which to make predictions, along with the results of various experiments.

Since the assessment will be based on the comparison of the dynamic time-history responses predicted by the linear model to those of the experimental results, the first step is to decide on some method for comparing the time-histories. Directly comparing two time-histories tends to be of little practical use, and the preferred method is to make the comparison based on one or more response features (Therrien, 1989; Jain and Zongker, 1997). Recall that the objective of model validation is to asses the quality of the model with regards to its intended use. Fortunately, for the challenge problem, the intended use of the model is clearly specified, as discussed in Section 6.2.4. Based on the model's intended use, the most natural comparison feature to work with is the maximum absolute acceleration of mass 3, which will be denoted here by $\tilde{a}$. This is also a very convenient feature to work with because it is a scalar quantity.

There are data available from a total of 120 tests on the subsystem with which to assess the quality of the given linear model. Each of these experiments corresponds to a different subsystem that is randomly selected from a population that contains variability. Sixty of these experiments subjected the subsystem to random vibration excitation (these are referred to as the "subsystem calibration" experiments by Red-Horse and Paez (2008)), and sixty of these experiments subjected the subsystem to shock excitations (referred to as the "subsystem validation" experiments). Again, in view of the intended use of the model, the validation assessment of

the subsystem model will be based only on those experiments which used shock excitations, because this is the excitation which corresponds to the target application.

These sixty experiments are further divided into three categories based on the nominal excitation level: low, medium, and high. As discussed below, each of these groups of experiments is treated as a separate "population," and the groups are compared separately with the predictions made by the linear model.

It is mentioned above that experimental tests on the subsystem can be classified as "fully characterized," in that any model input parameters which must be supplied in order to obtain corresponding predictions are known for each of the experiments. The inputs to the linear model consist of a) the excitation waveform and b) the modal parameters for the subsystem. The excitation waveform is known for each experiment. In addition, the modal parameters of each subsystem tested are also known because they can be back-calculated from the experimental data.

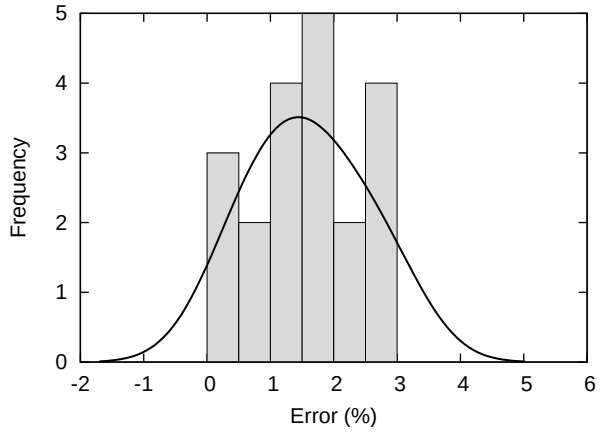
Since the experiments are fully characterized, there is one single model prediction corresponding to each. A direct comparison, based on the specified response feature, allows one to compute a scalar prediction error associated with each experiment. Let this error be defined as:

$$e = \tilde{a}_{\text{obs}} - \tilde{a}_{\text{pred}}, \tag{6.16}$$

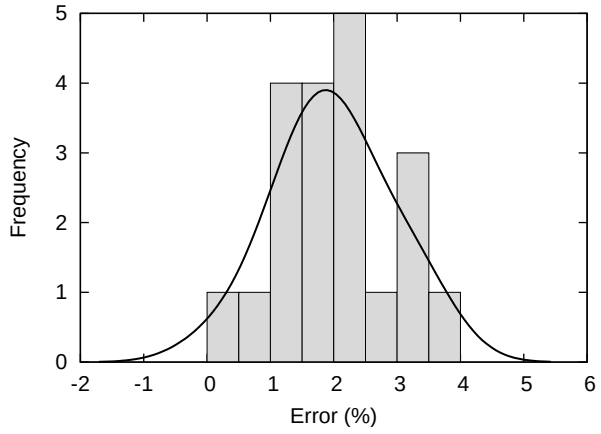
where $\tilde{a}_{\text{pred}}$ and $\tilde{a}_{\text{obs}}$ are the predicted and observed values of the response feature, respectively. In addition, for each of three nominal excitation levels, twenty randomly chosen specimens are tested. The proposed approach is to divide the data based on nominal excitation, and characterize the distribution of the error based on each.

The results are shown in Fig. 6.13, which shows histograms and non-parametric kernel

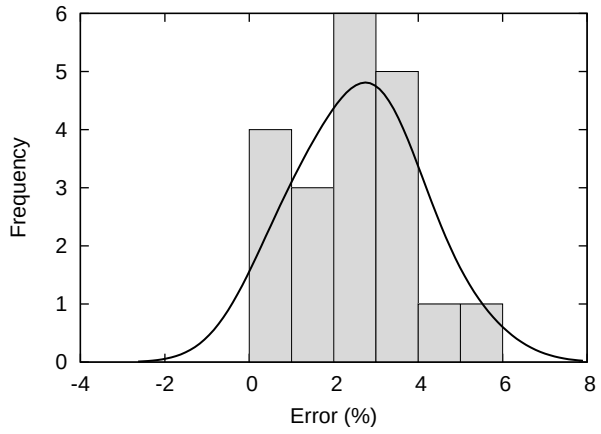
density estimates for the error, as a percentage of $\tilde{a}_{\text{obs}}$, for each of the three excitation levels.



(a) Low excitation level



(b) Medium excitation level



(c) High excitation level

Figure 6.13: Histogram and density estimate for prediction error at three excitation levels

The estimated distribution of the prediction error can now be used to make inferences about

the quality of the linear model. First, note that based on the available data, virtually all of the prediction errors are positive, indicating that the model has a strong tendency to under-predict the response, $\tilde{a}$. Further, these results also suggest that on a percentage basis, the distribution of the error does not appear to depend on the nominal excitation level. Finally, the magnitude of the error is generally observed in to be in the range of 0 to -4 percent.

System model assessment

In addition to experiments conducted on the subsystem, a small amount of experimental data is also available for the response corresponding to the "accreditation system" configuration. For this configuration, one test each has been done at low, medium, and high excitation levels.

The given model for predicting the behavior of this system takes as inputs an excitation waveform and a set of modal parameters describing the particular subsystem attached to the beam. With regards to validation, the fundamental difference from the case discussed above is that the modal parameters for the subsystem can not be derived from the response of the system. As a result, the modal parameters governing the subsystem, which are needed as inputs to the system model, are unknown. Thus, this validation analysis can be classified in the second category discussed in the beginning of Section 6.2.3: partially characterized experiments.

For the first case it was possible to make one-to-one comparisons between the predictions and observations because all of the model inputs corresponding to each experiment were known. However, for this case, the subsystem modal parameters associated with each experiment are unknown, but they are still needed as inputs to the model. Thus, for the purpose of model assessment, the following approach is adopted to deal with the case of partially characterized experiments:

1. Characterize the variability associated with the subsystem modal parameters.

2. Corresponding to the excitation of each experiment, propagate the subsystem variability through the system model using Monte Carlo simulation to obtain the predicted distribution of the response.

3. Compare the observed response with the predicted distribution obtained from the model.

Thus, for each of the three experiments, one observed value of the response is compared with an entire probability distribution associated with the model predictions. Clearly, this type of comparison makes for a much weaker assessment of the model's predictive capability than that of the first case. This analysis will not provide sufficient information to characterize the magnitude of the modeling error. In fact, the only conclusion that can be drawn is whether or not the experimental results are strongly inconsistent with the model predictions. Although the resulting inference about model quality is not as strong, it is the best that can be done given that the experiments are not fully characterized.

The first step, characterizing the probability distribution for the subsystem modal parameters, is discussed in Section 6.2.2. The second step is to use Monte Carlo simulation to propagate this variability through the given system models. As discussed in Section 6.2.1, the use of the system models directly inside a Monte Carlo simulation is computationally prohibitive. The approach taken here is to use the results from a reasonable number of runs of the given models to develop Gaussian process response surface approximations. The Gaussian process model is a powerful and flexible tool that has the ability to model a wide variety of functional forms, and its use is discussed in detail in Chapter III.

As with the first case, the validation comparisons are based on the scalar response feature $\tilde{a}$ only, so the response surface approximations are likewise constructed based on this feature only. Further, a separate Gaussian process model is constructed for predicting the response cor-

responding to each of the three excitations used in the accreditation experiments. A quadratic mean (a.k.a trend) function is also used for each response surface approximation.

For this work, the response approximations were found to give excellent fits using 150 training points. To assess the quality of the fits, the response approximations are used to predict a set of 50 held back data points. The mean absolute values of the errors were 265, 426, and 249 for the models corresponding to the first, second, and third accreditation experiments, respectively. In relationship to the magnitude of the response, these approximation errors are acceptably small (they correspond to 0.6%, 1.6%, and 0.6% of the experimentally observed response values, respectively).

The results of the Monte Carlo simulations for each of the three accreditation force levels are given in Figs. 6.14, 6.15, and 6.16. To assist with the visualization, the 95% highest density region (HDR; Lee, 2004) is shaded for each output distribution. The HDR indicates the most likely region for 95% of the responses, based on the model predictions. Similarly, the experimentally observed response is also plotted in each figure as a vertical line, to show where it lies in relation to the predicted output distribution.

Based on the results of Figs. 6.14, 6.15, and 6.16, the evidence does not suggest that the system-level model predictions are overly inconsistent with the experimental data. In all three cases, the experimental results lie within the 95% HDR's corresponding to the predicted response (although for the third case, the experimental response lies just at the upper bound of the predicted HDR). Further, for excitations 1 and 2, the experimentally observed response is near the mode (or most likely) value of the predicted response distribution. The gaps in the HDR's of Figs. 6.14 and 6.16 are the result of multimodality in the probability densities, in these cases causing a small separation between two regions of high probability. Finally, note

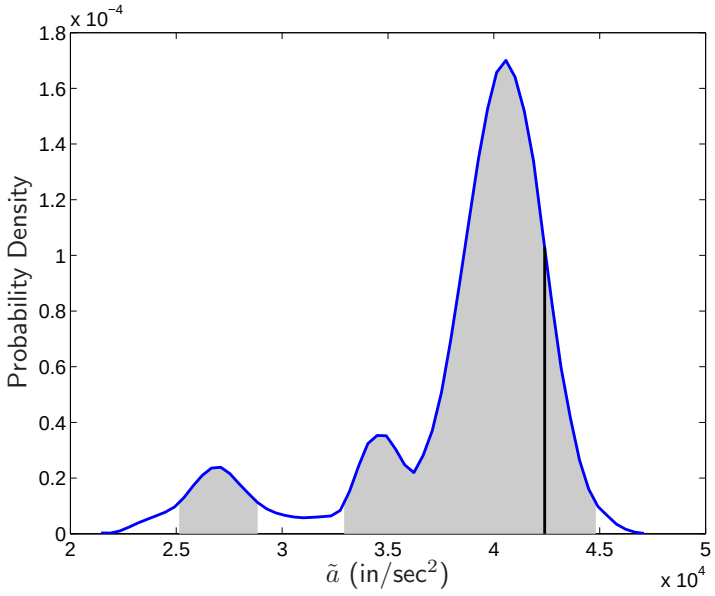


Figure 6.14: Predicted distribution for $\tilde{a}$ corresponding to excitation 1 for the accreditation configuration. The 95% highest density region is shaded, and the experimentally observed response is plotted as a vertical line.

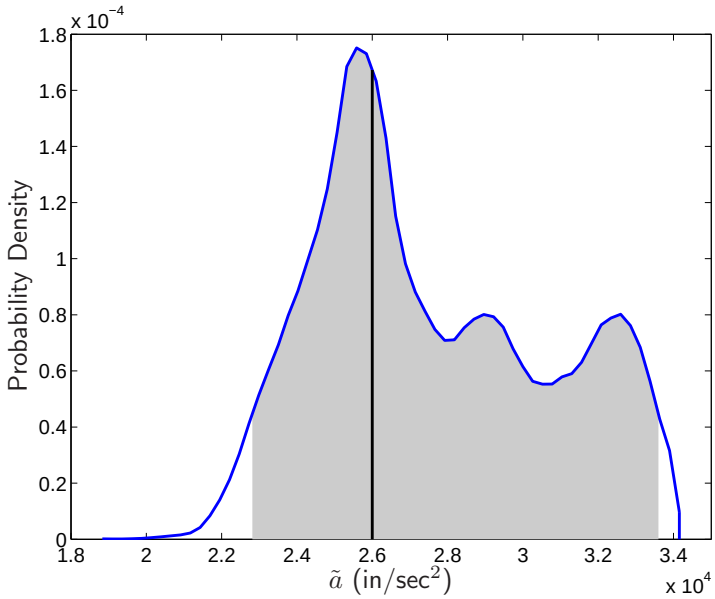


Figure 6.15: Predicted distribution for $\tilde{a}$ corresponding to excitation 2 for the accreditation configuration. The 95% HDR is shaded, and the experimentally observed response is plotted as a vertical line.

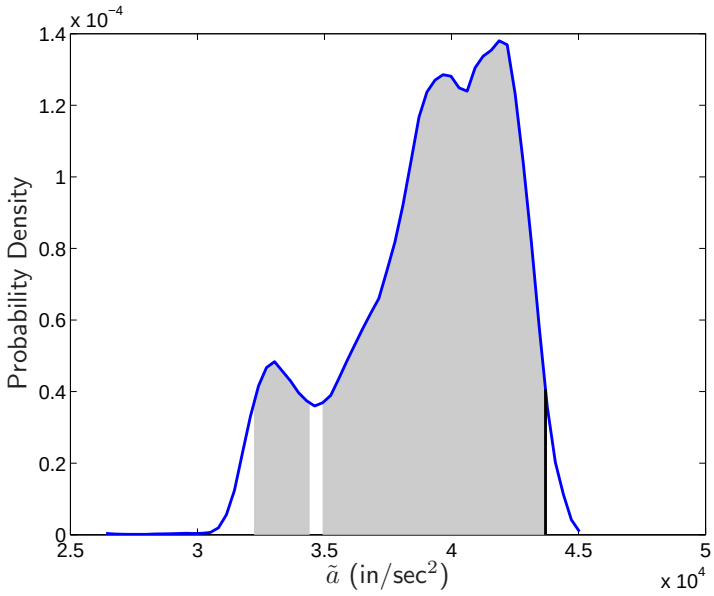


Figure 6.16: Predicted distribution for $\tilde{a}$ corresponding to excitation 3 for the accreditation configuration. The 95% HDR is shaded, and the experimentally observed response is plotted as a vertical line.

that as with the subsystem model, the system-level model tends to under-predict the magnitude of the response, $\tilde{a}$; this is particularly evident for excitations 1 and 3.

6.2.4 Prediction of system failure probability

After considering the predictive capability of the given models, the ultimate objective of the challenge problem is to predict whether or not a population of system devices will meet a specified probabilistic regulatory condition. The regulatory condition is defined in terms of a failure probability, which is specified for the "target configuration" of the system. The target system configuration differs slightly from the "accreditation configuration" (used for validation assessment discussed in the second case above) in terms of the support and loading conditions.

For the target configuration, the failure probability condition is defined in terms of the maximum absolute acceleration of mass 3 on the subsystem, $\tilde{a}$, as follows:

$$p_f = \text{Prob}\left(\tilde{a} > 1.8 \times 10^4 \text{ in/sec}^2\right). \tag{6.17}$$

The regulatory condition states that the failure probability shall not exceed 10^{-2}, and the analyst's objective is to use the given system model to predict whether or not this regulatory condition will be met.

Recall that the variability in the system response is due to specimen-to-specimen variability associated with the subsystem, which is treated here by characterizing a joint probability distribution for the subsystem modal parameters (the excitation and all other modeling conditions are known). The details associated with the characterization of and random sampling from this probability distribution are given in Section 6.2.2.

As was done in the second part of Section 6.2.3, the response $\tilde{a}$ predicted by the given system model is approximated with a Gaussian process surrogate to enable the large number of model evaluations necessary for Monte Carlo simulation. The surrogate model is constructed using 200 training points based on evaluations of the given dynamics model. The quality of the response approximation is evaluated by predicting the response at 50 new points, and the mean absolute prediction error is found to be 173, which is on the order of 1% of the response magnitude, and is deemed acceptable for the intended use of the model.

The predicted distribution of the response based on 25,000 random samples is shown in Fig. 6.17, and the corresponding Monte Carlo estimate of the failure probability is 0.14. It is apparent that the bulk of the predicted response values lie just inside the safe region, but there are still a significant number of cases which lie well inside the failure region.

Recall from Section 6.2.3 that both the subsystem model and the system model tended to under-predict the response measure $\tilde{a}$, which results suggest that, even for the application model, the true values are greater than or equal to the predicted values. Further, it is evident from Fig. 6.17 that only a slight shift to the right of the response distribution is needed to sig-

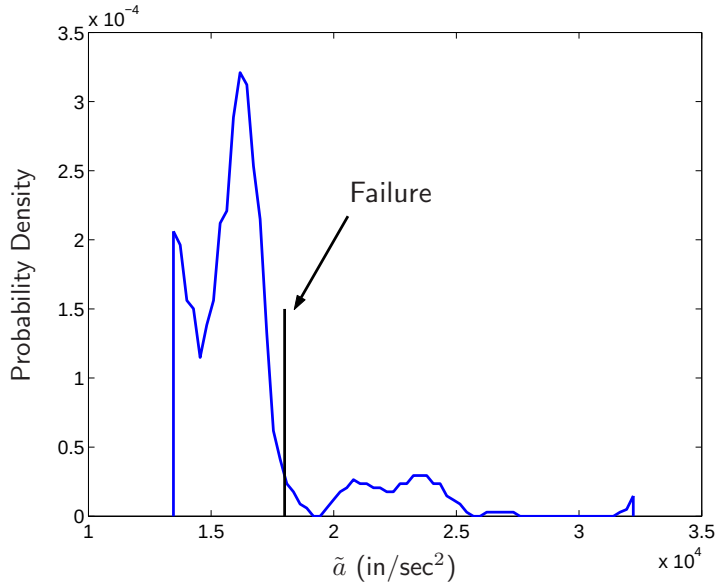


Figure 6.17: Predicted probability distribution of $\tilde{a}$ for the target system configuration.

nificantly increase the failure probability. Thus, the evidence suggests that there is substantial support for the conclusion that the regulatory condition $p_f < 10^{-2}$ will not be met.

6.2.5 Conclusions

Like the thermal challenge problem case study of the previous section, the structural dynamics challenge problem provides an excellent opportunity to explore a variety of uncertainty quantification techniques. One emphasis of the analysis discussed above is the development of model validation results that are informed strongly by the model's intended use. In addition, the uncertainty propagation activities undertaken to support both validation assessment and reliability prediction illustrate the use of Gaussian process surrogate models as inexpensive approximations to a complicated functional relationship between model inputs and outputs.

Model assessment is simplified by considering the relevant response feature only, which is a scalar quantity representing a maximum acceleration response. The task of model validation is broken into two cases. For the subsystem data, the experiments are fully characterized,

and one-to-one comparisons between observations and predictions are used to characterize the prediction error. For the system-level data, the subsystem modal parameters are not known, and a different approach is taken for model assessment. For this case, Monte Carlo simulation is employed to propagate the input uncertainty through the model, in order to compare the predicted output distribution against single experimental realizations. Admittedly, the strength of the conclusions that can be made from this form of model assessment are not as good as those of the case when the experiments are fully characterized.

Finally, one of the highlights of this particular case study is the development, illustration, and verification of a novel approach to characterizing and sampling from a complicated, high-dimensional probability distribution. The approach, described in Section 6.2.2, is especially powerful when a joint probability distribution function for a relatively large number of random model inputs is to be characterized using observed samples. Traditional density estimation techniques are often not appropriate for characterizing a non-normal joint density function with large dependencies among the variables (as is the case with the present case study). Principal component analysis and kernel density estimation are employed to surmount these issues, and Markov Chain Monte Carlo sampling may be used to sample from the resulting density for the purpose of uncertainty propagation.

6.3 Bayesian model calibration: QASPR simulation

This section illustrates the Bayesian model calibration approach, with the use of Gaussian process surrogates. The methodology is applied here to data from a modeling and simulation project at Sandia National Laboratories. Some of the noteworthy features of this particular case study are that there are a large number of calibration parameters (12 are considered here), and that the response quantity of interest was measured at multiple time instants during the

experiments.

6.3.1 Introduction

Qualification Alternatives to the Sandia Pulsed Reactor (QASPR) is a modeling a simulation project at Sandia National Laboratories. The purpose of the QASPR project is to develop validated computational simulations that model the interactions between nuclear radiation and electronic components. While the project itself consists of several levels of code ranging from the atomic scale to the electronic circuit scale, this study deals only with the calibration of one particular piece of the code, which is concerned with effects at the device level.

The objective of this study is to use available experimental observations of a response quantity of interest (which is a current ratio, or gain) to infer values and corresponding uncertainties for 12 calibration parameters governing the corresponding simulation model. The experiments were conducted at three different system configurations, which will be referred to as "Q1, Q2, and Q3"; these configurations represent different bias voltages that are applied to the transistor at the time of the radiation pulse. Additionally, the response quantity of interest is recorded at four time instants for each experiment.

One particular challenge with this study is that the calibration analysis must be performed without conducting any new simulator runs. Corresponding to each of the three system configurations of interest, the results of 300 runs of the simulator are available, and the relationship between the simulator inputs and outputs is known only in so much as it can be inferred from the previously observed simulator runs. As mentioned above, this calibration analysis is interested in making inference about 12 calibration parameters. The prior information about these parameters consists only of the reasonable bounds for each, as exhibited via the design of the original computer experiments.

An additional point of interest is that the same design of computer experiments was not used for the 300 simulator runs corresponding to each of the three system configurations (which was beyond this author's control). That is, different realizations of the calibration parameters were used in the simulator runs for each of configurations Q1, Q2, and Q3. The implications of this for the response surface approximations and the cross-validation assessments are discussed in Section 6.3.4.

Several different calibration analyses are considered, based on various combinations of the data, but the underlying methodology will be the same. In each case, a Gaussian process surrogate model is used in place of the simulator, and the Bayesian calibration approach presented in Section 5.3 is implemented. For each analysis, the prior distribution for $\boldsymbol{\theta}$ is independently uniform over the bounds that were used for the original computer experiments. This a logical choice for a vague prior distribution, and at the same time the use of bounds prevents the posterior from extending into regions in which the surrogate models can not be expected to be valid.

The error model associated with Eq. (5.11) is treated here as

$$\varepsilon_i \sim N(0, \sigma_i^2), \tag{6.18}$$

where the σ_i^2 are treated as known constants because there is insufficient experimental data to allow a meaningful treatment of the σ_i^2 as unknowns. In each case, σ_i is set equal to ten percent of the corresponding experimentally observed response value. This choice is admittedly highly arbitrary, and it is used here only to illustrate the calibration analysis. As such, the results must be interpreted accordingly, in that the resulting uncertainty magnitudes are valid only in so much as the specifications of σ_i^2 are valid. Nevertheless, effective parameter inference can

still be undertaken, and in cases in which more information were available (for example, error ranges or repeatability quantification associated with the experimental results), the values of σ_i^2 could be informed accordingly.

In the analysis that follows, the system response at 0.1, 1, 10, and 80 milliseconds (the only four time instants for which model predictions or experimental observations are available) will be referred to as "response 1", "response 2", "response 3", and "response 4", respectively. Accordingly, the 12 calibration parameters will be referred to as "input 1", ..., "input 12".

For each analysis, the associated Gaussian process surrogate models use a constant trend function, and the governing parameters are estimated using maximum likelihood, as discussed in Section 3.3. In addition, the posterior distribution for each analysis is constructed using the component-wise version of the Metropolis sampling algorithm (discussed in Section 2.3).

The calibration analysis will be broken down into a "nominal" analysis (Section 6.3.2), which considers only one data point, and two extensions to consider additional data, discussed in Section 6.3.3. Some additional analysis of the results is presented in Section 6.3.4, including cross-validation.

6.3.2 Calibration analysis: nominal case

The "nominal case" that is discussed here is defined as the calibration analysis that considers only the data for the Q1 scenario and response 1, such that $n = 1$. The first step is to use the 300 observed simulator runs to construct a Gaussian process approximation that relates the calibration parameters, $\boldsymbol{\theta}$, to response 1. The resulting maximum likelihood estimates of the normalized correlation lengths indicate that the simulation is probably most sensitive to inputs 6, 11, 2, and 8 (in that order).

To illustrate the form of the response, several input/output plots based on the Gaussian

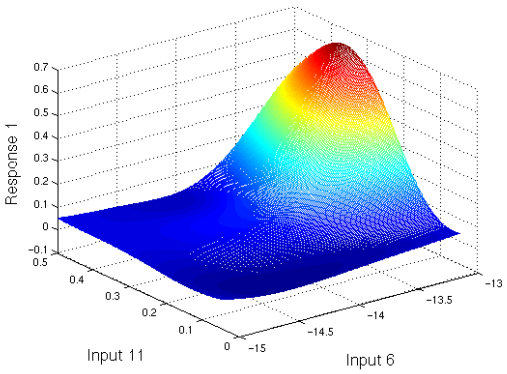
process model are constructed. These plots display response 1 versus inputs 6 and 11 (the 2 most important inputs) and response 1 versus inputs 2 and 8 (the next 2 most important inputs). In each case, the values of the 9 other inputs are held constant.[2] Figure 6.18 plots response 1 as a function of inputs 6 and 11 as both a mesh plot and a contour plot. The contour plot helps to illustrate the particular region of these inputs that matches well with the experimental observation (which observation is 0.41 for this case). A mesh/contour plot of σ_{GP} is also given to illustrate how the response surface approximation uncertainty varies in this domain. The corresponding 3 plots are also given for inputs 2 and 8 in Figure 6.19.

After specifying the GP response surface model, 25,000 MCMC samples are used to construct the posterior distribution of the calibration inputs. The marginal posterior distributions for the two most important inputs (6 and 11) are shown in Figures 6.20 and 6.21.[3] The marginal posteriors for inputs 6 and 11 both suggest that the upper bounds for these variables should possibly be increased in the future. The remaining ten inputs show less deviation from their marginal prior distributions, and five of the inputs show almost no change from their priors (which means that marginally, all values within the respective ranges are equally effective at yielding a response consistent with the observation).

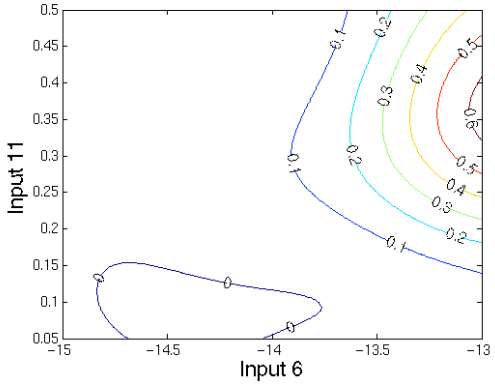
Two of the largest correlations between the updated inputs are between inputs 2 and 6, and between 5 and 11. Multivariate kernel density estimation (Section 4.2.2) is used to display contour plots of the two bivariate marginal densities in Figures 6.22 and 6.23. Using Spearman's ρ, a non-parametric correlation measure, the correlations are -0.26 and -0.29, which seem fairly mild. However, as more variables are considered together, the correlation struc-

[2]The particular values of the non-varying inputs are chosen based on the results of the calibration analysis itself. The values used are the estimated joint mode of the inputs, based on the observation of response 1. This means that the response plots will be relevant to the posterior distributions of the inputs.

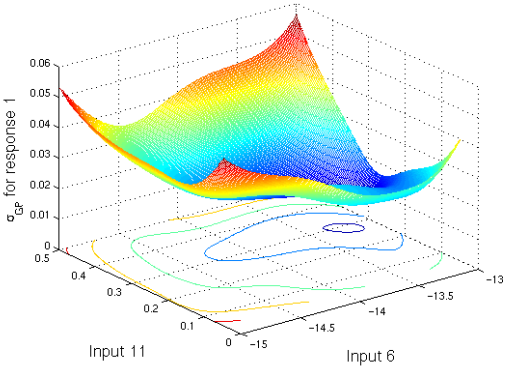
[3]To display the marginal posteriors, a beta distribution is fit to the posterior MCMC samples. The beta distribution is suitable because the variables have an upper and lower bound, and because the distributions are unimodal.



(a) Response based on Gaussian process model (μ_{GP})

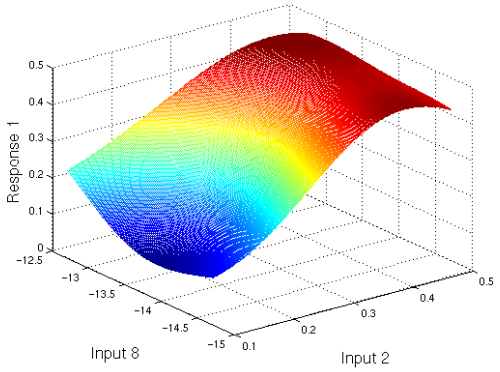


(b) Contour plot of response based on Gaussian process model (μ_{GP})

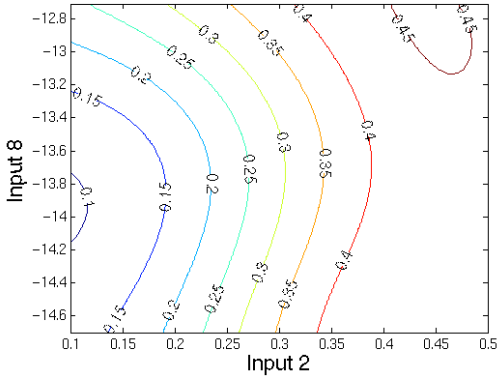


(c) Uncertainty (σ_{GP}) associated with Gaussian process model

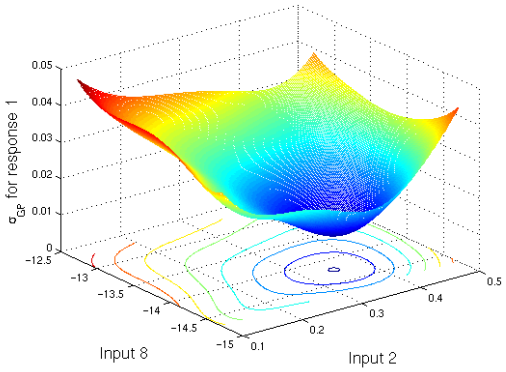
Figure 6.18: Gaussian process approximation to response 1 based on inputs 6 and 11



(a) Response based on Gaussian process model (μ_{GP})



(b) Contour plot of response based on Gaussian process model (μ_{GP})



(c) Uncertainty (σ_{GP}) associated with Gaussian process model

Figure 6.19: Gaussian process approximation to response 1 based on inputs 2 and 8

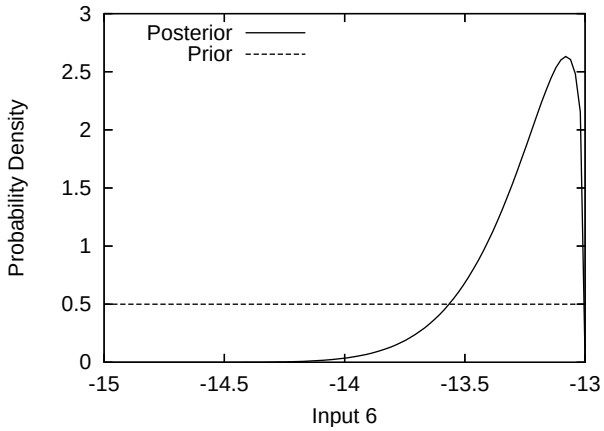


Figure 6.20: Marginal prior and posterior distributions for input 6 based on nominal calibration analysis

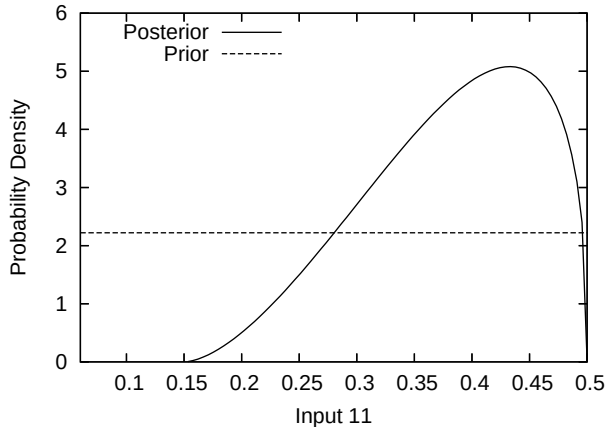


Figure 6.21: Marginal prior and posterior distributions for input 11 based on nominal calibration analysis

ture can only become more complicated, further reinforcing the fact that it is dangerous to assume the updated probability distributions to be independent of each other (as discussed in Section 6.3.4).

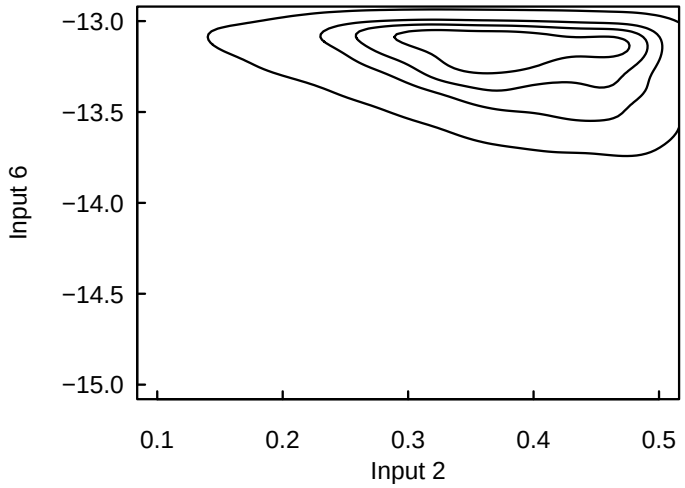


Figure 6.22: Estimated joint density of inputs 2 and 6

As a check on the calibration analysis, the posterior predictive distribution is compared to the experimental observation. The posterior predictive distribution is simply obtained by propagating the posterior distribution of $\boldsymbol{\theta}$ through the Gaussian process approximation to the simulator (recall that access to the simulator itself is not available). This comparison is illustrated below in Figure 6.24. The experimental observation is represented via a normal

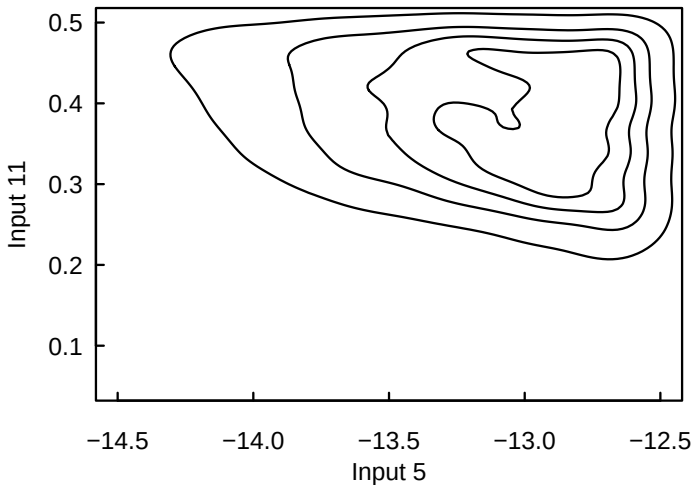


Figure 6.23: Estimated joint density of inputs 5 and 11

distribution with variance σ^2.[4] In addition, Figure 6.24 also shows the empirical probability density of the response value corresponding to the original 300 model runs. This distribution is included as a point of reference, and can be used to gauge the improvement associated with the calibrated parameter estimates. It is clear that the calibration analysis has resulted in model predictions (albeit, predictions based on the response surface approximation model) that agree well with the observation, particularly in comparison to the original simulator runs.

Since an independently uniform prior has been used for $\boldsymbol{\theta}$, it is expected that posterior predictive distribution will be proportional to the experimental uncertainty distribution (assuming the calibration analysis is successful in matching the predictions with the observation). This would be the case, except for the fact that the uncertainty in the response surface approximation is included in the Bayesian updating (as in Eq. (5.14)). The additional uncertainty added by the response surface approximation causes the variance of the posterior to be greater than the variance/uncertainty of the experiments. In addition, the posterior is "pulled" very slightly away from the datum towards the area where there is less response surface approximation un-

[4]Recall from Section 5.3 that σ^2 can represent both error/uncertainty in the experimental observations and error/uncertainty in the model output. To simplify the visualization here, σ^2 is considered to be uncertainty associated with the experiment.

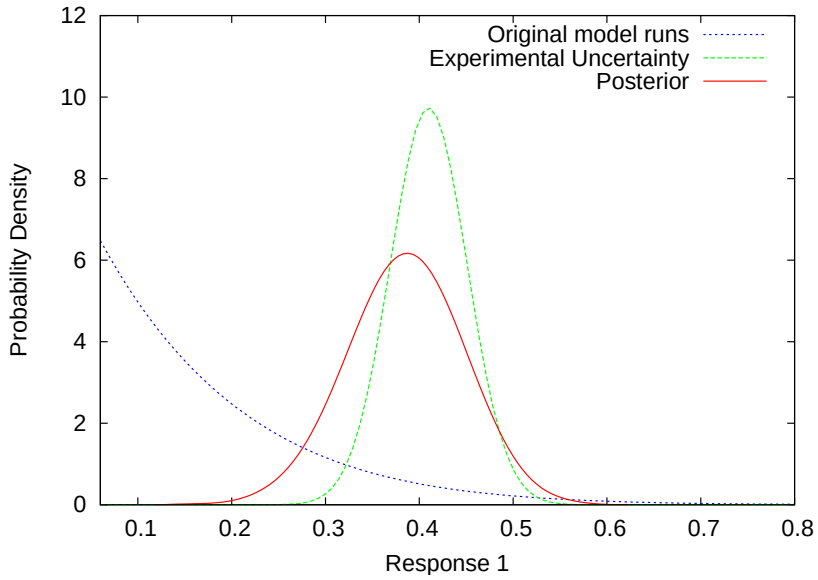


Figure 6.24: Distribution of original model runs, experimental uncertainty, and posterior predictive distribution for the nominal calibration analysis

certainty. Since most of the original model runs correspond to values of the response that are less than the observation, the posterior shifts slightly away from the datum in this direction.

6.3.3 Calibration based on multiple observations

This section presents two extensions to the nominal analysis, both of which are based on the inclusion of additional experimental observations. The first extension discusses a calibration analysis that considers all four response measurements for Q1. For this case, the responses are not independent, and a novel approach based on Principal Component Analysis is proposed. The second extension is to consider calibration analyses that account for data from multiple scenarios.

Multiple time responses

The preceding nominal analyses is based on one response value only, whereas both the simulation model and the experimental observations consist of measurements of the response at

4 distinct time instants. Ideally, one would like the calibration analysis to take account of all observed response quantities. This section will discuss a method for updating the inputs based on all four response measures (the data for which are shown in Figure 6.25).

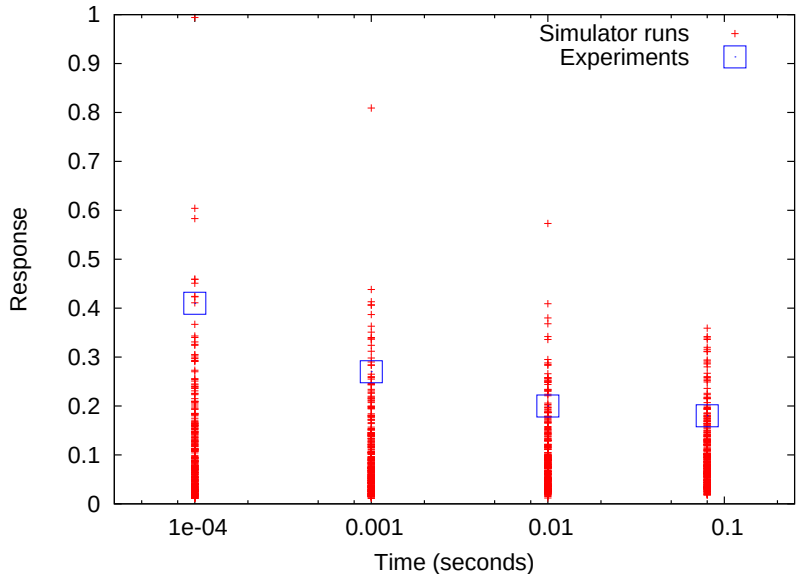


Figure 6.25: Predicted and measured response for Q1 configuration, as a function of time

For a calibration analysis based on time-history output that makes use of Gaussian process surrogates, two possible approaches would be:

1. The surrogate model captures time as an input.

2. A separate, independent, surrogate model is constructed to represent the response for each time instant of interest.

The first approach can be computationally cumbersome, although the use of the point selection algorithm proposed in Section 3.4 is shown to be both efficient and effective via the application discussed in Section 6.4. The second approach is more straightforward and more appropriate when the number of time instants is small. For example, this calibration analysis could be conducted with only four independent surrogates. The approach taken here will be based

on multiple surrogate models, but a method based on principal component analysis (PCA) is presented that is applicable even when the time response is highly multivariate.

For this analysis, the usual assumption of independence for the ε_i is not made, because dependencies among response 1, 2, 3, and 4 are envisioned, and random noise associated with the experimental measurements is not envisioned to be a dominant factor (which itself might justify the independence assumption).

Denote the joint distribution for $\boldsymbol{\varepsilon}$ as $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. The first step is to characterize the covariance structure for $\boldsymbol{\varepsilon}$, which is one of the primary difficulties with discarding the usual assumption of independence. Since there is only a small amount of experimental data available for this analysis (and no repeated experiments), the approach will be to estimate the correlation structure using the observed simulator runs.

To do so, the covariance matrix for $\boldsymbol{\varepsilon}$ is constructed as

$$\boldsymbol{\Sigma} = \rho_{i,j}\sigma^{(i)}\sigma^{(j)} \tag{6.19}$$

where $\sigma^{(i)}$ is the assumed standard deviation for response i, and $\rho_{i,j}$ is the sample correlation coefficient computed based on the 300 original simulator runs. The reason for not estimating $\boldsymbol{\Sigma}$ exclusively from the simulator data is that the distributions (in this case, uniform) used to generate realizations of the calibration inputs do not have any tangible meaning in terms of actual variability.

A principal component analysis (Section 4.2.3) based on the correlation matrix is now applied to arrive at a more compact representation of the simulator output. The sample correlation matrix of the 4 time responses based on the original simulator runs is

$$\boldsymbol{R} = \begin{bmatrix} 1.00 & 0.97 & 0.86 & 0.56 \\ 0.97 & 1.00 & 0.95 & 0.69 \\ 0.86 & 0.95 & 1.00 & 0.86 \\ 0.56 & 0.69 & 0.86 & 1.00 \end{bmatrix}. \tag{6.20}$$

Using PCA, the transformation is given by the eigenvectors, $\boldsymbol{A}$; the corresponding eigenvalues, $\boldsymbol{\lambda}$, of $\boldsymbol{R}$ represent the amount of variance explained by each principal component:

$$\begin{aligned} \boldsymbol{A} &= \begin{bmatrix} 0.49 & 0.54 & -0.55 & 0.40 \\ 0.53 & 0.29 & 0.18 & -0.78 \\ 0.53 & -0.13 & 0.69 & 0.47 \\ 0.45 & -0.78 & -0.43 & -0.09 \end{bmatrix} \\ \boldsymbol{\lambda} &= \begin{bmatrix} 3.458 & 0.504 & 0.0336 & 0.0042 \end{bmatrix} \end{aligned} \tag{6.21}$$

First, the eigenvalues indicate that by using the first two principal components only, 99.1% of the variance of the original variables can be explained. This allows the number of variables to be reduced from 4 to 2. Also, the columns of $\boldsymbol{A}$ represent the transformations corresponding to each component. It is apparent that the first component is effectively an average of the 4 original variables. This is typical when the variables are highly correlated. The second component is made up mostly of the 1st and 4th original variables, which makes sense because they each contain slightly different information.

Thus, using only the first two principal components, the transformation matrix is given by

$$\boldsymbol{A}_{(2)} = \begin{bmatrix} 0.49 & 0.54 \\ 0.53 & 0.29 \\ 0.53 & -0.13 \\ 0.45 & -0.78 \end{bmatrix}. \tag{6.22}$$

Let $\boldsymbol{y}$ denote the response vector and $\boldsymbol{z}$ denote the principal components. The variables are first centered based on the experimentally observed response values, so that the PCA transformation

is given by

$$\boldsymbol{z} = \boldsymbol{A}_{(2)}^T \boldsymbol{y}' \tag{6.23}$$

and

$$\boldsymbol{y}' = \boldsymbol{A}_{(2)} \boldsymbol{z}, \tag{6.24}$$

where $\boldsymbol{y}' = \boldsymbol{D}_s^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{y}}_{obs})$, ($\tilde{\boldsymbol{y}}_{obs}$ is used here to represent the fixed realization of the experiment, *not* a random variable), and $\boldsymbol{D}_s$ is the diagonal matrix containing the assumed standard deviations of the experimental measurements.

Since $\boldsymbol{z}$ is a linear transformation of $\boldsymbol{y}$, it is straightforward to show that based on Eq. (5.11) and the multivariate normal error model, the sampling distribution for $\boldsymbol{z}$ (not yet accounting for surrogate uncertainty) is

$$\boldsymbol{z} \sim N_2 \left(\boldsymbol{A}_{(2)}^T \boldsymbol{D}_s^{-1} \left(\boldsymbol{G}(\boldsymbol{\theta}) - \tilde{\boldsymbol{y}}_{obs} \right), \ \boldsymbol{A}_{(2)}^T \boldsymbol{R} \boldsymbol{A}_{(2)} \right). \tag{6.25}$$

Since the response is now represented by a two-dimensional quantity, only two response surface approximation models are now needed. One surrogate captures the relationship between $\boldsymbol{\theta}$ and z_1, and another captures the relationship between $\boldsymbol{\theta}$ and z_2. The likelihood function, including the GP surrogate variance, can be expressed as

$$L(\boldsymbol{\theta}) \propto |\boldsymbol{\Sigma}_2|^{-1/2} \exp \left[-\frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu} \right], \tag{6.26}$$

where

$$\boldsymbol{\mu} = \boldsymbol{A}_{(2)}^T \boldsymbol{D}_s^{-1} \left(\boldsymbol{G}(\boldsymbol{\theta}) - \tilde{\boldsymbol{y}}_{obs} \right), \tag{6.27}$$

and

$$\boldsymbol{\Sigma}_2 = \boldsymbol{A}_{(2)}^T \boldsymbol{R} \boldsymbol{A}_{(2)} + \boldsymbol{\Sigma}_{GP}. \tag{6.28}$$

Note that $\boldsymbol{\Sigma}_{GP}$ is diagonal (because the two surrogate models are independent of each other); as a result, $\boldsymbol{\Sigma}_2$ is also diagonal, which allows the likelihood function to be computed more efficiently.

The results of this calibration analysis indicate that compared to the calibration results based only on 1 response, a much more precise combination of inputs is required to get good agreement for all 4 response values simultaneously, and this is expected. Unlike the nominal case, almost all of the inputs now have refined posterior distributions, as opposed to having support along their entire bounds.

Figure 6.26 illustrates the agreement between the model predictions and the experimental observations, after updating the input distributions. The posterior predictive distributions of three of the response values are plotted (response 3 is omitted for clarity), along with the experimental uncertainty, as before. The fact that the posterior predictive distributions agree well with all of the response features suggests a successful calibration. Recall that MCMC simulation was conducted on the transformed variables (principal components). Thus, to plot the posterior distributions of the original variables, the reverse transformation given by Eq. (6.24) is employed.

Data from multiple scenarios

In this section, the calibration analysis attempts to account for data from multiple scenarios. For simplicity, though, only response 1 is considered. Since the experiments corresponding to each configuration are independent, it is now safe to assume independence for the error terms,

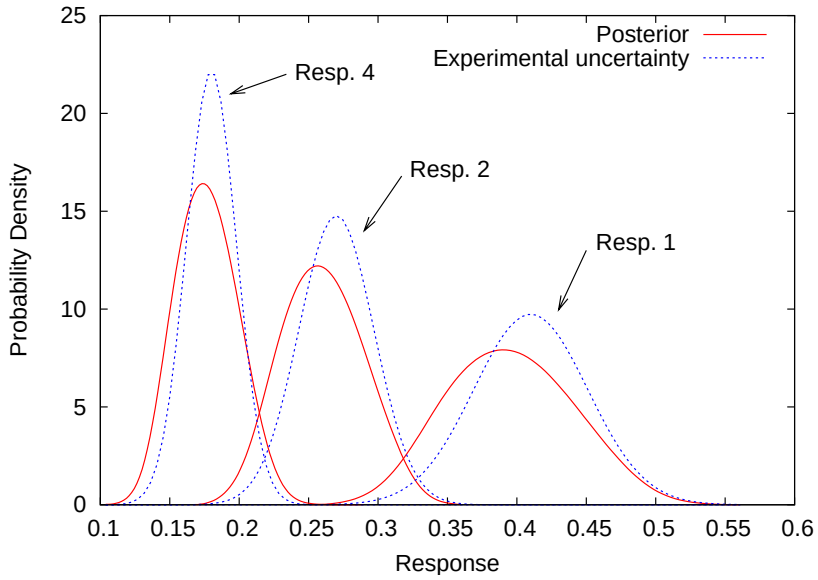


Figure 6.26: Posterior predictive distributions for responses 1, 2, and 4 (response 3 omitted for clarity) resulting from the calibration based on 2 principal components of all 4 response measures.

which simplifies the analysis.

The corresponding data for the three scenarios (response 1 only) are shown in Figure 6.27. It is readily apparent that on average, the simulator is under-predicting for Q1, shows little bias for Q2, and is over-predicting for Q3. This is a preliminary indication that there may be modeling bias that can not be captured via the calibration inputs alone. In particular, it may not be possible to simultaneously calibrate to both the Q1 and Q3 data.

The first analysis is to calibrate based on the data from both the Q1 and Q2 scenarios ($n = 2$). As mentioned above, the error terms are now independent. A separate Gaussian process surrogate model is created to capture the simulator response for each of the two configurations. The results of this analysis are illustrated in Figure 6.28, which compares the posterior predictive distributions for both scenarios to the corresponding experimental observations. It is apparent that the resulting posterior distribution for $\boldsymbol{\theta}$ results in model predictions that agree well for both of these two configurations.

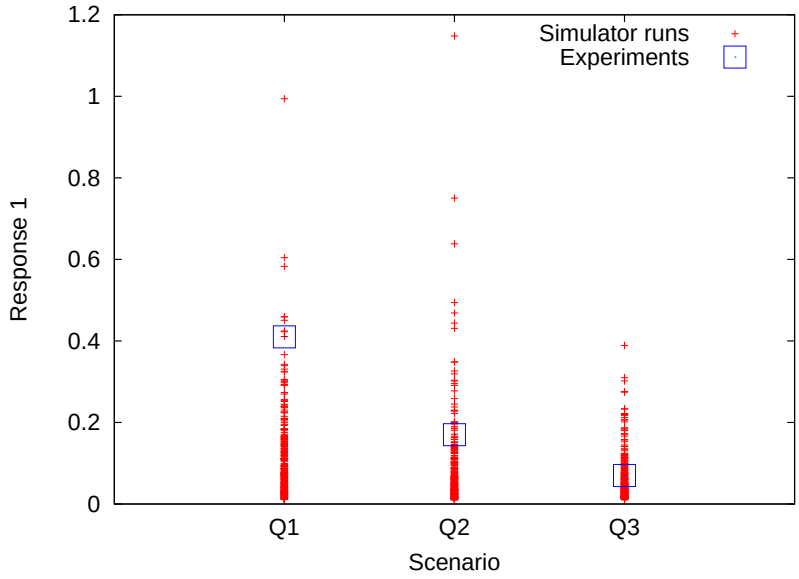


Figure 6.27: Predicted and measured values of response 1, for each of the three configurations

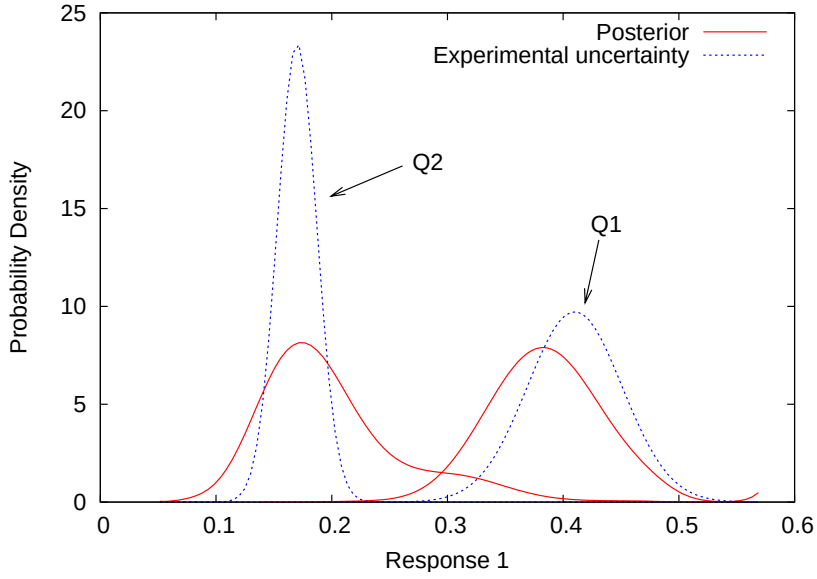


Figure 6.28: Posterior predictive distributions for responses 1 at Q1 and Q2 resulting from the calibration based on both the Q1 and Q2 measurements

The next step is to attempt the calibration based on both the Q1 and Q3 data. As was previously discussed regarding Figure 6.27, there is less of an expectation that this calibration will be successful because there appears to be a discrepancy between these two scenarios. This analysis is repeated as before, but the values of σ_i are now set to 20% (as opposed to 10% previously) of the corresponding observations to allow for more leeway in the calibration. The resulting posterior predictive distributions are plotted in Figure 6.29.

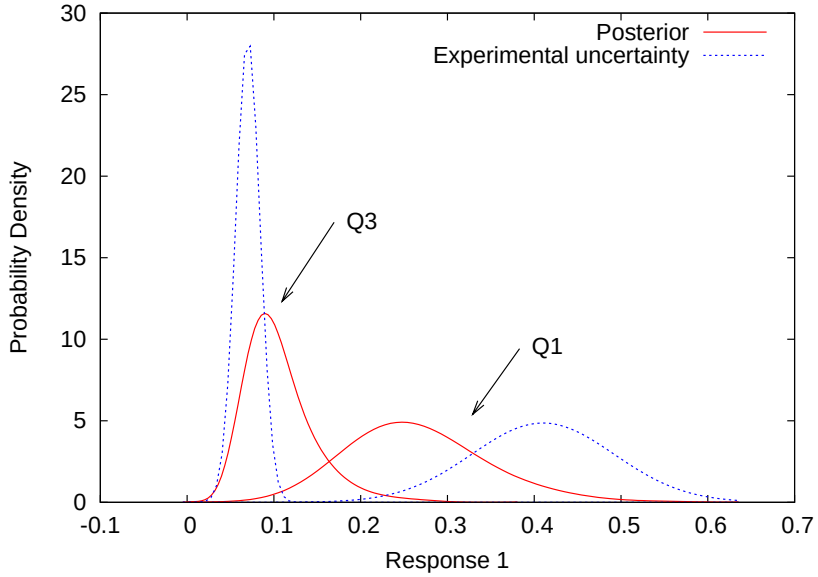


Figure 6.29: Posterior predictive distributions for responses 1 at Q1 and Q3 resulting from the calibration based on both the Q1 and Q3 measurements

The results indicate that, as expected, the simulator does not calibrate nearly as well to the Q1 and Q3 scenarios simultaneously. While there is not a complete mismatch between the predicted distributions and the experimental uncertainty distributions, the agreement is not nearly as good as before. In particular, the simulator has maintained its prior tendency to under-predict the response for the Q1 scenario and over-predict it for the Q3 scenario.

While the results for the calibration based on both Q1 and Q3 suggest that there may be an inadequacy associated with the simulator itself, it was later revealed that some of the data used for this analysis were faulty. In particular, it was later found that the simulator data provided

for the Q3 scenario was incorrect, which confirms the suspicion that the given data can not be made to calibrate in terms of both Q1 and Q3 simultaneously.

6.3.4 Further analysis of results

This section will briefly consider some cross-validation analyses of the calibration results. The purpose is to attempt to develop confidence in the interpretations and usages of the resulting input distributions.

Interpretation of posterior distributions

First, consider the fact that one-dimensional marginal posterior distributions (see, for example, Figures 6.20 and 6.21) for the calibration inputs only provide summary information, and do not tell the entire story of the joint posterior distribution. Various linear and nonlinear dependencies can exist among the inputs, and the fact that the posterior is not well approximated by a multivariate normal distribution further complicates its representation (in which case, even marginal distributions and pairwise correlation coefficients do not fully specify the distribution). The point is that while marginal distributions (and other summaries) are useful for graphical display of certain features of the posterior distribution, the only reliable representation of the full joint posterior is given by the MCMC samples.

For example, consider what information is lost when one attempts to ignore the dependencies among the calibration parameters in the posterior distribution (the pairwise linear correlations are not large, the strongest being only $\rho = -0.29$). Figure 6.30 illustrates the predictive distribution obtained when independent beta distributions are used to represent the posterior. Although the location of the response does not change too much (on average, the predictive distribution agrees with the observation), the uncertainty has increased significantly. This example

illustrates how much information can be lost by ignoring (even seemingly weak) dependencies.

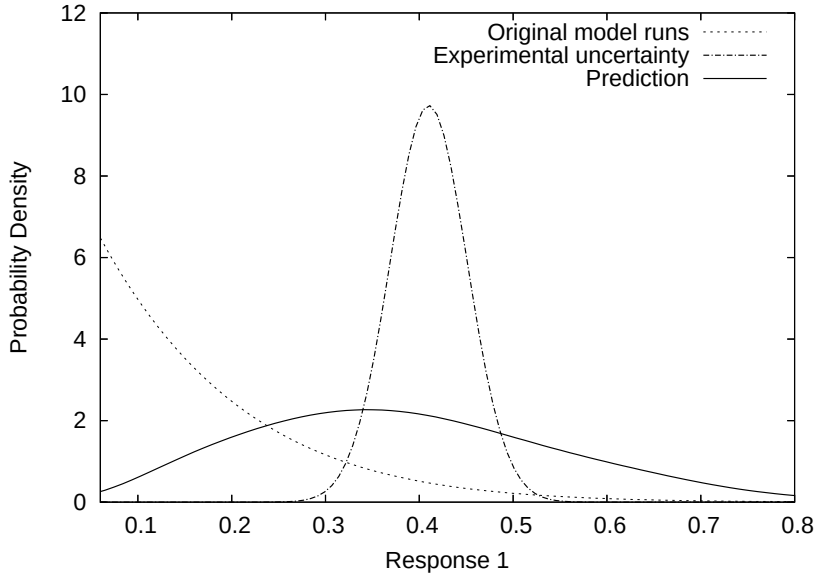


Figure 6.30: Predictive distribution obtained when the dependencies among the calibrated inputs are ignored

Cross-validation

This section discusses the results of three cross-validation exercises which test the simulation's ability to predict one result when calibrated using another. The nominal (Q1, response 1) model is used as the baseline, and the objective is to test the predictive capability of calibration inputs that are estimated using observations based on other time responses or configurations. It is first noted, however, that as shown in Figures 6.25 and 6.27, there may be a scenario-dependent component to the simulator bias that can not be captured via the calibration inputs. As such, this cross-validation is a test of both the calibration process and the simulator itself, since poor cross-validation results may indicate simulator inadequacy.

Specifically, three cases of cross-validation are considered (in each case, the objective is to predict the observed value for configuration Q1, response 1):

(a) Inputs are estimated using the observed value of response 1 for the Q2 configuration

(b) Inputs are estimated using the observed value of response 1 for the Q3 configuration

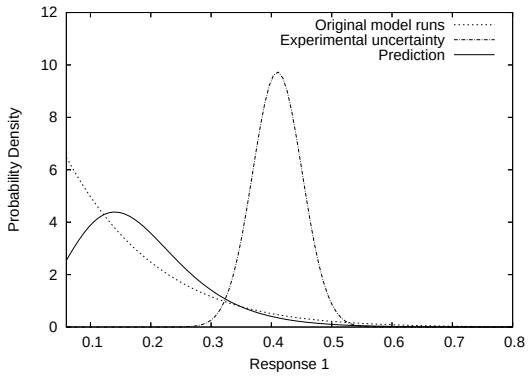
(c) Inputs are estimated using the observed value of response 4 for the Q1 configuration

The corresponding predictive distributions are shown in Figure 6.31. Unfortunately, the calibrated input distributions do not give accurate performance for predicting scenarios or response measures other than those with which they were calibrated. This could be an indication that the physics simulation is not modeling the experimental data correctly. However, it is also important to keep in mind that in each case, the predictive distribution is computed using the GP surrogate for the nominal case, whereas the calibrated input distributions were computed using different surrogates (which surrogates also used different training data). As a result, it is very possible that poor cross-validation agreement could simply be because the various GP models were trained using different inputs (in any case, the use of the same design of computer experiments for each configuration is recommended, as it would allow for more compatibility among the various surrogate models).

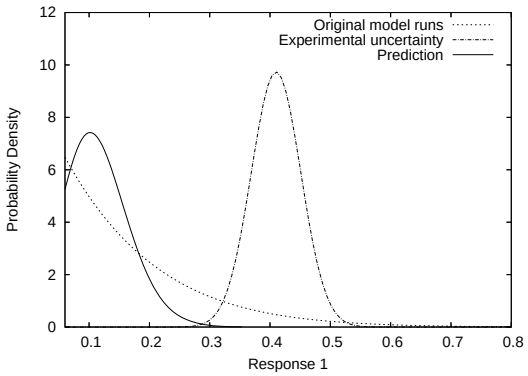
Also of note is that as mentioned previously, it was revealed after the analysis that the provided simulator data for the Q3 configuration were erroneous. As such, a successful cross-validation between the Q1 and Q3 configurations would not be expected.

6.3.5 Conclusions

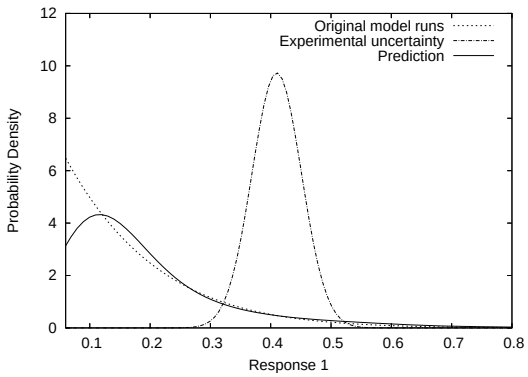
Even though the cross-validation results from this case study do not support the conclusion that the calibrated simulation model is also "validated" (and it was discovered after the fact that some of the data used for this analysis were faulty), there is still much to learn from this case study from a research perspective. For example, the use of a powerful (in this case Gaussian process) surrogate modeling technique can enable comprehensive calibration analysis and



(a) Inputs calibrated based on Q2, response 1



(b) Inputs calibrated based on Q3, response 1



(c) Inputs calibrated based on Q1, response 4

Figure 6.31: Resulting predictive distributions when various calibrated input distributions are used to predict response 1 for configuration Q1

parameter exploration, even when the knowledge of the relationship between the simulator inputs and outputs is limited to previously observed code runs.

While only mild success is achieved here in attempting to make use of data from multiple system configurations and/or response measures simultaneously for calibration, this is an important pursuit, since most analyses will want to make use of all available data when calibrating a simulation. In particular, the principal component-based approach presented in the first part of Section 6.3.3, while not particularly powerful for decomposing response quantities that are already low-dimensional (as here), is envisioned to be a a useful tool for calibrating simulators based on time-series output (an alternative approach is illustrated via the case study of Section 6.4).

Another interesting conclusion is that the interpretation and presentation of the results of the calibration analysis are not trivial. As discussed in Section 6.3.4, ignoring the full correlation structure of the updated distributions will result in a large overestimate of the uncertainty. Given the importance of the joint structure, marginal posterior distributions and confidence intervals should be used with care. Additionally, it is difficult to comprehensively express or summarize the posterior distribution when it is high-dimensional. For example, it is difficult to visualize joint distributions and their confidence regions in more than two dimensions. Thus, there exists the potential for future work to address the task of constructing useful summary statistics based on random samples (in the case of a posterior distribution that is constructed with MCMC sampling) of a high-dimensional random variable.

6.4 Bayesian model calibration: thermally decomposing foam

6.4.1 Introduction

A series of experiments have been conducted at Sandia National Laboratories in an effort to support the physical characterization and modeling of thermally decomposing foam (Erickson et al., 2004). An associated thermal model is described in by Romero et al. (2006). The system considered here, often referred to as the "foam in a can" system, consists of a canister containing a mock weapons component encapsulated by a foam insulation. Several illustrations of this setup are shown in Figures 6.32 and 6.33.

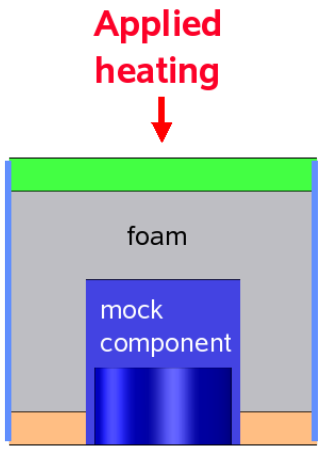


Figure 6.32: Schematic of the "foam in a can" system

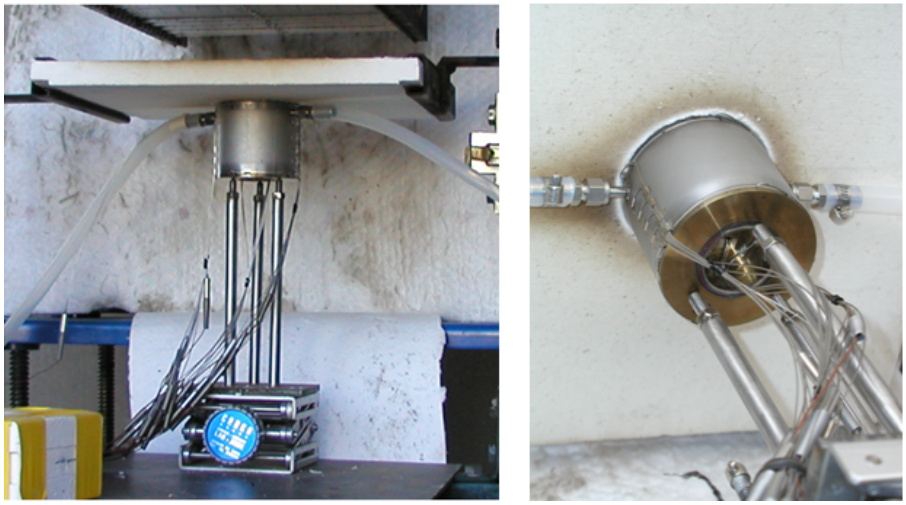


Figure 6.33: Experimental setup

The simulation model is a finite element model developed for simulating heat transfer through decomposing foam. The model contains roughly 81,000 hexahedral elements, and has been verified to give spatially and temporally converged temperature predictions. The heat transfer model is implemented using the massively parallel code Calore (CALORE-MAN), which has been developed at Sandia National Laboratories under the ASC (Advanced Simulation and Computing) program of the NNSA (National Nuclear Security Administration).

The simulator has been configured to model the "foam in a can" experiment, but several of the input parameters are still unknowns (either not measured or not measurable). In particular, five calibration parameters are considered first: q_2, q_3, q_4, q_5, and FPD. The parameters q_2 through q_5 describe the applied heat flux boundary condition, which is not well-characterized in the experiments. The last calibration parameter, FPD, represents the foam final pore diameter, and is the parameter of most interest, because it will play a role in the ultimate modeling and prediction process. The calibration analysis will be based on the empirically observed temperature response of the system from 0 to 2200 seconds at nine different locations on the structure (six external and three internal).

6.4.2 Preliminary analysis

The first step is to collect a database of simulator runs for different values of the calibration parameters, from which the surrogate model will be constructed. Ideally, the design of computer experiments should provide good coverage for the posterior distribution of the calibration inputs. However, since the form of the posterior is not known beforehand, it is necessary to begin with an initial guess for the appropriate bounds. Fortunately the Bayesian method provides feedback, so if the original bounds are not adequate, they can be revised appropriately. This type of sequential approach has previously been used for Bayesian model calibration and

other studies (Kennedy and O'Hagan, 2001; Bernardo et al., 1992; Craig et al., 1996; Aslett et al., 1998).

The DAKOTA (Eldred et al., 2006) software package is used for the design and collection of computer experiments. DAKOTA is an object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis that can be configured to interface with the thermal simulator via external file input/output and a driver script. For the initial design, DAKOTA is used to generate an LHS sample of size 50 using the variable bounds listed in Table 6.6.

Table 6.6: Original design of computer experiments

Variable	Lower bound	Upper bound
FPD	2.0×10^{-3}	15.0×10^{-3}
q_2	25,000	150,000
q_3	100,000	220,000
q_4	150,000	300,000
q_5	50,000	220,000

The Bayesian calibration using these bounds illustrates that some adjustment to the bounds would be useful, because the resulting posterior distribution directly indicates which regions of the parameter space are feasible, including whether or not the parameter space should be expanded in the subsequent design. Thus, a new LHS sample of size 50 is constructed using the revised design described in Table 6.7. The revised bounds are chosen so that they will cover the entire range of the posterior distribution for the calibration inputs.

Table 6.7: Revised design of computer experiments

Variable	Lower bound	Upper bound
FPD	4.0×10^{-3}	6.0×10^{-3}
q_2	25,000	150,000
q_3	0	200,000
q_4	100,000	400,000
q_5	120,000	160,000

Using the results from the simulation runs, a first check is to compare the ensemble of pre-

dicted time histories against the experimental time histories to see if the experimental data are "enveloped" by the simulation data. Figures 6.34 and 6.35 compare the envelope of simulator outputs against the experimental data for locations 1 and 9, respectively. In general, the experimental observations are enveloped by the simulator outputs, although at locations 5 and 6, the experimental response exceeds the maximum of the simulator outputs for $t < 800$ seconds, as seen in Figure 6.36.

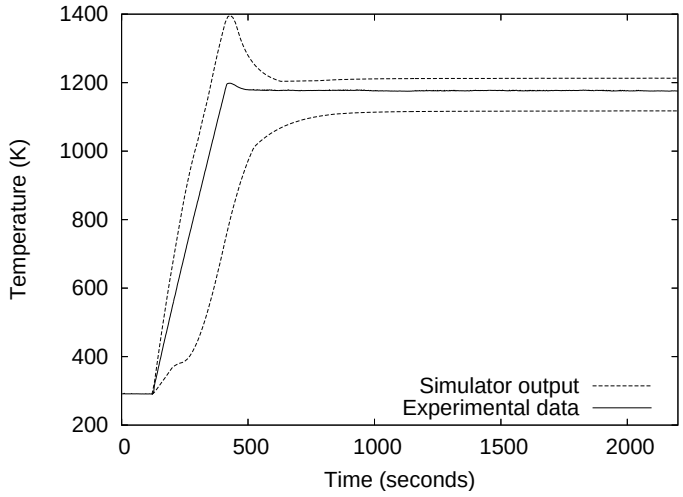


Figure 6.34: Temperature response comparison for envelope of 50 simulator outputs with observed data for location 1 (average lid temperature)

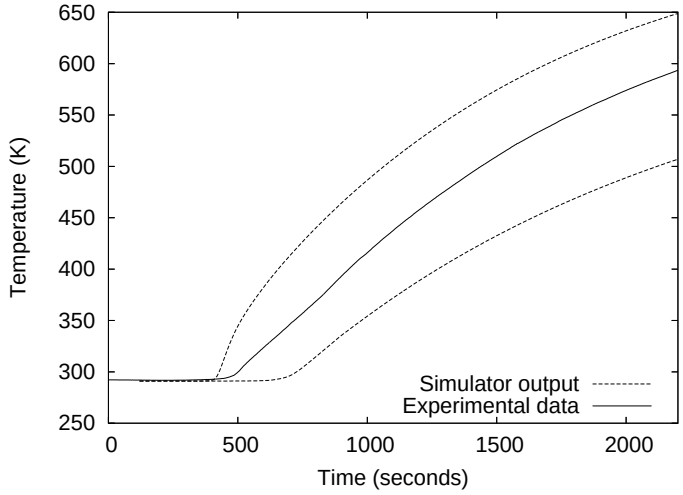


Figure 6.35: Temperature response comparison for envelope of 50 simulator outputs with observed data for location 9 (internal thermocouple)

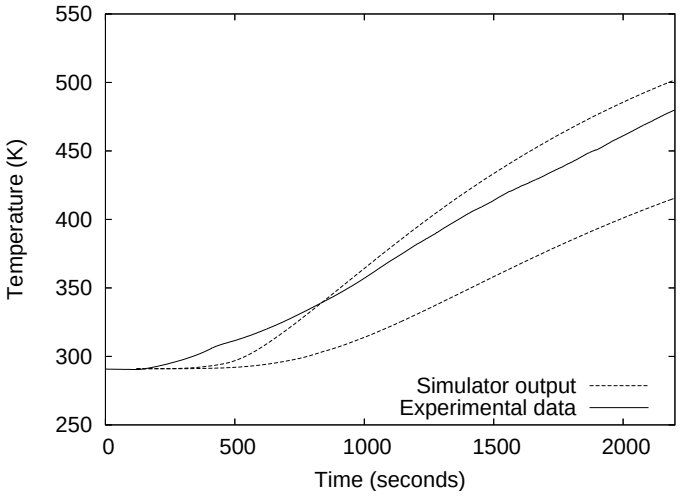


Figure 6.36: Temperature response comparison for envelope of 50 simulator outputs with observed data for location 6 (average of thermocouples 13 through 16)

6.4.3 Bayesian calibration analysis: nominal case

This section presents a "nominal" Bayesian calibration analysis of the CALORE simulator using data from all nine "locations" of interest. Some of these "locations" (for example, location 1) are averages of multiple thermocouple readings, while others represent single thermocouple readings. The application of the Bayesian calibration extensions discussed in Sections 5.3.2 and 5.3.3 will be presented in Sections 6.4.6 and 6.4.7.

The variance of ε in Eq. (5.11), σ^2, is not considered as a function of time or location. It would be straightforward to incorporate a parametric dependence for the variance on temporal or spatial coordinates if such a formulation were desired. Nevertheless, σ^2 is treated as an object of Bayesian inference, making use of the standard reference prior (Lee, 2004):

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}. \tag{6.29}$$

The prior distribution for $\boldsymbol{\theta}$ is taken to be independently uniform, as in Eq. (2.5), where the initial bounds for each parameter are as listed in Table 6.6. After revising the design of

computer experiments, the prior bounds are adjusted to reflect those listed in Table 6.7.

Each of the nine "locations" are modeled separately with two independent surrogates representing the response before and after 500 seconds, which results in a total of 18 surrogate models for the simulator output. Multiple Gaussian process surrogate models are used because a single stationary Gaussian process representation of the response at all locations and time instances does not seem to be appropriate. The choice of dividing the surrogates at 500 seconds is admittedly subjective (and a more comprehensive approach might choose different time divisions for different locations), but on average for the different locations, there is a significant change in the response behavior around 500 seconds (for example, the process variance increases; see Figures 6.34, 6.35 and 6.36).

For each surrogate, the point selection algorithm discussed in Section 3.4 is implemented to select an optimal subset of points with which to build the surrogate. At each location, the first surrogate is based on 75 points chosen optimally from the 1,950 available points (39 time instances $\times$ 50 LHS samples), while the second is based on 100 points chosen optimally from 8,550 points. It should be emphasized that the process for constructing these surrogate models is not trivial: the iterative MLE process described in Section 3.4 is applied separately for each of 18 surrogate models. This results in approximately 3,000 numerical MLE optimization problems in six dimensions, which is why an efficient MLE scheme is critical, and the use of gradient information, as discussed in Section 3.3, can be very important.

For the experimental data, 21 points evenly spaced at time intervals of 100 seconds are used for each of the 9 locations. The MCMC simulation is adjusted appropriately and run for 100,000 iterations. The resulting marginal posterior distributions for the two parameters of most interest, FPD and q_5, are shown in Figures 6.37 and 6.38, where the plotting ranges are

representative of the bounds of the prior distribution. Recall that the prior distribution for $\boldsymbol{\theta}$ is independently uniform over the ranges listed in Table 6.7.

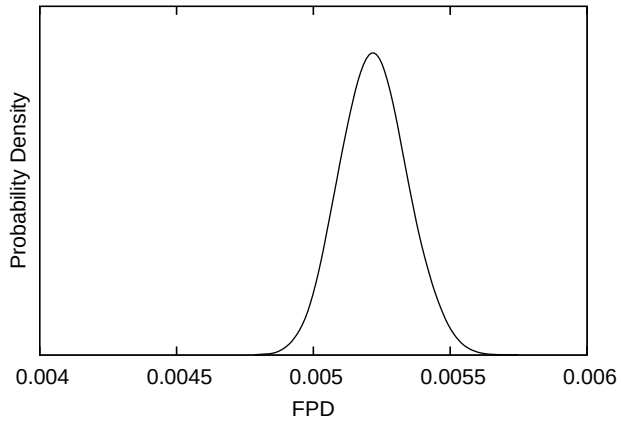


Figure 6.37: Posterior distribution of FPD (x-range represents prior bounds)

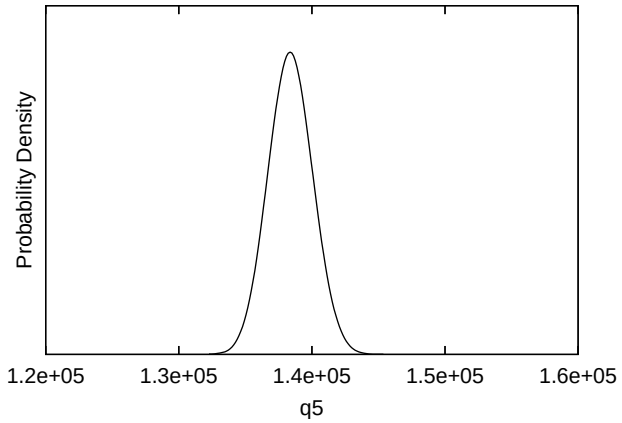


Figure 6.38: Posterior distribution of q_5 (x-range represents prior bounds)

The statistics of the marginal posteriors are given in Table 6.8, and the pairwise correlation coefficients are given in Table 6.9. The correlation coefficients indicate a strong negative relationship between q_2 and q_3, as well as moderate negative relationships between FPD and q_5, and q_3 and q_4. For a more visual interpretation of these relationships, kernel density estimation (Silverman, 1986) can be used to visualize the two-dimensional density functions. For example, Figure 6.39 plots the 95% confidence region for FPD and q_5 based on a kernel density estimate to the two-dimensional posterior of these two variables.

Table 6.8: Posterior statistics based on the nominal calibration analysis

Variable	Mean	Std. Dev.
FPD	5.22×10^{-3}	1.17×10^{-4}
q_2	88,546	16,977
q_3	113,100	11,307
q_4	246,270	11,652
q_5	138,390	1,565

Table 6.9: Pairwise correlation coefficients within the posterior distribution for nominal analysis

	FPD	q_2	q_3	q_4	q_5
FPD	1.00	0.02	0.02	-0.25	-0.67
q_2	0.02	1.00	-0.80	0.18	-0.02
q_3	0.02	-0.80	1.00	-0.58	-0.01
q_4	-0.25	0.18	-0.58	1.00	0.00
q_5	-0.67	-0.02	-0.01	0.00	1.00

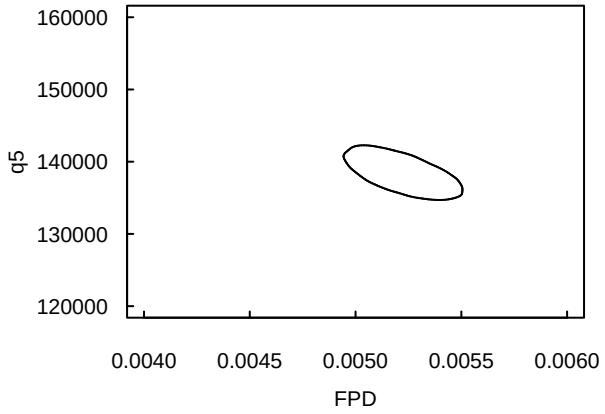


Figure 6.39: 95% confidence region for FPD and q_5. Plotting bounds represent prior bounds.

Finally, as a check on the surrogate models, the total RMS difference (see Eq. (6.30) below) between the surrogate output and the true simulator output is computed at the posterior mean of the calibration inputs. This RMS difference is found to be only 2.4 K, which suggests that the surrogates have accurately captured the relationship between the simulator inputs and outputs. Figure 6.40 illustrates how the surrogate compares to the actual simulator output at location 9. The discrepancy is visually almost indistinguishable. The experimental observations have been plotted as well, for illustration.

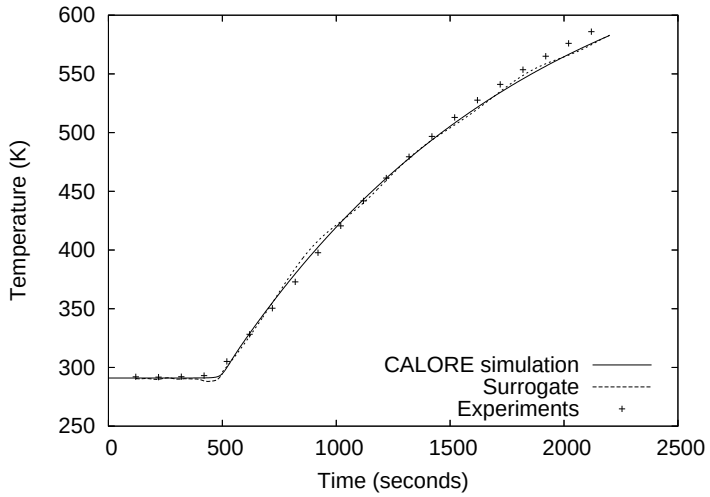


Figure 6.40: Comparison of surrogate model output to actual CALORE output for location 9, based on the posterior mean of the calibration inputs.

6.4.4 Comparison to classical parameter estimation

This section briefly discusses how the results of the above nominal Bayesian calibration analysis compare to the results obtained using the methods of nonlinear regression analysis. As discussed in Section 5.2, classical nonlinear regression analysis provides methods to compute a point estimate to the calibration parameters (referred to as $\hat{\boldsymbol{\theta}}$), as well as various types of confidence regions to summarize the uncertainty in this estimate. Note that the nonlinear regression approach is often attractive because it is based on the minimization of a simple error

measure between the predictions and observations (although see Section 5.5, which discusses the theoretical connection between Bayesian and least-squares estimation).

Two analyses are considered here: first, a global search algorithm is interfaced directly with the physics simulation in order to minimize the sum of squared errors. Second, the same surrogates created for the Bayesian analysis are used in conjunction with a gradient-based algorithm to compute both a least-squares estimate and associated confidence regions.

DIRECT approach

The purpose of this analysis is to develop an alternative estimate of the calibration parameters that does not make use of surrogate models or Bayesian inference, but employs a comparable number of simulator evaluations. This will allow the results of Section 6.4.3 to be gauged on an objective basis.

In order to quantify the accuracy of the Bayesian estimate when only a small number of simulator runs are available, the posterior mean is considered, based on the analysis with the original bounds for the calibration parameters (Table 6.6), which analysis corresponded to only 50 runs of the simulator. The measure of agreement considered here will be the sum of squared errors between the predictions and observations. This measure can also be expressed in terms of what is known as the root-mean-squared (RMS) error, which has the same units as the data and is computed as

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - G(\boldsymbol{\theta}, \boldsymbol{s}_i)\right)^2}. \tag{6.30}$$

The RMS agreement between the simulator predictions at the Bayesian posterior mean and the experimental data is 19.4 Kelvin (based on an actual simulator run, not the surrogate approximation).

The alternative estimate of the calibration parameters is obtained by attempting to minimize the sum of squared errors (equivalently, the RMS). This minimization is conducted here using the global optimization algorithm DIRECT (Jones et al., 1993), and the optimization algorithm interfaces directly with the simulation code, so that no surrogate models are used. The convergence criteria are adjusted to limit the number of objective function evaluations (equivalently, runs of the CALORE simulation) to a number comparable to that used in the Bayesian calibration analysis (50). In order to keep the comparison fair, the DIRECT algorithm is provided with the same variable bounds that were available to the Bayesian analysis (the prior bounds, listed in Table 6.6). As mentioned above, the reason for doing this analysis is so that the results can be used as a comparison against the more complicated Bayesian approach: it is easy to understand the motivation for picking the calibration parameters by minimizing the sum of squares error measure, but does this result in a more accurate point estimate than the surrogate-based Bayesian approach (using a comparable number of total function evaluations)?

After 65 function evaluations, the DIRECT algorithm reduces the RMS error to 32.3 Kelvin, which is significantly worse than the RMS error achieved using the Bayesian approach (19.4 Kelvin at the Bayesian posterior mean). These results suggest that while the Bayesian approach provides a comprehensive framework for representing uncertainty in the parameter estimates, the Bayesian framework is still capable of providing an efficient (in terms of number of simulator runs) means of obtaining accurate point estimates to the calibration parameters, using a comparable number of simulator evaluations. It is acknowledged that a surrogate-based optimization approach might be preferred to interfacing directly with the expensive simulator, and such an approach is presented below.

Surrogate-based approach

While the above least squares analysis is presented to emphasize the point estimation accuracy of the surrogate-based Bayesian approach, this section presents a comparison of the uncertainty quantification capabilities of the classical and Bayesian approaches.

The nonlinear regression analysis applied here makes use of the same Gaussian process surrogate approximations to the simulator, and the same experimental data that were used in the Bayesian analysis. One difference is that there is not a natural way of accounting for the uncertainty introduced by the surrogates in the classical framework (because the surrogate uncertainty is a function of $\boldsymbol{\theta}$); however, for this particular application the surrogate uncertainty is relatively small, so it is not expected to have a large effect on the parameter estimation uncertainty.

As described in Section 5.2, the point estimate $\hat{\boldsymbol{\theta}}$ is found by minimizing the sum of squares function, $S(\boldsymbol{\theta})$. Following the preceding Bayesian analysis, the errors ε_i are taken to be independently and identically distributed, so that weights are not needed and the sum of squares representation given by Eq. (5.4) becomes the objective function (as with the above "DIRECT approach").

The minimization of Eq. (5.4) can be made more efficient by incorporating gradient information associated with $G(\boldsymbol{\theta}, \boldsymbol{s}_i)$. This is typically expressed in terms of the Jacobain matrix, which is given by

$$\boldsymbol{J} = \begin{bmatrix} \dfrac{\partial G(\boldsymbol{\theta}, \boldsymbol{s}_1)}{\partial \theta_1} & \cdots & \dfrac{\partial G(\boldsymbol{\theta}, \boldsymbol{s}_1)}{\partial \theta_p} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial G(\boldsymbol{\theta}, \boldsymbol{s}_n)}{\partial \theta_1} & \cdots & \dfrac{\partial G(\boldsymbol{\theta}, \boldsymbol{s}_n)}{\partial \theta_p} \end{bmatrix}. \tag{6.31}$$

Fortunately, since $G(\cdot, \cdot)$ is being approximated with Gaussian process interpolation, all of the

corresponding partial derivatives can be obtained analytically, using Eq. (3.11).[5] The numerical method used here to find the minimizer of the sum of squares function is a Levenberg-Marquardt method, as described in Seber and Wild (2003); in particular, the implementation of More' et al. (1999) is employed.

Both the linear approximation and "F-test" methods are used to obtain confidence regions for the estimated parameters. As in the previous section, confidence regions are computed for the two-dimensional parameter subset $\boldsymbol{\theta}_2 = (FPD, q_5)$ to enable visualization. In order to take appropriate account of the three nuisance parameters in computing these two-dimensional confidence regions, inequalities (5.9) and (5.10) are employed for the linear and "exact" methods, respectively. Note that each evaluation of inequality (5.10) requires a minimization of the sum of squares function over the nuisance parameters; as such, construction of a confidence region subset using the exact method can be very expensive, underlying the need for an efficient minimization routine. However, it is found that for this application the linear approximation and exact confidence regions provide indistinguishable results; this is not surprising, though, because this study has a large number of experimental observations, and the two confidence regions are asymptotically equivalent.

A comparison of the Bayesian and classical results is given in Figure 6.41, which shows the simultaneous 95% confidence region for FPD and q_5 constructed using each approach. The corresponding point estimates obtained using each approach are also plotted (for the Bayesian analysis, the point estimate is the posterior mean). Clearly, both the point estimates and the confidence regions obtained using the two approaches agree very closely. This is expected, however, since both the Bayesian approach (assuming an appropriately vague prior distribu-

[5] Although some analysts might be tempted to do so, the use of finite differencing to obtain approximate gradients for a response quantity being modeled using Gaussian process interpolation should always be avoided. This is because for most realistic finite difference step sizes, there is a substantial possibility that numerical error will dominate the computations, resulting in unreliable gradient estimates.

tion) and the nonlinear regression "exact" approach define confidence regions in terms of contours of the likelihood function (the only difference between the two likelihood functions is that the Bayesian analysis has included the surrogate uncertainty in the likelihood, as in Eq. (5.14); for this particular problem, though, the magnitude of the surrogate uncertainty is fairly small).

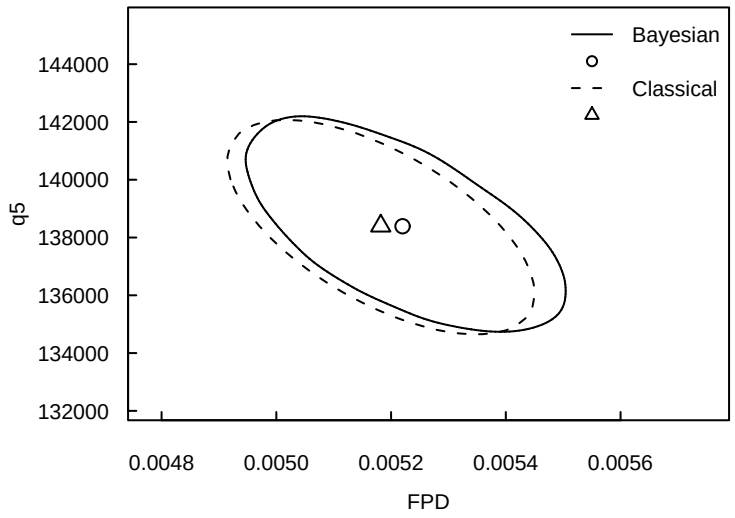


Figure 6.41: Comparison of Bayesian and classical results showing 95% confidence regions and point estimates for FPD and q_5 constructed using each approach.

While the visualizations of the results obtained using each approach are basically indistinguishable, there are differences that recommend the use of the Bayesian approach. One advantage of the Bayesian approach is that when MCMC sampling is used to construct the posterior distribution, the resulting samples can be used in a variety of ways. With the samples, the computation of quantitative summary statistics and the marginalization over nuisance parameters are trivial, whereas the corresponding classical computations can be quite cumbersome. Further, the samples can be propagated to new analyses in order to aid in the quantification of uncertainty associated with new simulator predictions.

In addition to the above considerations, the Bayesian approach is readily extensible to account for additional uncertainty sources. Several extensions to the nominal calibration analysis

are discussed in the following sections. Note that the use of a correlated error model, discussed in Section 6.4.5, can be handled nicely in both the Bayesian and nonlinear regression frameworks; however, the explicit treatment of the parameters governing the error autocorrelation model as additional unknowns is not accommodated by the nonlinear regression framework. Further, two additional extensions are presented in Sections 6.4.6 and 6.4.7 that showcase the flexibility of the Bayesian approach.

6.4.5 Accounting for correlated errors

The "nominal" calibration analysis of Section 6.4.3 is based on a probabilistic model that treats all of the errors, ε_i, as being independent. While this is often a reasonable assumption, its validity comes into question when a response value is observed at closely spaced points in time or space. Since this particular case study deals with a response quantity that is observed at multiple time instants and locations, the effect of considering a dependency structure for the errors is discussed here.

While a comprehensive analysis of this data set would consider correlations in the response for points that are closely spaced in both time *and* space, only autocorrelation in time is considered here. However, this calibration framework is certainly flexible enough to consider correlations in space as well, and one might include such correlations by first quantifying the geometric coordinates of each of the nine locations on the structure for which observations are available.

Consider Figures 6.42 and 6.43, which plot the residuals from the nominal analysis (at the posterior mean of the calibration inputs) as a function of time, for the response at locations number one and nine. Clearly, in each case there is a significant amount of serial correlation among the residuals, and the assumption that the errors are independent is certainly not valid

for this particular analysis.

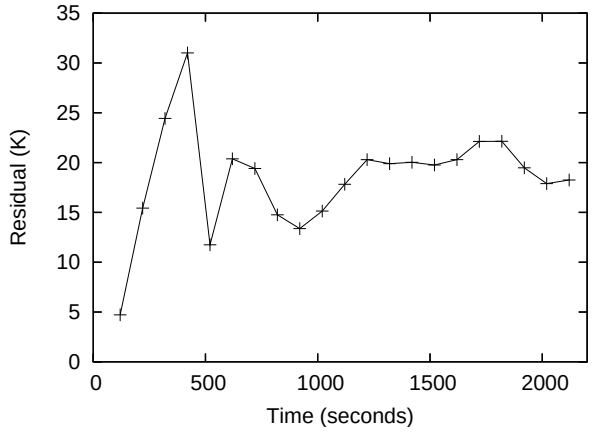


Figure 6.42: Residuals from the nominal analysis (at the posterior mean of the calibration inputs) at location number one

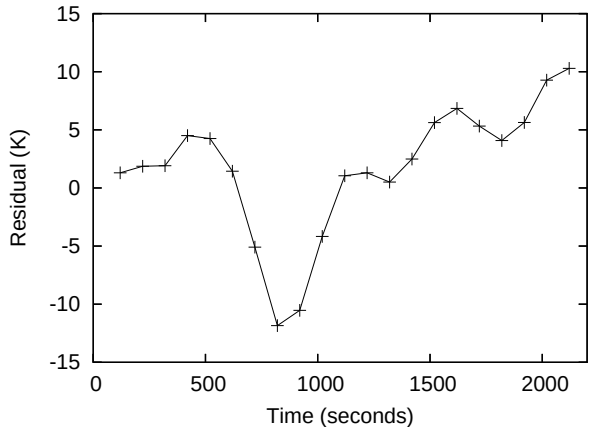


Figure 6.43: Residuals from the nominal analysis (at the posterior mean of the calibration inputs) at location number nine

In order to illustrate one approach for accounting for serially correlated errors, a first-order auto-regressive model, known as an "AR(1)" model, is adopted to model the autocorrelation in the errors over time. This model has the form

$$\varepsilon_i = \phi\varepsilon_{i-1} + \nu_i, \tag{6.32}$$

where the correlation parameter ϕ satisfies $|\phi| < 1$, and the "innovation errors" are indepen-

dently and identically distributed as $\nu_i \sim N(0, \sigma_\nu^2)$. The parameter ϕ describes the correlation that exists between ε_i and ε_{i-1} (i.e. the "lag-1" autocorrelation), and the correlation between any two errors is given by

$$\rho_{i,j} = \phi^{|i-j|}. \tag{6.33}$$

Also, the error variance is given by

$$\text{Var}\left[\varepsilon_i\right] = \frac{\sigma_\nu^2}{1-\phi^2}. \tag{6.34}$$

Note that in order to use this model, the observations must be spaced evenly in time (although see Glasbey, 1979, for an extension that allows for unequally spaced time intervals). For this study, the errors for observations at different locations on the structure are still assumed to be independent of each other, but the errors at the same location will be modeled using the AR(1) model of Eq. (6.32). The same autocorrelation model is used for all nine locations.

In the Bayesian calibration framework, the adoption of this error model is achieved by using the AR(1) model to construct the error covariance matrix. Let the full data covariance matrix associated with the likelihood function of Eq. (5.14) be denoted $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{AR} + \boldsymbol{\Sigma}_{GP}$, where $\boldsymbol{\Sigma}_{AR}$ is the error covariance matrix, and $\boldsymbol{\Sigma}_{GP}$ contains the covariance associated with the Gaussian process surrogates. The elements of $\boldsymbol{\Sigma}_{AR}$ that correspond to errors from the same location on the structure are computed as

$$\text{Cov}\left[\varepsilon_i, \varepsilon_j\right] = \frac{\sigma_\nu^2}{1-\phi^2}\phi^{|i-j|}. \tag{6.35}$$

Because this analysis is treating errors at different locations on the structure as being independent, those elements of $\boldsymbol{\Sigma}_{AR}$ that correspond to different locations on the structure are zero.

One of the main difficulties in adopting this model (or any parametric model for the error dependencies) is that it adds an additional parameter that must be estimated, ϕ (both this approach and the previous nominal approach require the estimation of one error variance term). Outside of a Bayesian framework, ϕ is sometimes estimated using maximum likelihood or iterative methods such as "two-stage" estimation[6] (Seber and Wild, 2003). Fortunately, within the Bayesian framework, the fact that ϕ is an unknown can be handled naturally by treating ϕ as an additional object of Bayesian inference. For this analysis, ϕ and σ_ν^2 are given the vague reference prior distribution

$$\pi(\phi, \sigma_\nu^2) \propto \begin{cases} 1/\sigma_\nu^2, & |\phi| < 1, \\ 0, & |\phi| \geq 1, \end{cases} \tag{6.36}$$

which is independently uniform in ϕ on $(-1, 1)$ and uniform in $\log \sigma_\nu^2$.

The posterior mean for the autocorrelation parameter, ϕ, is 0.985, and its posterior distribution is shown in Figure 6.44. The resulting statistics of the posterior distribution for $\boldsymbol{\theta}$ are given in Table 6.10, and a comparison of the joint posterior of FPD and q_5 with the nominal results is shown in Figure 6.45. Clearly, accounting for correlated errors has largely increased the amount of uncertainty present in the resulting inference about the parameter of interest, FPD (although the uncertainty in the nuisance parameters has actually decreased). An increase in uncertainty is to be expected, though, because if the errors truly are dependent, then there is less information present in the experimental data.

In addition to an increase in the posterior uncertainty for FPD (and q_5), the inclusion of

[6] In two-stage estimation, the errors are first assumed to be independent, and the resulting residuals at the least-squares estimate, $\hat{\boldsymbol{\theta}}$, are used to obtain an estimate of ϕ. The analysis is then repeated using the AR(1) model and the estimated value of ϕ to obtain a new estimate of $\boldsymbol{\theta}$. In fact, this procedure can be iterated further to refine the estimates of ϕ and $\boldsymbol{\theta}$.

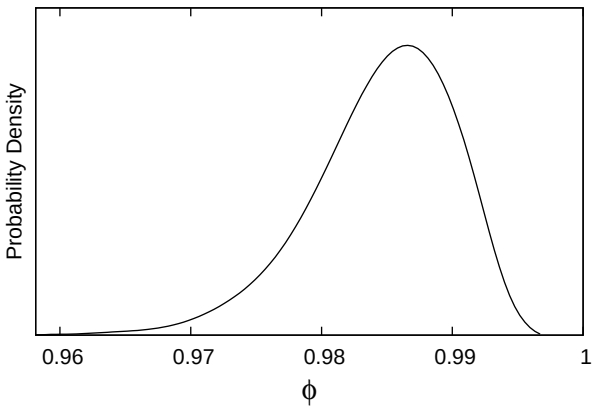


Figure 6.44: Posterior distribution for autocorrelation parameter, ϕ

Table 6.10: Posterior statistics accounting for autocorrelated errors

Variable	Mean	Std. Dev.
FPD	4.55×10^{-3}	2.05×10^{-4}
q_2	80,164	4,844
q_3	114,320	3,678
q_4	247,630	3,952
q_5	150,920	3,315

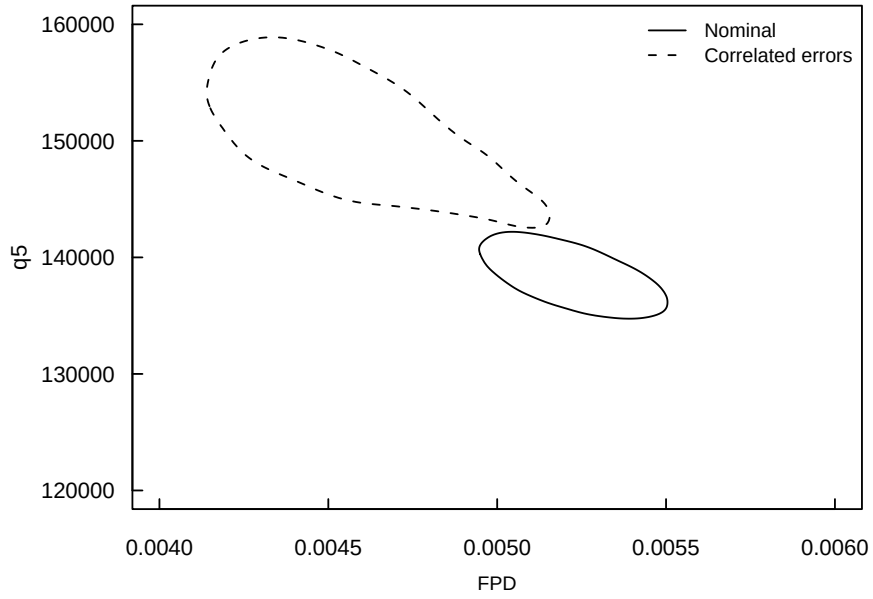


Figure 6.45: Illustration of the effects of accounting for correlated errors on the joint posterior distribution of FPD and q_5 (95% confidence regions)

the correlated error model has resulted in a shift in the location of the posterior distribution. In fact, it is apparent from Figure 6.45 that the posterior distribution is being affected by the bounds for FPD and q_5, and a subsequent analysis with larger bounds (requiring a new design and analysis of computer experiments) would be recommended.

Finally, note that the purpose of this study is not to illustrate a comprehensive autocorrelation analysis, but instead to emphasize the flexibility of the Bayesian calibration framework. In fact, the first-order autoregressive model is one of the simplest choices, and much more general models can be achieved, such as higher-order autoregressive moving average (ARMA) models. Specialized techniques also exist for choosing the orders of an ARMA process, which often involve constructing the sample autocorrelation and partial autocorrelation functions. While there is an extensive body of literature dealing with time-series analysis, Seber and Wild (2003) provide a comprehensive and practical coverage of the relevant issues for nonlinear regression models. In addition, Glasbey (1979, 1980) provide excellent examples of applied work in nonlinear regression with autocorrelated errors.

6.4.6 Accounting for characterized measurement uncertainty

This section illustrates the approach proposed in Section 5.3.3 to account for characterized measurement uncertainty associated with the thermocouple readings. One would expect this addition to be reflected by a broadening of the posterior distribution of the calibration inputs. In addition, since some of the thermocouples are biased, a shift in the location of the posterior is also expected.

For the thermocouples on the sides and bottom of the structure (corresponding to "locations" two through six), the experimentalists characterize the measurement uncertainty as -2 to 0% (Nakos, 2004). Because this measurement uncertainty is bounded, uniform random vari-

ables are employed: $u_i \sim \text{Uniform}(-0.02 \times y_i,\, 0)$, which is a time-dependent percentage of the measured temperature, y_i (the distributions for the u_i are not to be confused with prior distributions, since the thermocouple error is not an object of Bayesian inference). As is apparent from Eq. (5.18), negative values of u correspond to measurements that underestimate the actual value.

The remaining thermocouples (corresponding to "locations" 1, 7, 8, and 9) are located internally within the system, so they do not share the same -2 to 0% error specification as the external thermocouples. For the internal thermocouples, the FEM simulation itself is used to estimate the measurement uncertainty. This is possible because these thermocouples are explicitly modeled in the FEM simulation,

The FEM model for each of the internal thermocouples contains an associated contact parameter, which represents the amount of contact between the thermocouple and the structure. By varying the contact parameter, one is able to use the simulator to estimate the magnitude of the effect that imperfect contact might have on the thermocouple reading. For this study, the contact parameter is varied from perfect contact to zero contact (with all other model parameters held constant) in order to assess the maximum possible effect of imperfect thermocouple contact on the thermocouple reading. Since this result also characterizes the thermocouple error in terms of bounds, the uniform distribution is again employed, but this time the uncertainty is characterized as $u_i \sim \text{Uniform}(-\delta_i,\, \delta_i)$, where δ_i is the difference between the simulator output for perfect and zero contact. Note that δ_i varies with time and thermocouple location. The internal thermocouple uncertainties are not largely dissimilar in magnitude to those of the external thermocouples: the maximum value of δ at location 9 (near the mock weapons component) is about 1.7% of the corresponding measured temperature value.

The use of the FEM model itself to characterize the instrumentation error/uncertainty in this manner may seem somewhat counterintuitive. However, it is justifiable in this case on the basis that the primary source of error with the internal thermocouple readings is believed to be due to imperfect contact. Since the FEM simulation is capable of modeling the thermal response at each thermocouple location for the range of possible contact values, the simulation can be used to assess the magnitude of the effect that imperfect contact may have on the thermocouple reading. An additional justification for doing this is that the simulation is only being used to predict relative changes in the temperature response between the two cases (perfect and no contact). As such, the absolute accuracy of the FEM model does not come into play, only its ability to model the relative effect that imperfect contact has on temperature.

Also, note that several of the "locations" are averages of multiple thermocouple readings. For example, location one is the average of four thermocouples mounted on the lid. In these cases the thermocouple measurement errors average as well, and the generation of random realizations from such averages is handled using simulation.

The resulting statistics of the posterior distribution for this case (based on 100,000 MCMC samples) are reported in Table 6.11, which indicates small shifts in the means and small increases in the variance. This is illustrated graphically for FPD and q_5 in Figure 6.46, which compares a contour of the posterior density with and without the effect of characterized measurement uncertainty.

Table 6.11: Posterior statistics based on the calibration analysis with characterized measurement uncertainty

Variable	Mean	Std. Dev.
FPD	5.27×10^{-3}	1.24×10^{-4}
q_2	87,477	16,723
q_3	116,900	12,223
q_4	242,680	12,864
q_5	140,290	1,647

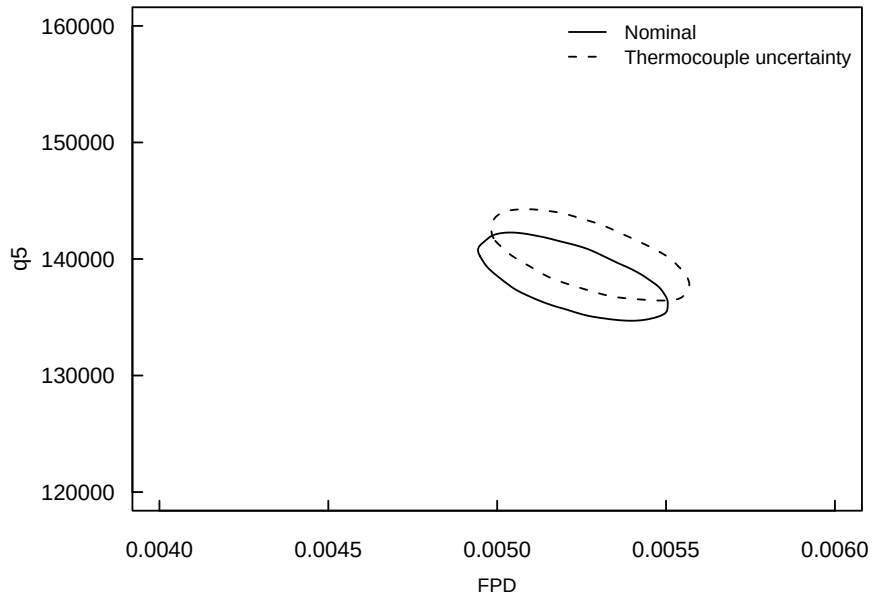


Figure 6.46: Comparisons of joint posterior distribution for FPD and q_5 with and without characterized thermocouple uncertainty (95% confidence regions)

The shift in the means for both q_5 and FPD is explainable in terms of the thermocouple uncertainty. Since q_5 represents applied heat flux, it is positively related to temperature response. Similarly, the negative correlation between q_5 and FPD suggests that the foam final pore diameter is also positively related to temperature response. Since the external thermocouples are known to provide readings that underestimate the actual temperature response, it is expected that accounting for this bias will result in an increase in the estimates for q_5 and FPD, and this is in fact what is seen.

6.4.7 Incorporating prescribed input uncertainties

This section extends the nominal analysis to include additional modeling uncertainties, as discussed in Section 5.3.2. Up to this point, only five model inputs have been considered as calibration inputs, but there are in fact many additional inputs to the simulator that are subject to uncertainty or lack of knowledge. Thus, this section will study the effect on the calibration

results when thirteen additional model inputs are treated as having prescribed uncertainties (in this case simply feasible bounds, represented by uniform probability density functions).

While it is also possible to treat these additional model inputs as calibration parameters, along with the original five, the primary reason for holding their uncertainties fixed is simply because there is an interest in knowing what effect this will have on the results. On the other hand, if they are treated as additional calibration parameters, their prior uncertainties may be reduced in light of the data $\boldsymbol{d}$, which would not give a picture of the effect of the prescribed uncertainties. Nevertheless, each of these analyses are conducted, as well as one "control" analysis, for comparison:

1. To make a fair comparison, the analysis is first conducted while holding the additional uncertain inputs fixed at their mean values. Although conceptually the same as the analysis discussed in Section 6.4.6, it is based on a different set of training data, and the surrogates must now model the relationship between the additional thirteen inputs and the response, which is expected to result in additional overall uncertainty.

2. Using the method outlined in Section 5.3.2, the analysis is performed while allowing the additional inputs to vary according to their prescribed uncertainty distributions.

3. For comparison, an analysis is also performed in which the additional thirteen inputs are treated as calibration parameters, along with the original five.

The first step is to collect a new set of simulator data, which is necessary because the Gaussian process surrogates must now model the relationship between the temperature response and the thirteen new inputs, in addition to the five original calibration inputs. This results in a design of computer experiments over eighteen variables, and surrogates that are based on

nineteen inputs (since time is an input to the surrogates). A random LHS sample of size 50 is used, with the bounds for the original parameters shown in Table 6.12 (for brevity, the information on the thirteen additional parameters is not shown). Generous bounds are used for the calibration parameters, since it is not known how much extra uncertainty will be introduced by the additional uncertain inputs.

Table 6.12: Design of computer experiments for study with additional prescribed input uncertainties (specifications for additional thirteen inputs not listed)

Variable	Lower bound	Upper bound
FPD	2.0×10^{-3}	10.0×10^{-3}
q_2	0	200,000
q_3	0	200,000
q_4	100,000	400,000
q_5	50,000	200,000

With the new code runs, the surrogate models are structured in the same manner as before: two surrogates (for response before and after 500 seconds) are used at each of nine locations on the structure, for a total of eighteen surrogate models. Note that the surrogates capture the temperature response as a function of time, the five original calibration inputs, and the thirteen additional uncertain inputs. The point selection process discussed in Section 3.4 is again employed, and this time between 40 and 128 points are used for each surrogate, depending on the complexity of the response.

Each of the three analyses described above are then conducted. For each case, 50,000 MCMC samples are used to construct the posterior. Note that these analyses are considerably more expensive than those described in Sections 6.4.3 and 6.4.6. Of the three, the most expensive is the third case, in which the new inputs are treated as calibration inputs: the computational cost here is high because the MCMC sampler must evaluate the likelihood ratio (see Eq. (5.14)) once per iteration for each calibration input. Running on a Linux machine with a 64-bit, 2.4GHz processor, the third analysis took approximately 30 hours, while the first two

took on the order of 10 hours each.

Since the calibration parameter FPD is of most interest for the thermal simulation, its marginal posterior distribution is illustrated in Figure 6.47, comparing each of the three analyses listed above. As expected, the posterior distribution for analysis 1 (holding the additional uncertain parameters fixed to their nominal, mean, values) is basically the same posterior that was obtained in the nominal analysis described in Section 6.4.3. The results also indicate that allowing the additional parameters to vary on their prescribed uncertainty bounds leads to a significant increase in the posterior uncertainty for FPD. Finally, the least amount of uncertainty in FPD is obtained when the additional uncertain parameters are treated as calibration parameters, and this is to be expected as well, since that analysis effectively increases the number of degrees of freedom in the calibration.

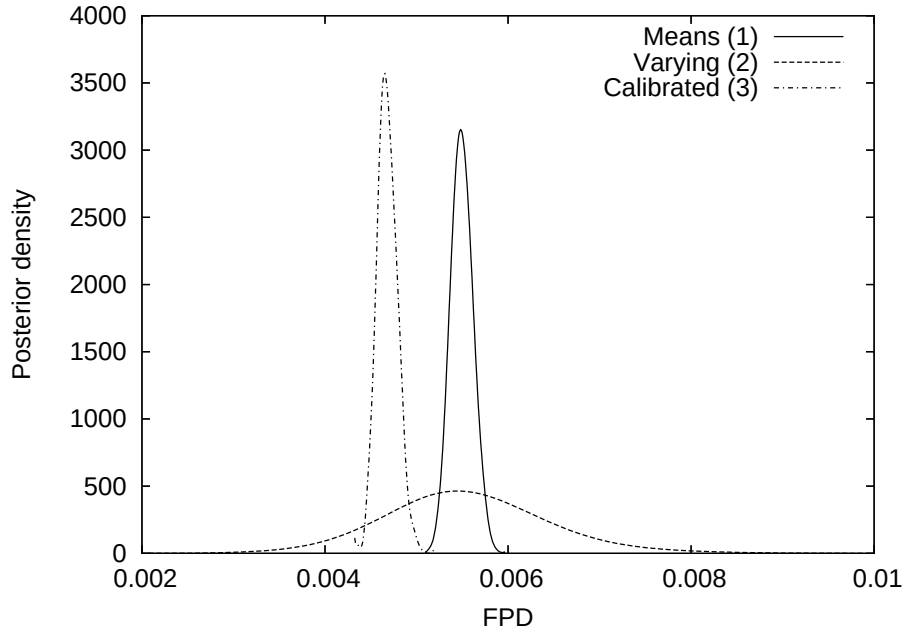


Figure 6.47: Comparison of posterior distribution of FPD for each of three approaches for treating the thirteen additional uncertain model inputs

The preceding analyses were also re-done using 100 LHS samples over the eighteen variables, in order to assess whether or not the results differ significantly from those found using

50 LHS samples. It is determined that increasing the number of simulator runs to 100 does not significantly alter the posterior distribution for this analysis.

6.4.8 Conclusions

This case study has provided the opportunity to illustrate a variety of techniques. First, the point selection algorithm described in Section 3.4 is implemented in order that the surrogate models may capture the response as a function of time, allowing the calibration analysis to consider a full time-history of the observed thermal response. This approach allows for highly compact surrogate model representations (roughly 100 points per GP surrogate, to capture the response as a function of five calibration inputs and time) that are also very accurate (see Figure 6.40).

While the emphasis is on the Bayesian approach to parameter estimation and uncertainty quantification, the classical nonlinear regression approach is compared in Section 6.4.4. While the resulting confidence regions obtained using the nonlinear regression approaches are roughly equivalent to the confidence regions obtained using the Bayesian approach (see Figure 6.41), several differences between the approaches are apparent. One major advantage of using the Bayesian approach with MCMC simulation is that the resulting parameter uncertainty is represented using samples. This makes calculating marginal statistics or displaying confidence region subsets (e.g., a two-dimensional confidence regions subset from a total of five calibration parameters) trivial, as compared to the classical approach. The Bayesian approach also provides a natural construct for enforcing parameter bounds (via the prior information), whereas this can not be done rigorously in classical analysis when computing confidence regions. Finally, the Bayesian approach is also broadly extensible to account for additional uncertainty sources (as discussed in Sections 6.4.6 and 6.4.7).

Another important result that is evident from this study is that the presence of serially correlated errors may have a significant effect on the estimation of the calibration parameters, both with regards to the point estimates and the uncertainty. The existence of serially correlated errors is most likely to occur when the calibration data consist of the same response quantities measured at multiple points in time, as here. While the assumption that the errors are independent it the simplest and most convenient, it does not necessarily take proper account of the amount of information that the calibration data bring to bear on the calibration parameters. The purpose of Section 6.4.5 is to illustrate that the Bayesian calibration framework can easily handle dependent errors.

The incorporation of characterized thermocouple measurement uncertainty (in addition to uncharacterized Gaussian noise) is discussed in Section 6.4.6. This is a powerful addition to the analysis. The known 0–2% bias associated with the external thermocouples is accounted for, and the uncertainty associated with the internal thermocouples is estimated using a finite element model contact parameter study. Thus, the thermocouple uncertainty varies with both time and location on the structure. As expected, the resulting parameter estimation displays a change from that of the original analysis that is consistent with the fact that the external thermocouples provide readings that underestimate the actual temperature (see Figure 6.46).

Finally, the methodology for incorporating prescribed input uncertainties proposed in Section 5.3.2 is illustrated in Section 6.4.7. This is somewhat of an unusual concept, because additional uncertain parameters would normally be treated as additional calibration parameters. However, this particular case study is a good example of when the prescribed uncertainty treatment might be of interest. After considering the "nominal" set of five calibration inputs, the analysts are interested in quantifying the effect of thirteen additional model parameters

having uncertain ranges. In Section 6.4.7, the results are considered both when the additional inputs are treated as calibration inputs and when they are given prescribed uncertainty distributions. As seen in Figure 6.47, the treatment of the additional parameters as calibration parameters results in a small decrease in uncertainty, while their treatment using prescribed uncertainty results in a large increase in overall uncertainty. This information encourages the model developers to consider whether or not the thirteen additional inputs can be viewed conceptually as additional "degrees of freedom." If not, the results of Section 6.4.7 suggest that more effort should be dedicated to reducing the uncertainty in these parameters, so that a more accurate estimation of the parameter of interest (here FPD) might be possible.

6.5 Top-down calibration: bolted joint "three-leg" system

Modeling and simulation projects often consist of code that is hierarchical, such that top-level system simulations are made up of one or more separate pieces of simulation code that describe lower-level physical processes that contribute to the behavior of the system. For example, the simulation may contain various constitutive material models that are themselves used to build one or more subsystem models, which are then put together to form the top-level system model that is ultimately of interest. In such cases, the modeling parameters governing each level of the simulation are typically estimated using experimental data that correspond to that particular level. That is, the parameters governing a material model would typically be estimated using experiments whose purpose is to isolate the behavior of the material in question.

The purpose of this section is to illustrate why this "bottom-up" approach to calibration may not always be optimal in terms of achieving the most predictive system model. The alternative approach considered here is termed a "top-down" approach, because all of the calibration parameters, including the low-level modeling parameters, are estimated using data observed

at the top (system) level. Admittedly, such an approach will be subject to the availability of system-level data.

To illustrate the application of the top-down approach, and to compare its performance to that of the usual approach, a case study based on a system of nonlinear joints is considered. This particular system provides an excellent testbed for illustrating the approach for several reasons. First, this is a multi-level system, in which the bottom-level simulation is the model of an individual bolted joint. Second, there exists a large database of well-controlled, repeated experiments to study the response of both the three-leg system and the individual bolted joints when subject to a variety of excitations.

The bolted joint system is described in Section 6.5.1, and Section 6.5.2 outlines the application of the previous, bottom-up, approach to parameter estimation for this system. In Section 6.5.3, the proposed top-down calibration approach is introduced, and its application to the bolted joint system is presented. Finally, Section 6.5.4 provides a discussion about how the two approaches compare.

6.5.1 Physical system description

The physical system being considered is representative of an aerospace component and consists of a conic shell supported on three legs by a cylindrical shell support structure (three nominally identical sets are shown in Figure 6.48). The conic shell is attached to the support structure via three bolted joints, which play an integral role in the dynamic response of the system.

The bolted joint connections are characterized by an absence of macroslip (relative motion between the upper and lower mating surfaces). However, it is well known that microslip still occurs. Microslip consists of small levels of relative motion between the mating surfaces that occur only over some small portion of the contact surface. The result of microslip is friction,

Figure 6.48: Experimental hardware for the three-leg system

which causes energy dissipation, which in turn tends to damp out the motion of the structure. As a result, accurate modeling of the microslip phenomenon plays an important role in the dynamical modeling of structures containing such connections.

Several energy dissipation models have been proposed for such joints, including the Iwan model (Segalman, 2002) and the Smallwood model (Smallwood, 2000). Both of these joint models are parametrized by a small number of parameters (4 and 3, respectively), which can be estimated based on experimental data. Previous work that has considered parameter estimation and stochastic modeling with the Iwan and Smallwood models includes Urbina et al. (2003b,a, 2004). For this work, however, only the Iwan model is considered.

6.5.2 Previous approach to calibration and prediction: bottom-up

The traditional approach to parameter estimation for multi-level simulation models has been what might be termed a "bottom-up" approach. As mentioned above, this approach involves estimating the unknown modeling parameters that describe each "level" of code using observed data that isolate the behavior corresponding to that particular level. The previous approach for parameter estimation in the three-leg system provides a good illustration.

In this approach, the Iwan parameters are estimated without regard to the behavior of the three-leg system as a whole. As reported by Urbina et al. (2003b), these parameters are estimated using experiments designed to isolate the behavior of the joints themselves. The experimental setup employed for this study is illustrated in Figure 6.49. The mechanical joint was tested by attaching the lower sub-element to a shaker, and attaching the upper sub-element to a 200 pound mass. The 200 pound mass was suspended from springs such that the static force on the joint was approximately zero when no dynamic force was being applied.

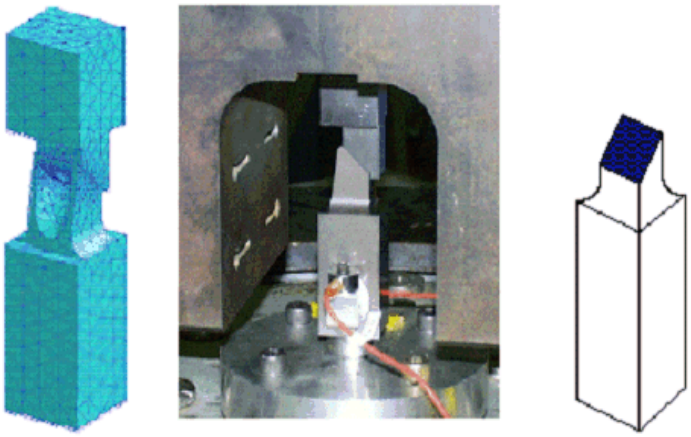


Figure 6.49: Experimental setup for bolted joint tests

The joints were then tested with sine sweep excitations controlled to provide input forces to the lower sub-element equal to 60, 120, 180, 240, and 320 lb. The response accelerations were then measured and used to infer the energy dissipated per cycle in the system, at resonance. This procedure was repeated for a total of 12 experiments, in which the system was disassembled and reassembled between each experiment. Each of these energy dissipation curves (corresponding to 5 points: one for each force level) can then be compared against the curve predicted by the Iwan model, for a particular combination of the Iwan parameters.

The parameter estimation approach taken by Urbina et al. (2003b) was to find the set of Iwan parameters that give the best fit to each individual data point (in a least-squares sense).

This analysis resulted in 12 realizations of the Iwan parameters, one corresponding to each experimental observation. The means and standard deviations of these parameter estimates are listed in Table 6.13, and the observed correlation coefficients are listed in Table 6.14.

Table 6.13: Means and standard deviations of Iwan parameters identified by Urbina et al. (2003b)

Parameter	Mean	Std. Dev.
R	3.35×10^6	1.07×10^6
S	1.58×10^6	1.40×10^5
χ	-0.538	0.0338
ϕ_{max}	3.27×10^{-4}	8.34×10^{-5}

Table 6.14: Observed correlations for Iwan parameters identified by Urbina et al. (2003b)

	R	S	χ	ϕ_{max}
R	1	0.29	0.65	-0.22
S	0.29	1	-0.15	-0.90
χ	0.65	-0.15	1	0.09
ϕ_{max}	-0.22	-0.90	0.09	1

Once the parameters of the Iwan joint model are identified, the Iwan joint model may be included as a part of the system model of the three-leg system. Urbina et al. (2005) provide a validation assessment of the three-leg system, using the above identified Iwan parameters. The simulation model used for the system is a simple lumped-mass representation, a schematic of which is given in Figure 6.50. The "attachment" and "correction" stiffnesses (K_{corr} and $K_{attachment}$ in Figure 6.50) must also be estimated, in what might be termed a second calibration phase. The attachment stiffness was calibrated to match the axial frequency of the structure. The correction stiffnesses were treated, as their name suggests, as correction factors, and were simply used in an ad hoc manner to adjust the output of the three-leg model to agree more closely with the observations.

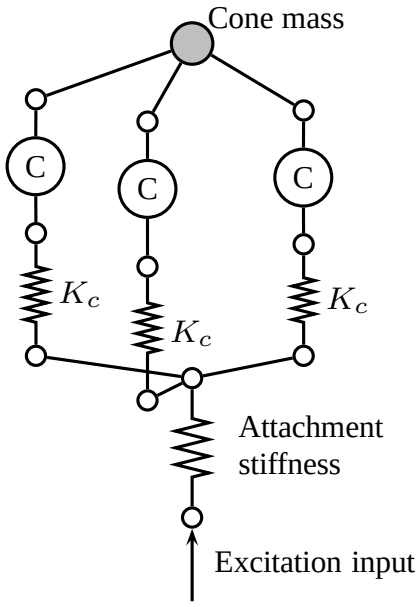


Figure 6.50: Schematic of "lumped-mass" model of the three-leg system (each C represents an Iwan model of a bolted joint connection)

6.5.3 Proposed "top-down" calibration approach

The proposed approach differs from the previous approaches in the process that is used to estimate the parameters governing the bolted joints. As discussed in Section 6.5.2, previous approaches have used only single-joint experiments to estimate the parameters governing the Iwan model for the bolted joints. The present hypothesis, however, is that a more predictive *system* model might be obtained by using the observed response of the three-leg system itself to estimate the governing parameters of the Iwan joint model. A pre-requisite for this "top-down" approach is then of course that experimental data are available at the system level (which is the case for the three-leg system). The philosophy behind the top-down approach is that by considering the component-level parameters as degrees of freedom (as opposed to known values) during the system-level calibration analysis, a more predictive system-level model may be obtained.

An additional difference is that the top-down approach is a completely "black-box" method.

That is, knowledge about the physical modeling issues is neither required nor applicable. In contrast, the traditional approaches to the calibration of the Iwan and Smallwood models are based on the physical meanings of the unknown parameters. For the top-down approach, the system-level model is simply viewed as a black-box function that takes a vector of input parameters (among which are those parameters associated with the component-level models, i.e. the Iwan parameters) and produces an output. If the corresponding output has been observed experimentally, the problem of parameter estimation (including component-level parameters) can be cast in a straightforward manner as an inverse problem.

The top-down calibration will be conducted by considering the response of the three-leg system when it is subjected to a "wavelet" excitation (as shown in Figure 6.51). The response quantity under consideration is the acceleration time-history of the conic shell. Various possibilities exist for calibration based on time-series output. One option is to discretize the response and compare the predicted and observed output at a set of discrete time instances. While this might provide the most comprehensive representation of the output, point-wise comparisons are generally not appropriate for oscillatory, dynamic responses. The problem is that small phase shifts between the two time histories will result in drastic inconsistencies for a point-wise comparison, even when the important characteristics of the two responses match well.

A simpler, more appropriate approach is to reduce the output to one or more scalar "features" of interest, such as the peak value. This work considers a feature set that attempts to capture the behavior of the response with respect to acceleration decay and natural frequency. The acceleration decay features are chosen to emphasize the importance of the model's ability to predict the maximum acceleration, as well as the amount of decay in the response. For simplicity, two acceleration features are used: the maximum acceleration (generally the height of

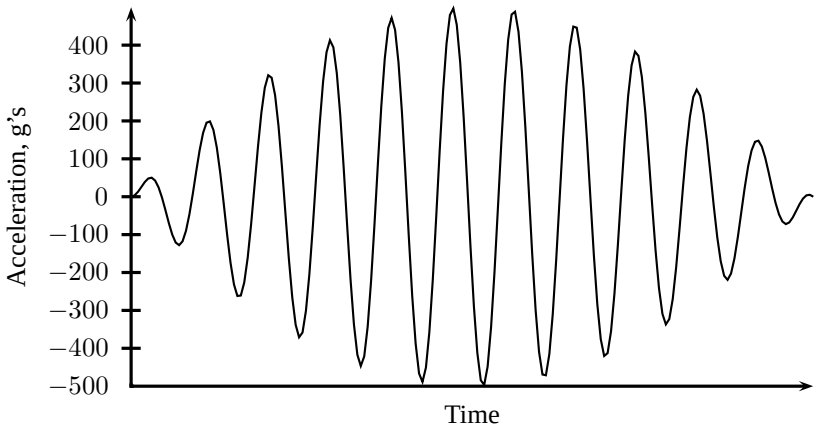


Figure 6.51: Wavelet input excitation waveform

the first peak), and the final recorded acceleration peak (which represents the amount of decay in the response). An example acceleration time history associated with the response of the three-leg system is shown in Figure 6.52. In addition to the two acceleration features, a third feature is included in the calibration analysis, which is the value of the first natural frequency of the response.

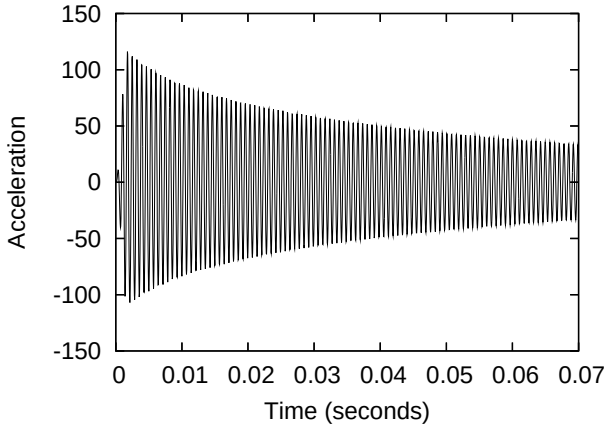


Figure 6.52: Example of acceleration time history associated with three-leg system subject to the wavelet excitation

For the top-down calibration approach, the Bayesian calibration methodology described in Section 5.3 is implemented. While the Bayesian approach provides a comprehensive representation of the uncertainty in the resulting parameter estimates, it is not a requirement of the top-down approach, and in fact any appropriate parameter estimation technique might be used

(such as nonlinear regression analysis). The unknown parameters to be estimated are the four parameters governing the Iwan joint, R, S, χ, and ϕ_{max}.

Recall that the three "correction" stiffness terms were also added to the model of the three-leg system. The effect of these terms is to soften the stiffness of the structure, and they were originally included in the model simply as correction factors. For this work, two cases are considered:

1. The correction stiffness terms are included in the model of the three-leg system, as with the previous work, but they are treated as three additional unknown calibration parameters, along with the Iwan parameters, so that there are a total of 7 calibration parameters.

2. The correction stiffnesses are not used in the model, which is equivalent to setting their stiffness values to infinity. In this case, the 4 Iwan parameters are the only calibration parameters.

The Bayesian approach requires the specification of a prior distribution for the unknowns. As discussed in Section 2.2, bounded uniform distributions are often desirable as vague priors because they capture the notion that any value of the parameters within the prescribed bounds is equally likely. When the parameters do not have physical limits, generous bounds may be specified to allow for a comprehensive exploration of the parameter space. Although there may be little guidance regarding the choice of appropriate bounds, the method does provide feedback, and the resulting marginal posterior distributions will show whether or not the bounds should be expanded or contracted. (As will be discussed shortly, the bounds will also be used to define the design of computer experiments, and this is why the prior distribution can not be given support over the entire real line.)

The bounds listed in Table 6.15 for the Iwan parameters are chosen to allow the posterior

distribution to deviate significantly from the previous estimates discussed in Section 6.5.2. The parameter χ is physically constrained to be in the range of $(-1, 0)$, while the bounds for the remaining three Iwan parameters are chosen subjectively. Urbina et al. (2005) used a nominal value of the correction stiffnesses of 7.1×10^6 lb/in, so its bounds are constructed at approximately $+/-$ 40% of this nominal value, as listed in Table 6.15.

Table 6.15: Prior bounds for the parameters in the top-down calibration

Parameter	Lower bound	Upper bound
$\log R$	13.8	20
S	1×10^5	1×10^7
χ	-1	0
ϕ_{max}	5×10^{-5}	1×10^{-3}
K_{c1}, K_{c2}, K_{c3}	4×10^6	1×10^7

The experimental data are 27 independent, repeated measurements of the acceleration response of the conic shell piece in the three-leg system, when the "wavelet" excitation is applied. To obtain 27 repeated observations, each of three nominally identical conic shells is paired with each of three nominally identical base plates (see Figure 6.48), for a total of nine combinations. For each combination, the experiment is repeated three times, disassembling and reassembling the system between each experiment. An illustration of the experimental setup is given in Figure 6.53.

Gaussian process surrogate models are used in place of the actual three-leg simulation for the Bayesian calibration analysis. The initial training data for the surrogates is based on 200 Latin hypercube samples of the calibration parameters, in accordance with the bounds listed in 6.15. The simulation runs have to be computed separately with and without the correction stiffnesses. For the case in which the correction stiffnesses are considered as unknowns, the 200 samples of the Iwan parameters are augmented with 200 independent samples of K_{c1}, K_{c2}, and K_{c3}.

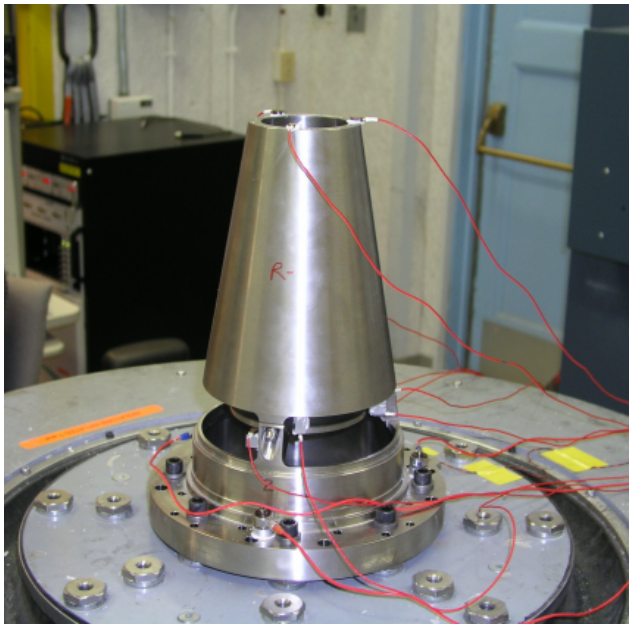


Figure 6.53: Experimental setup for the three-leg system

The three-leg simulation is configured for the same wavelet excitation waveform as the experiments and is exercised for each parameter combination (200 runs are made with the correction stiffnesses, and 200 runs are made without them). For the model that includes the correction stiffnesses, five out of the 200 parameter combinations cause the simulation to fail, and it is determined that the five failed parameter combinations all have values of χ very near -1, which is a physical lower bound. The corresponding surrogates for this case are thus built using the 195 training points that do not fail the simulator. None of the simulations fail when the model is exercised without the correction stiffness terms.

Instead of creating a surrogate to the time-history acceleration response itself, three independent surrogates are created for the three response features (maximum acceleration, final acceleration peak, and frequency). Each Gaussian process surrogate uses a constant trend, and the GP parameters are estimated using maximum likelihood, as discussed in Section 3.3.

The Bayesian calibration formulation is as given by Eq. (5.11), where in this case the scenario variable $\boldsymbol{s}$ is simply an indicator for each of the three response features. Consider that

the experimental data vector is partitioned as $\boldsymbol{d} = (\boldsymbol{d}_1^T, \boldsymbol{d}_2^T, \boldsymbol{d}_3^T)^T$, where each $\boldsymbol{d}_i$ contains the 27 observations of the ith response feature. The 81×81 data covariance matrix is partitioned as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 \boldsymbol{I} + \sigma_{GP1}^2 \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \boldsymbol{I} + \sigma_{GP2}^2 \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_3^2 \boldsymbol{I} + \sigma_{GP3}^2 \mathbf{1} \end{bmatrix}, \tag{6.37}$$

where $\boldsymbol{I}$ is the 27×27 identity matrix and $\mathbf{1}$ is the 27×27 matrix of all ones. The variance of ε_i in Eq. (5.11) is one of σ_1^2, σ_2^2, or σ_3^2, depending on which response features observation i corresponds to. Similarly, σ_{GP1}^2, σ_{GP2}^2, and σ_{GP3}^2 are the surrogate model variances from the surrogate model corresponding to each response feature. Each $\sigma^2 \boldsymbol{I}$ term indicates that the experimental observations of the same response feature are independent of each other. Each $\sigma_{GP}^2 \mathbf{1}$ indicates that the surrogate predictions for a given response feature are perfectly correlated, with variance (surrogate uncertainty) σ_{GP}^2. The off-diagonal $\mathbf{0}$ blocks in Eq. (6.37) indicate the assumption that the response features are not correlated with each other. This is true for the surrogate predictions (because the three GP surrogates are independent), but it is not necessarily true for the experimental data. However, the three features for this analysis are approximately orthogonal, and the largest sample correlation coefficient for the experimental data is only 0.37, between the maximum acceleration and the first natural frequency. It thus seems appropriate to treat the response features as independent for this analysis.

The three error variances, σ_1^2, σ_2^2, and σ_3^2, are treated as additional objects of Bayesian inference, as opposed to known constants. They are each independently given the standard reference prior distribution (Lee, 2004) $\pi(\sigma^2) \propto 1/\sigma^2$, which yields

$$\pi(\sigma_1^2, \sigma_2^2, \sigma_3^2) \propto \frac{1}{\sigma_1^2 \sigma_2^2 \sigma_3^2}. \tag{6.38}$$

Not only does this approach naturally accommodate response features with incompatible units, but it takes account of the fact that the variance of the difference between the predictions and observations is not known *a priori*.

The posterior distribution for each case (with and without the correction stiffnesses) is constructed using the component-wise version of the Metropolis sampling algorithm. The acceptance rate for each parameter is adjusted to an optimal value of about 0.25.

The resulting marginal posterior distributions for the Iwan parameters are shown below in Figures 6.54 through 6.57. Each figure compares the marginal posteriors obtained with and without the correction stiffnesses included in the three-leg simulation. In each figure, the range used for the x-axis is the same as the bounds used for that parameter's prior distribution. For the analysis that includes the correction stiffnesses as calibration parameters, their marginal posteriors do not show much of a change from the priors.

A variety of dependencies and correlations are found within the resulting posterior distributions. Perhaps the most significant is a nearly linear relationship between $\log(R)$ and χ. This relationship is illustrated in Figure 6.58, which shows 95% simultaneous confidence regions for these two parameters, with and without the correction stiffnesses. Note that Urbina et al. (2003b) also found a positive relationship between R and χ.

6.5.4 Validation assessment

The validity of the previously obtained parameter estimates is now addressed. To do so, the simulation is exercised using a new type of excitation, specifically a "blast" excitation (Figure 6.59), as opposed to the "wavelet" excitation (Figure 6.51) which was used for the calibration experiments. The response of the three-leg system to the "blast" excitation was also observed experimentally, and as with the calibration data, 27 measurements are available. In

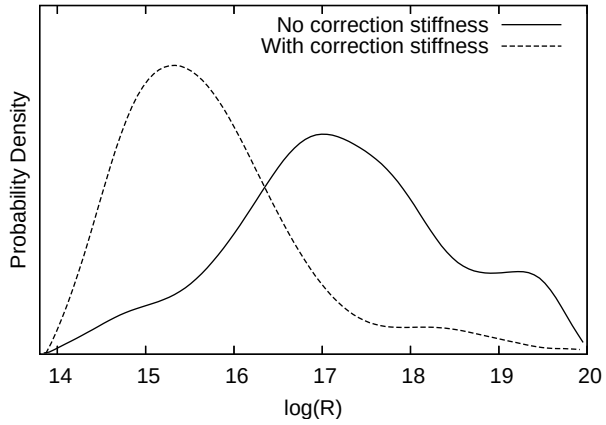


Figure 6.54: Marginal posterior distributions of $\log R$

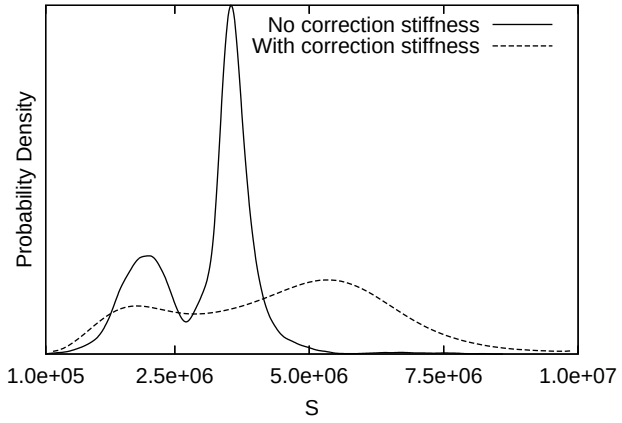


Figure 6.55: Marginal posterior distributions of S

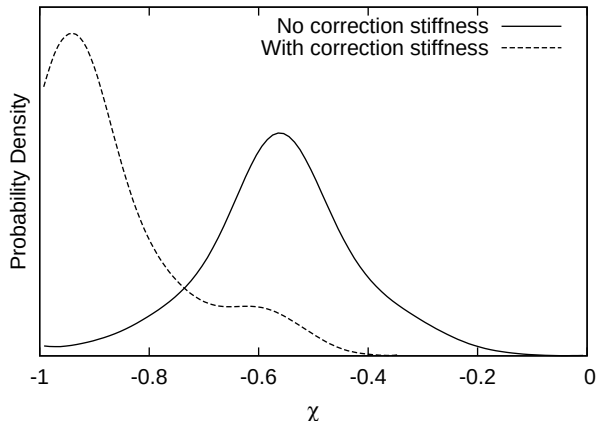


Figure 6.56: Marginal posterior distributions of χ

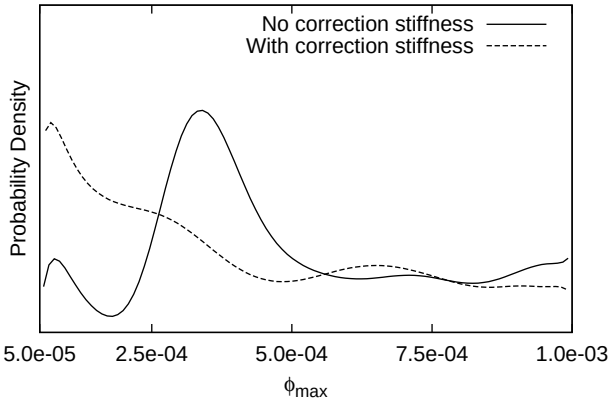


Figure 6.57: Marginal posterior distributions of ϕ_{max}

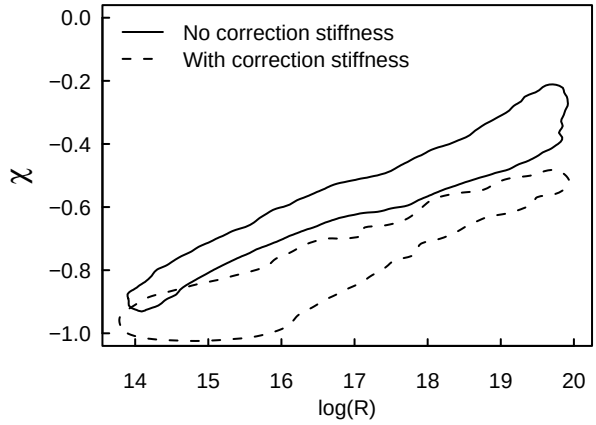


Figure 6.58: 95% posterior confidence regions for $\log(R)$ and χ. The plotting bounds represent the prior bounds.

addition, comparison against previous model prediction results is possible, specifically comparison with 20 model predictions obtained using realizations of the Iwan parameter discussed in Section 6.5.2.

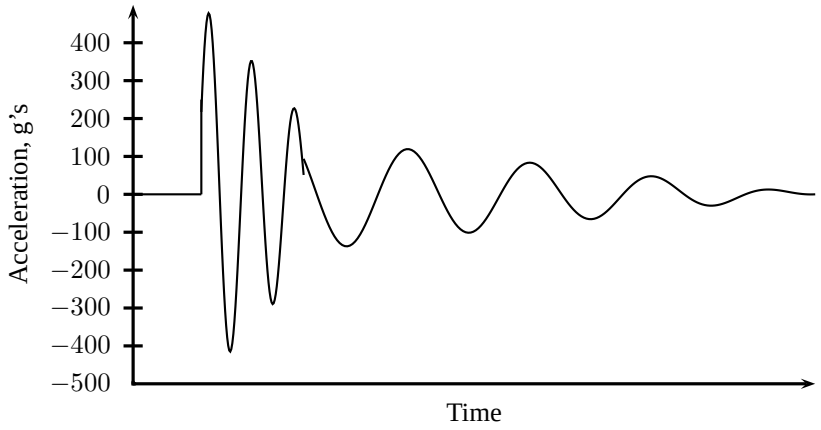


Figure 6.59: Blast input excitation waveform

An example of the experimentally observed acceleration response of the three-leg system to the blast excitation is given in Figure 6.60.

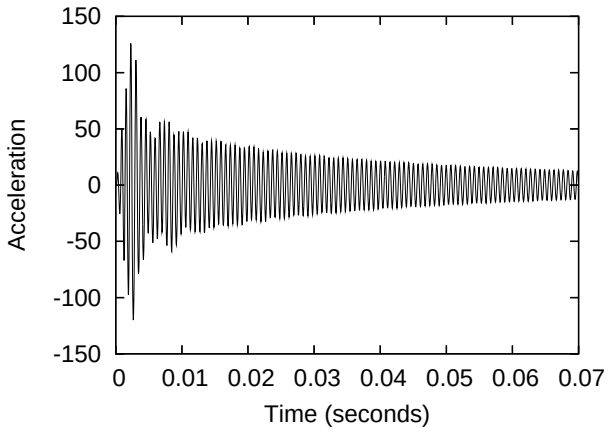


Figure 6.60: Example of experimentally observed acceleration response of the three-leg system subject to the blast excitation

The validation assessment will also be based on response features. The validity of the predictions are considered with respect to the maximum absolute acceleration and the final recorded acceleration peak (up to 0.074 seconds). The maximum absolute acceleration is con-

sidered here as opposed to maximum positive acceleration because the asymmetry in the response is slightly more pronounced for the blast excitation than for the wavelet (recall that for the top-down calibration based on the wavelet excitation, maximum positive acceleration was considered).

The validation assessment will be made by considering simultaneous confidence regions for the mean values of these two features. For the 27 experimentally observed responses and the 20 model predictions based on realizations of the original Iwan estimates, each data set is considered to contain samples from a bivariate normal population, and the simultaneous confidence regions for the population means are based on the inequality (Srivastava, 2002)

$$\frac{(f-p+1)n}{fp}\bar{\boldsymbol{x}}^T\boldsymbol{S}^{-1}\bar{\boldsymbol{x}} \geq F_{p,f-p+1,\alpha}, \tag{6.39}$$

where $f = n-1$, n is the number of samples, $p = 2$, $\bar{\boldsymbol{x}}$ is the sample mean of the data, $\boldsymbol{S}$ is the sample covariance, and $F_{p,f-p+1,\alpha}$ is the upper $100\alpha\%$ point of the F-distribution with $(p, f-p+1)$ degrees of freedom. It is acknowledged that both the 27 experimental samples and the 20 originally calibrated samples show strong non-normality, but the confidence region of (6.39) is still applied, under the assumption that the central limit theorem holds approximately.

The confidence region for the means based on the Bayesian top-down calibration results must be computed differently, because the Bayesian approach does not yield samples of a population of model responses representing inherent variability. Instead, when the joint posterior distribution of the calibration inputs is propagated through the model for blast excitation, one directly obtains confidence regions for the true mean value of the response. This confidence region is estimated here by choosing 100 random samples from the joint posterior distribution of the calibration inputs (for both cases: with and without the correction stiffnesses) and us-

ing these samples as inputs to the three-leg simulation with blast excitation. In doing so, 10 simulator runs fail for the model with correction stiffnesses, but the remaining 90 realizations are used to construct the posterior distribution of the response for that case. Recall that this posterior distribution is not a representation of response variability, but is instead a representation of the posterior uncertainty in the "true" value of the response. As such, this posterior distribution itself (based on the 90 samples in the first case and 100 samples in the second) is used (via multivariate kernel density estimation) to construct a 95% confidence region for the mean response based on the Bayesian top-down calibration results.

The resulting confidence regions are plotted in Figure 6.61. The Bayesian top-down calibration results are plotted both with and without the inclusion of the corrections stiffnesses.

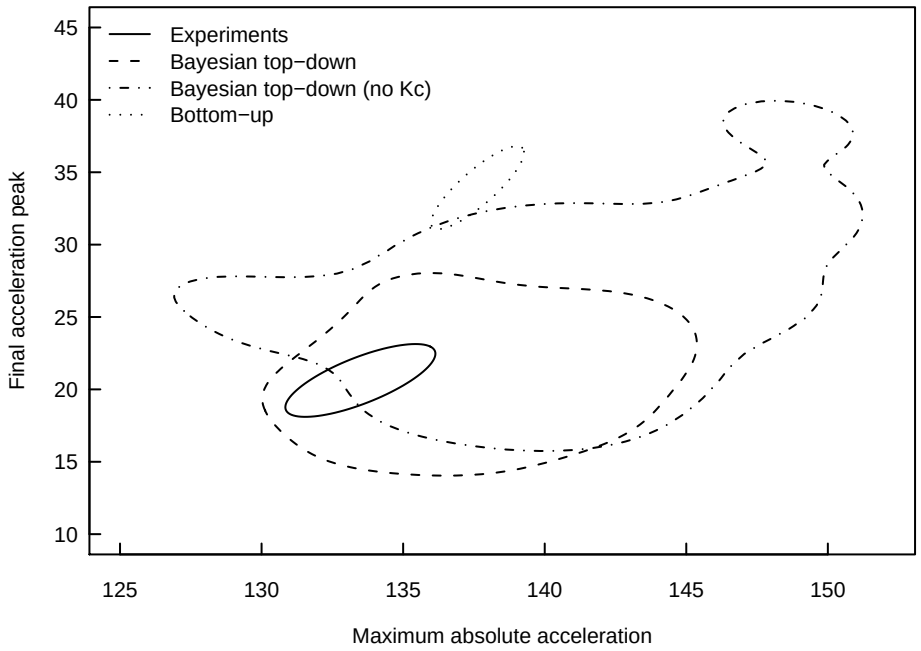


Figure 6.61: 95% confidence regions for means of response features for three-leg system with blast excitation

There are several noteworthy features of the results shown in Figure 6.61. First, one notices that the Bayesian confidence regions are much larger than the confidence region based on the original calibration approach (as well as that of the experiments). This is because the Bayesian

approach has taken a broad account of various uncertainties, including the surrogate model uncertainty, as well as the uncertainty in the response feature variability (in so much as σ_1, σ_2, and σ_3 are each given vague prior distributions and treated as objects of Bayesian inference). Along these lines, one also notices that the Bayesian confidence regions either intersect or enclose the confidence region for the experiments, which is a result of the Bayesian approach's comprehensive uncertainty representation.

In any case, the primary objective of this section is to discuss the validity of the top-down calibration approach, not to tout the uncertainty quantification capabilities of Bayesian inference. While there is a relatively large amount of uncertainty in the top-down calibration results, the validation results indicate that overall, the predictions calibrated via the top-down approach provide a closer agreement to the experiments for the blast excitation. In particular, the original calibration results provide predictions that significantly over-estimate the experimentally observed minimum acceleration, whereas the the top-down results agree more closely with the experiments with regards to this feature. Overall, the top-down calibration that included the correction stiffnesses provides predictions that agree very closely to the experiments. The top-down calibration that did not included the correction stiffnesses does not agree quite as well in a means sense, but the experiments are still not outside the predicted uncertainty band.

Despite its title, the purpose of this section is not to make an absolute statement about the validity of the three-leg model, only to illustrate the potential of the top-down calibration approach. As such, the confidence regions presented in Figure 6.61 are not intended to show "validity" or non-validity of any particular model. While the confidence regions indicate that the experiments are not outside of the uncertainty ranges predicted by the models calibrated with the top-down Bayesian approach, it may be more telling to compare the top-down approach to

the original results. It has already been stated that the top-down results compare better with the experiments than the original results, but this comparison can also be quantified.

Consider here the Mahalanobis squared distance between to vectors $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$, which is given by

$$\Lambda^2 = (\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})^T \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}). \tag{6.40}$$

The Mahalanobis distance takes account of the covariance structure associated with the population under consideration, in effect penalizing deviations that are inconsistent with the principal directions of variation. This makes the Mahalanobis distance a good scalar metric for assessing the "closeness" to the experiments of the various predictions.

For this assessment, the covariance matrix will be estimated using the experimental data, and this same covariance matrix will be used for each calculation, so that a fair comparison can be made. In each case, the distances are between the sample means of the experiments and the predictions, so that the locations, not the uncertainties, are being compared. The resulting distances between the experiments and the predictions are listed in Table 6.16.

Table 6.16: Mahalanobis squared distances between the means of the experiments and predictions (based on two response features) for each calibrated model

Approach	Λ^2
Original	10.39
Top-down	0.848
Top-down (no K_c)	1.39

The Mahalanobis distance computations provide clear evidence that the model predictions obtained using the top-down calibration approach agree more closely to the experimental data than those predictions obtained using the original calibration approach.

6.5.5 Conclusions

The purpose of this section is to illustrate an alternative to the usual "bottom-up" approach to the calibration of multi-level simulation codes. The usual approach to calibration involves multiple stages of parameter estimation, in which the modeling parameters for each "piece" of the simulation are estimated using experimental observations whose relevance is limited to that piece of the simulation. The alternative approach presented here, which is referred to as a "top-down" calibration approach, is a parameter estimation approach in which the modeling parameters at all of the simulation levels are estimated using only those experimental data at the highest level (most likely, the system level).

One of the reasons that the bottom-up approach to calibration is so prevalent is that it is an intuitive approach. It makes sense from a modeling perspective that each piece of the code should be able to operate independently, and that if the relevant features of the physical behavior are captured correctly at each modeling level, then the high-level model predictions when the different pieces are put together should be trustworthy. While this approach is intuitively attractive from a physical modeling perspective, it will not necessarily yield the most predictive system-level model predictions. All models contain approximations and errors, and the bottom-up parameter estimation approach may render the high-level model particularly susceptible to the errors and approximations pertaining to the interfaces and interactions among the various model components.

The primary limitation of the top-down approach is that it requires that the experimental measurements of the system-level response are available, and this is admittedly often not the case. However, when such measurements are available, they effectively provide the ultimate benchmark for the simulation model as a whole, and as such, limiting the use of the system-

level observations to validation exercises is in essence ignoring an opportunity to improve the predictive capability of the entire simulation model.

The three-leg system case study presented in this section is an example in which the system-level observations are used to infer *all* of the unknown modeling parameters (those at both the component and system levels). This is somewhat of an extreme example of the top-down calibration philosophy, but it is used here to make the point that the top-down approach has the potential to result in a more predictive system-level model (refer to the results of Section 6.5.4). In less extreme cases, the parameter estimation process may involve both lower and higher-level data, and there is no doubt that future work could develop a methodology to accommodate such an approach. One cautionary note, though, would be that when parameter estimates based on lower-level data versus those based on higher-level data show strong discrepancies, it is likely that an attempt to accommodate all pieces of data in a unified approach will result in a less-predictive system-level model than might be obtained by "yielding" to the parameter estimates suggested by the system-level data. In any case, I encourage modelers to consider whether or not higher-level experimental observations might provide relevant information for the estimation of lower-level modeling parameters.

CHAPTER VII

CONCLUSION

7.1 Summary

Given the importance that is placed on modeling and simulation in engineering and the sciences, it is essential that analysts make efforts to consider the uncertainty and validity of their models. While increasing interest in model assessment has given rise to the field of *Verification and Validation*, there has also been recent work to study how experimental observations may be used for model calibration. In fact, calibration provides not only an opportunity to improve the predictive capability of a model by aligning model output more closely with observed response values, but as emphasized throughout this dissertation, calibration is also an opportunity to quantify contributors to the uncertainty in model predictions. Through the use of a variety of case studies, this dissertation aims to illustrate, enhance, and develop methods that support the quantification of uncertainty in the modeling and simulation process.

One of the primary goals of this dissertation is to emphasize the power of both Bayesian inference and Gaussian process interpolation as tools that can greatly enhance the uncertainty quantification process. Bayesian analysis allows one to develop rigorous, mathematical representations of the uncertainty present in parameters that are estimated using observed data. Such a capability is fundamental to the uncertainty quantification process, and Bayesian inference is applied in four out of the five case studies of Chapter VI. Similarly, as a surrogate modeling technique, Gaussian process interpolation is indispensable because the technique enables the construction of accurate, inexpensive approximations to the functional relationship between

simulator inputs and outputs (consider that the technique is employed in all of the five case studies of Chapter VI). Gaussian process interpolation is especially valuable in the uncertainty quantification arena because the technique allows one to directly quantify the uncertainty introduced by the use of the surrogate model (in fact, this uncertainty can be accounted for in a calibration analysis using Bayesian inference, as outlined in Section 5.3.1).

While model validation has been recognized for some time as an important phase of the modeling and simulation process, there has been little success to establish broadly applicable tools, or *validation metrics*, for constructing quantitative statements about model validity in light of observed data. While significant previous efforts have been devoted to developing and comparing various quantitative metrics (see, for example, Rebba, 2005), this dissertation takes a different approach to advancing the state of knowledge with respect to model assessment. Instead of focusing on individual metrics, the contention here is that analysts should focus on finding an approach that is appropriate to the particular model validation scenario. This work also emphasizes the importance of interpreting quantitative results in an a manner that is meaningful to the ultimate model validation objective, which is to determine whether or not a given model is suitable for its intended use.

The five case studies presented in Chapter VI are intended to illustrate the point that there may be a variety of constructive ways of thinking about quantitative validation assessment, and that the choice of which method to apply is very much situation-dependent. For example, statistical significance testing is employed in Section 6.1, where it is illustrated that the strength of the model assessment conclusion is not particularly strong because of the high variability in the experimental response and the small number of repeated experimental observations. On the other hand, two different error characterization approaches are used for the structural dy-

namics problem of Section 6.2, where the difference between fully- and partially-characterized experiments is discussed. Validation assessment can also be an extension of the parameter estimation or model calibration process, as illustrated via the cross-validation exercises illustrated in Section 6.3. Finally, multivariate distance measures and confidence region plots are used for model assessment in the case study of Section 6.5.

While model assessment is certainly an important research area, this dissertation focuses more on model calibration, and in particular how the calibration process can be viewed as an additional opportunity to quantify contributors to the uncertainty associated with model predictions. Traditionally, calibration has been seen as a process whose primary objective is to improve the predictive capability of a model, by aligning its output more closely with observed data. However, since the seminal work of Kennedy and O'Hagan (2001), there has been an increasing interest in calibration approaches that support the quantification of modeling uncertainty. And as elucidated by Kennedy and O'Hagan, Bayesian inference is particularly well-suited for this endeavor.

One of the objectives of this dissertation is to illustrate the Bayesian framework for model calibration for a variety of applications, and this is done in Sections 6.1, 6.3, 6.4, and 6.5. The power of the Bayesian approach for uncertainty quantification is clearly evident. For example, in Section 6.1.4, the uncertainty associated with the calibration parameter estimates is employed via uncertainty propagation to compute an *uncertainty distribution* for the failure probability of a system of devices (see Figures 6.6 and 6.7). The Bayesian framework is also readily extensible, and two extensions to account for uncertainty sources beyond those treated in the traditional framework are proposed in Sections 5.3.2 and 5.3.3 and illustrated in Sections 6.4.6 and 6.4.7.

The formulation proposed by Kennedy and O'Hagan (2001) is also considered here with respect to what it adds to the analysis beyond the traditional Bayesian framework. In particular, Kennedy and O'Hagan's framework incorporates a model inadequacy function that captures the simulator bias as a function of independent variables, such as boundary conditions or geometry. In fact, this approach is compared to the more traditional formulation in Sections 6.1.3 and 6.1.4. It turns out that the inclusion of the model inadequacy function (modeled as a Gaussian process over the independent variables) adds a significant amount of complexity to an already demanding analysis. Not only this, but the Gaussian process covariance parameters governing this function can be especially difficult to estimate (unless there is a wealth of experimental data, as in the example problem presented by Kennedy and O'Hagan, 2001), which difficulty may lead to inaccurate uncertainty estimation. And as discussed in Section 5.4, the inclusion of a scenario-dependent model inadequacy function is not even appropriate in many cases (in fact, of the five case studies presented in Chapter VI, the incorporation of the model inadequacy function is only appropriate in one).

The Bayesian framework is not the only means of quantifying uncertainty in the calibration process, and in fact the more traditional approach known as nonlinear regression is discussed in Section 5.2. However, it is found that the development of nonlinear regression confidence regions is not nearly as flexible as the uncertainty representations used in Bayesian inference. For example the Bayesian approach allows one to explicitly incorporate parameter constraints or bounds (via the prior distribution). Even more, the nonlinear regression confidence regions can be in some cases (perhaps counterintuitively) even more expensive to compute and display than the Bayesian counterparts (see nonlinear regression confidence region subsets, as in Eq. (5.10)), and highly nonlinear models can cause problems for classical, asymptotic ap-

proaches to inference (Seber and Wild, 2003).

Perhaps most importantly, the propagation of parameter uncertainties that are characterized via Bayesian posterior distributions is much more straightforward than with classical analysis. This is because the Bayesian parameter uncertainty is represented using probability distributions, and if the posterior distribution is constructed using MCMC sampling, then the samples needed for uncertainty propagation are obtained through the calibration process itself. In fact, once these samples are obtained, the calculation of statistics, the display of complete or marginal confidence regions, and marginalization over nuisance variables are all trivial tasks. Even in the midst of a text focused on classical approaches to parameter inference, Seber and Wild (2003, Sec. 5.11) acknowledge that "with the development of efficient methods for finding posterior density functions..., it is clear that Bayesian methods of inference based on the joint posterior density of $\boldsymbol{\theta}$ and its marginal densities can have considerable advantages."

7.2 Recommendations for future work

Model validation, uncertainty quantification, and calibration are all active research areas, and many excellent opportunities exist for researchers to make additional contributions. Fundamental work to develop new or improve existing response surface approximation methods will be extremely valuable. While the Gaussian process interpolation method considered in this work is quite powerful, there are certainly opportunities to extend its capabilities. In particular, the development of practical approaches for constructing non-stationary models could allow for the creation of a more broadly applicable and more accurate class of surrogate models. In addition, the problem with redundant data and ill-conditioned correlation matrices, while addressed in this dissertation through the development of an iterative point selection algorithm, might still warrant additional consideration.

With respect to model validation, future work should focus on the development and implementation of approaches that extract the most information possible from available validation data and use this information to draw meaningful conclusions about a model's suitability for its intended use. Because validation data are so often limited, it is important to be able to make the most of the available data. Additional future work that considers the optimal design of validation experiments would also be a significant contribution to this field.

Another interesting aspect of model validation is that validation data are rarely 100% relevant to the model's intended use; as such, there is usually some degree of extrapolation that exists when one makes use of available data to draw conclusions about the utility of a particular model. For example, if validation data show that a model's accuracy is acceptable for predicting system response in laboratory conditions, what conclusions can be made about the confidence in the model's accuracy for more extreme conditions that are representative of its ultimate application? Even further, what aspects of this question are better addressed by domain experts, and what aspects can be addressed by validation analysts? It may be the case that the model developers can use physical arguments to show that if a model is validated in the laboratory conditions, then its performance will remain satisfactory in the application domain because all of the relevant physics are captured in the validation experiments. Such a conclusion might be accessible only through the use of physical arguments, in which case it could not be reached via a strictly statistical analysis of the validation data.

Perhaps progress in model validation might be best achieved by developing approaches that better involve the physical modelers and model users in the validation process; in the end, they are probably the ones who can make the most sense out of the validation results. Validation analysts should strive to communicate the steps of the validation process, as well as the statis-

tical tools and metrics, in such a way that the model builders and users are fully apprised of the statistical information that can be gleaned from the validation data, allowing them to make the most informed decision possible about the capabilities of the model in question.

Additional efforts might also be made to consider more closely the relationship between validation and calibration. A commonly asked question is, "If I validate my model, should I still do calibration?" Similarly, the implication of a successful calibration analysis on model validity may also be of interest. Towards these ends, researchers might think more carefully about *cross-validation*, and how "different" the validation domain must be from the calibration domain in order that substantive conclusions about validity might be drawn.

As a new and promising field, the Bayesian calibration of computer simulations is an area where plenty of opportunities exist to make significant research contributions. One particular area of interest might be the study of formulations that allow one to correct for systematic model bias. While Kennedy and O'Hagan (2001) include such a consideration in their work, their use of a Gaussian process for the model bias term results in computational difficulties and a potentially inaccurate uncertainty representation, as discussed in Sections 5.4 and 6.1. A simpler formulation, where the simulator bias is modeled as a linear function of the independent variables, say, might very well be worth considering.

Finally, the central theme of this work is uncertainty quantification. It seems that there will always be room for additional work to develop more efficient methods for uncertainty propagation or more accurate approaches to model uncertainty and variability. The broad applicability and extensibility of the Bayesian method, as illustrated herein, suggests that with a little effort and creativity, the tools of Bayesian inference might be brought to bear on modeling and simulation in a manner far more profound and insightful than the ideas presented here.

REFERENCES

AIAA. Guide for the verification and validation of computational fluid dynamics simulations. Technical Report AIAA-G-077-1998, American Institute for Aeronautics and Astronautics, Reston, VA, 1998.

ANS. Guidelines for the verification and validation of scientific and engineering computer programs for the nuclear industry. Technical Report ANSI/ANS-10.4-1987, American Nuclear Society, 1987.

R. Aslett, R. J. Buck, S. G. Duvall, J. Sacks, and W. J. Welch. Circuit optimization via sequential computer experiments: design of an output buffer. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(1):31–48, 1998.

ASME. Guide for verification and validation in computational solid mechanics. Technical Report V&V 10-2006, American Society of Mechanical Engineers, New York, NY, 2006.

I. Babuska, F. Nobile, and R. Tempone. Formulation of the static frame problem. *Computer Methods in Applied Mechanics and Engineering*, 2008. In press, available online.

S. Balakrishnan, A. Roy, M. G. Ierapetritou, G. P. Flach, and P. G. Georgopoulos. Uncertainty reduction and characterization for complex environmental fate and transport models: an empirical Bayesian framework incorporating the stochastic response surface method. *Water Resources Research*, 39(12):1350, 2003.

O. Balci. Verification, validation, and accreditation of simulation models. In *Proceedings of the 29th Conference on Winter Simulation*, Atlanta, GA, 1997.

O. Balci and R. Sargent. Validation of simulation models via simultaneous confidence intervals. *American Journal of Mathematical and Management Sciences*, 4(3 & 4):375–406, 1984.

O. Balci and R. Sargent. Validation of multivariate response simulation models by using Hotelling's two-sample T^2 test. *Simulation*, 39(3):185–192, 1982.

O. Balci and R. Sargent. A methodology for cost-risk analysis in the statistical validation of simulation models. *Communications of the ACM*, 24(4):190–197, 1981.

H. Banks. Remarks on uncertainty assessment and management in modeling and computation. *Mathematical and Computer Modeling*, 33(1–3):39–47, 2001.

M. J. Bayarri, J. O. Berger, D. Higdon, M. C. Kennedy, A. Kottas, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C. H. Lin, and J. Tu. A framework for validation of computer models. Technical Report 128, National Institute of Statistical Sciences, Research Triangle Park, NC, 2002.

M. C. Bernardo, R. J. Buck, L. Liu, W. A. Nazaret, J. Sacks, and W. J. Welch. Integrated circuit design optimization using a sequential strategy. *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, 11(3):361–372, 1992.

K. Beven and A. Binley. The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, 6:279–298, 1992.

B. J. Bichon, M. S. Eldred, L. P. Swiler, S. Mahadevan, and J. M. McFarland. Efficient global reliability analysis for nonlinear implicit performance functions. *AIAA Journal*, 2008. submitted for publication.

K. Campbell. Statistical calibration of computer simulations. *Reliability Engineering and System Safety*, 91(10–11):1358–1363, 2006.

W. Chen, L. Baghdasaryan, T. Buranathiti, and J. Cao. Model validation via uncertainty propagation. *AIAA Journal*, 42:1406–1415, 2004.

S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *American Statistician*, 49(4):327–335, 1995.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2001.

P. S. Craig, M. Goldstein, A. H. Seheult, and J. A. Smith. Bayes linear strategies for matching hydrocarbon reservoir history. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 5*, pages 69–95. Oxford University Press, Oxford, 1996.

C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86(416):953–963, 1991.

B. Debusschere, H. Najm, P. Pebay, O. Knio, R. Ghanem, and O. Le Maitre. Numerical challenges in the use of polynomial chaos representations for stochastic processes. *SIAM Journal of Scientific Computing*, 26(2):698–719, 2004.

J. L Devore. *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole, 5^{th} edition, 2000.

J. Donaldson and B. Schnabel. Computational experience with confidence regions and confidence intervals for nonlinear least squares. *Technometrics*, 29(1):67–82, 1987.

K. Dowding, R. G. Hills, I. Leslie, M. Pilch, B. M. Rutherford, and M. L. Hobbs. Case study for model validation: assessing a model for thermal decomposition of polyurethane foam. Technical Report SAND2004-3632, Sandia National Laboratories, Albuquerque, NM, 2004.

K. Dowding, M. Pilch, and R. Hills. Formulation of the thermal problem. *Computer Methods in Applied Mechanics and Engineering*, 2008. In press, available online.

D. Dubois and H. Prade. Possibility theory, probability theory and multiple-valued logics: A clarification. *Annals of Mathematics and Artificial Intelligence*, 32:55–66, 2001.

M. S. Eldred, S. L. Brown, B. M. Adams, D. M. Dunlavy, D. M. Gay, L. P. Swiler, A. A. Giunta, W. E. Hart, J. P. Watson, J. P. Eddy, J. D. Griffin, P. D. Hough, T. G. Kolda, M. L. Martinez-Canales, and P. J. Williams. DAKOTA, a multilevel parallel object-oriented framework for

design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: Version 4.0 reference manual. Technical Report SAND2006-4055, Sandia National Laboratories, October 2006.

K. L. Erickson, S. M. Trujillo, K. R. Thompson, A. C. Sun, M. L. Hobbs, and K. J. Dowding. Liquefaction and flow behavior of a thermally decomposing removable epoxy foam. In A. A. Mammoli and C. A. Brebbia, editors, *Computational Methods in Materials Characterisation*, pages 217–242. WIT press, Southampton-Boston, 2004.

C. Farrar and H. Sohn. Pattern recognition for structural health monitoring. In *Proceedings of the Second MCEER Workshop on Mitigation of Earthquake Disaster by Advanced Technologies*, Las Vegas, NV, November 2000.

J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–141, 1991.

R. Ghanem. Ingredients for a general purpose stochastic finite elements formulation. *Computer Methods in Applied Mechanics and Engineering*, 168(1-4):19–34, 1999.

R. Ghanem and S. Dham. Stochastic finite element analysis for multiphase flow in heterogeneous porous media. *Transport in Porous Media*, 32:239–262, 1998.

R. Ghanem, A. Doostan, and J. Red-Horse. A probabilistic construction of model validation. *Computer Methods in Applied Mechanics and Engineering*, 2008. In press.

R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag, New York, 1991.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, Boca Raton, 1996.

A. A. Giunta, J. M. McFarland, L. P. Swiler, and M. S. Eldred. The promise and peril of uncertainty quantification using response surface approximations. *Structure and Infrastructure Engineering*, 2(3–4):175–189, 2006.

C. A. Glasbey. Correlated residuals in non-linear regression applied to growth data. *Applied Statistics*, 28:251–259, 1979.

C. A. Glasbey. Nonlinear regression with autoregressive time series errors. *Biometrics*, 36: 135–140, 1980.

R. F. Guratzsch. *Sensor Placement Optimization under Uncertainty for Structural Health Monitoring Systems of Hot Aerospace Structures*. PhD thesis, Vanderbilt University, 2007.

C. Haas. On modeling correlated random variables in risk assessment. *Risk Analysis*, 19: 1205–1214, 1999.

A. Haldar and S. Mahadevan. *Probability, Reliability, and Statistical Methods in Engineering Design*. John Wiley and Sons, Inc., New York, 2000.

D. Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2):383–385, 1974.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

R. G. Hills and I. Leslie. Statistical validation of engineering and scientific models: validation experiments to application. Technical Report SAND2003-0706, Sandia National Laboratories, Albuquerque, NM, 2003.

R. Iman and W. Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics – Simulation and Computation*, 11:311–334, 1982.

ISO. Quality management and quality assurance standards – Part 3: Guidelines for the application of ISO 9001 to the development, supply and maintenance of software. Technical Report ISO 9000-3, International Standards Organization, Geneva, Switzerland, 1991.

S. S. Isukapalli, A. Roy, and P. G. Georgopoulos. Stochastic Response Surface Methods (SRSMs) for uncertainty propagation: application to environmental and biological systems. *Risk Analysis*, 18:351–363, 1998.

A. Jain and D. Zongker. Feature selection: evaluation, application and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:153–158, 1997.

H. S. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, 3rd edition, 1961.

X. Jiang and S. Mahadevan. Bayesian risk-based decision method for model validation under uncertainty. *Reliability Engineering and System Safety*, 92:707–718, 2007.

M. E. Johnson. *Multivariate Statistical Simulation*. John Wiley, New York, 1981.

D. Jones, C. Perttunen, and B. Stuckman. Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Application*, 79(1):157–181, 1993.

D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

I. Kaymaz. Application of kriging method to structural reliability problems. *Structural Safety*, 27(2):133–151, 2005.

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

M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000a.

M. C. Kennedy and A. O'Hagan. Supplementary details on Bayesian calibration of computer codes. University of Sheffield, Sheffield, 2000b. Available from http://www.shef.ac.uk/~st1ao/ps/calsup.ps.

M. C. Kennedy, C. W. Anderson, S. Conti, and A. O'Hagan. Case studies in Gaussian process modelling of computer codes. *Reliability Engineering and System Safety*, 91:1301–1309, 2006.

P. Knupp. *Verification of Computer Codes in Computational Science and Engineering*. Chapman & Hall/CRC, Boca Raton, FL, 2002.

W. L. G. Koontz and K. Fukunaga. A nonlinear feature extraction algorithm using distance transformation. *IEEE Transactions on Computers*, C-21(1):56–62, January 1972.

P. Lee. *Bayesian Statistics, an Introduction*. Oxford University Press, Inc., New York, 2004.

Y. C. Liang, H. P. Lee, S. P. Lim, W. Z. Lin, K. H. Lee, and C. G. Wu. Proper orthogonal decomposition and its applications—Part I: Theory. *Journal of Sound and Vibration*, 252 (3):527–544, 2002.

H. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62:399–402, 1967.

P. Lurie and M. Goldberg. An approximate method for sampling correlated random variables from partially-specified distributions. *Management Science*, 44:203–218, 1998.

S. Mahadevan and R. Rebba. Validation of reliability computational models using Bayes networks. *Reliability Engineering and System Safety*, 87:223–232, 2005.

K. Mardia and R. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135–146, 1984.

J. Martin and T. Simpson. Use of kriging models to approximate deterministic computer models. *AIAA Journal*, 43(4):853–863, 2005.

Y. Marzouk, H. Najm, and L. Rahn. Stochastic spectral methods for efficient Bayesian solution of inverse problems. *Journal of Computational Physics*, 224(2):560–568, June 2007.

J. McFarland and S. Mahadevan. Multivariate significance testing and model calibration under uncertainty. *Computer Methods in Applied Mechanics and Engineering*, 2008. In press.

M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21 (2):239–245, 1979.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

J. More', B. Garbow, and K. Hillstrom. Minpack. Argonne National Laboratory, 1999. Software available online at http://www.netlib.org/minpack/.

M. D. Morris, T. J. Mitchell, and D. Ylvisaker. Bayesian design and analysis of computer experiments: use of derivatives in surface prediction. *Technometrics*, 35(3):243–255, 1993.

J. T. Nakos. Uncertainty analysis of thermocouple measurements used in normal and abnormal thermal environment experiments at Sandia's radiant heat facility and Lurance Canyon burn site. Technical Report SAND2004-1023, Sandia National Laboratories, Albuquerque, NM, 2004.

N. C. Nigam. *Introduction to Random Vibrations*. MIT Press, Cambridge, MA, 1983.

J. Oakley and A. O'Hagan. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89:769–784, 2002.

W. Oberkampf and M. Barone. Measures of agreement between computation and experiment: validation metrics. *Journal of Computational Physics*, 217:5–36, 2006.

W. Oberkampf and T. Trucano. Verification and validation in computational fluid dynamics. Technical Report SAND2002-0529, Sandia National Laboratories, Albuquerque, New Mexico, 2002.

A. O'Hagan. Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety*, 91:1290–1300, 2006.

T. Paez and A. Urbina. Validation of mathematical models of complex structural dynamic system. In *Proceedings of the Ninth International Congress on Sound and Vibration*, Orlando, FL, 2002.

A. Papoulis and S. U. Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, New York, 2002.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554, 1971.

R. Paulo. Default priors for Gaussian processes. *The Annals of Statistics*, 33:556–582, 2005.

C. Rasmussen. *Evaluation of Gaussian processes and other methods for non-linear regression*. PhD thesis, University of Toronto, 1996.

J. O. Rawlings, S. G. Pantula, and D. A. Dickey. *Applied Regression Analysis*. Springer-Verlag, New York, 1998.

R. Rebba. *Model Validation and Design Under Uncertainty*. PhD thesis, Vanderbilt University, 2005.

R. Rebba and S. Mahadevan. Computational methods for model reliability assessment. *Reliability Engineering and System Safety*, 2007. In press, available online August 10, 2007.

R. Rebba and S. Mahadevan. Validation of models with multivariate output. *Reliability Engineering and System Safety*, 91:861–871, 2006.

R. Rebba, S. Mahadevan, and S. Huang. Validation and error estimation of computational models. *Reliability Engineering and System Safety*, 91:1390–1397, 2006.

J. R. Red-Horse and T. L. Paez. Sandia National Laboratories validation workshop: structural dynamics application. *Computer Methods in Applied Mechanics and Engineering*, 2008. In press, available online.

B. Ripley. *Spatial Statistics*. John Wiley, New York, 1981.

P. Roache. *Verification and Validation in Computational Science and Engineering*. Hermosa Publishers, Albuquerque, NM, 1998.

V. J. Romero, J. W. Shelton, and M. P. Sherman. Modeling boundary conditions and thermocouple response in a thermal experiment. In *Proceedings of the 2006 International Mechanical Engineering Congress and Exposition*, number IMECE2006-15046, Chicago, IL, November 2006. ASME.

J. Sacks and S. Schiller. Spatial designs. In S. S. Gupta and J. O. Berger, editors, *Statistical Decision Theory and Related topics IV*, volume 2, pages 385–399. Springer-Verlag, New York, 1988.

J. Sacks, S. Schiller, and W. Welch. Designs for computer experiments. *Technometrics*, 31(1): 41–47, February 1989a.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–435, 1989b.

A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. *Sensitivity analysis in practice: a guide to assessing scientific models*. Wiley, 2004.

J. W. Sammon. A nonlinear mapping algorithm for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401–409, May 1969.

CALORE-MAN. *Calore: A Computational Heat Transfer Program. Vol. 2 User Reference Manual for Version 4.1*. Sandia National Laboratories, Albuquerque, NM, 2005.

T. J. Santner, B. J. Williams, and W. I. Noltz. *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York, 2003.

R. Sargent. Validation and verification of simulation models. In *2004 Winter Simulation Conference*, volume 1, pages 1–28, 2004.

S. Schlesinger. Terminology for model credibility. *Simulation*, 32(3):103–104, 1979.

G. A. F. Seber and C. J. Wild. *Nonlinear Regression*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2003.

D. Segalman. A four-parameter Iwan model for lap-type joints. Technical Report SAND2002-3828, Sandia National Laboratories, Albuquerque, NM, 2002.

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

S. Shapiro and M. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, New York, 1986.

J. S. Simonoff. *Smoothing methods in statistics*. Springer-Verlag, New York, 1996.

T. W. Simpson, J. D. Peplinski, P. N. Koch, and J. K. Allen. On the use of statistics in design and the implications for deterministic computer experiments. In *Proceedings of the ASME Design Engineering Technical Conferences*, Sacramento, CA, September 1997.

T. W. Simpson, T. M. Mauery, J. J. Korte, and F. Mistree. Kriging models for global approximation in simulation-based multidisciplinary design optimization. *AIAA Journal*, 39(12): 2233–2241, 2001.

D. Smallwood. Damping investigations of a simplified frictional shear joint. Technical Report SAND2000-1929C, Sandia National Laboratories, Albuquerque, NM, 2000.

M. Srivastava. *Methods of Multivariate Statistics*. John Wiley and Sons, Inc., New York, 2002.

J. Stigter and M. Beck. A new approach to the identification of model structure. *Environmetrics*, 5(3):315–333, 1994.

C. W. Therrien. *Decision Estimation and Classification: An Introduction to Pattern Recognition and Related Topics*. John Wiley and Sons, New York, 1989.

T. Trucano, L. Swiler, T. Igusa, W. Oberkampf, and M. Pilch. Calibration, validation, and sensitivity analysis: what's what. *Reliability Engineering and System Safety*, 91(10–11), 2006.

A. Urbina, T. Paez, D. Gregory, and B. Resor. Probabilistic modeling of mechanical joints. Technical Report SAND2003-1456C, Sandia National Laboratories, Albuquerque, NM, 2003a.

A. Urbina, T. Paez, T. Hasselman, G. Wathugala, and K. Yap. Assessment of model accuracy relative to stochastic system behavior. Technical Report SAND2003-1279C, Sandia National Laboratories, Albuquerque, NM, 2003b.

A. Urbina, T. Paez, D. Segalman, F. Bitsie, and D. Gregory. Validation of a mechanical joint. In *9^{th} ASCE Specialty Conference on Probabilistic Mechanics and Structural Reliability*, Albuquerque, NM, July 2004.

A. Urbina, T. Paez, D. Gregory, and B. Resor. Validation of a multi-jointed mechanical system model. In *2005 Annual Meeting of the Society of Experimental Mechanics*, Portland, OR, June 2005.

E. Vazquez and E. Walter. Estimating derivatives and integrals with Kriging. In *44^{th} IEEE Conference on Decision and Control*, pages 8156–8161, December 2005.

A. Vecchia and R. Cooley. Simultaneous confidence and prediction intervals for nonlinear regression models, with application to a groundwater flow model. *Water Resources Research*, 23(7):1237–1250, 1987.

K. W. Vugrin, L. P. Swiler, R. M. Roberts, and N. J. Stucky-Mack. Confidence region estimation techniques for nonlinear regression in groundwater flow: Three case studies. *Water Resources Research*, 43, 2007.

J. Wang and N. Zabaras. Using bayesian statistics in the estimation of heat source in radiation. *International Journal of Heat and Mass Transfer*, 48:15–29, 2005.

W. J. Welch. A mean squared error criterion for the design of experiments. *Biometrika*, 70: 205–213, 1983.

L. Zadeh. Fuzzy sets as the basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.