Bayesian Uncertainty Analysis for Complex Computer Codes

Jeremy Oakley

Thesis submitted to the University of Sheffield

for the degree of Doctor of Philosophy

Department of Probability and Statistics

School of Mathematics and Statistics

December, 1999

Acknowledgements

Firstly, I would like to thank my supervisor, Tony O'Hagan, for his enthusiasm and guidance throughout my PhD studies. I am also grateful to the Engineering and Physical Sciences Research Council for the funding of my studentship and the National Radiological Protection Board for their additional financial support.

I would also like to thank the following people for their assistance: Richard Haylock for providing the iodine model example, and his general assistance regarding my time working at NRPB, Edward Glennie at WRc for providing the SIMPOL model data, Joanne Brown for her participation in the elicitation exercise and her advice regarding the FARMLAND model, and Marc Kennedy for his assistance with the APL programming language, and helpful discussions about Gaussian processes.

Finally, I would like to thank all my friends both in Nottingham and Sheffield for a fantastic three years. In particular, my housemates Mark, Brad and Sacha, and my fellow students Raquel, Martin, Caterina, Stefan, John, Julie and Sofia.

Summary

A deterministic computer model returns an output y when provided with an input $\mathbf{x}$. The model is to be used in a situation where the true value of the input, $\mathbf{X}$, corresponding to a particular set of circumstances is unknown. Our aim is to make inferences about the value of the output Y obtained if the model was to be run at the true input $\mathbf{X}$. An expert provides a distribution $G(\mathbf{x})$ describing their uncertainty about $\mathbf{X}$, and from this we then attempt to learn about the induced distribution of Y. This problem is known to users of computer models as uncertainty analysis.

A simple Monte Carlo solution to this problem involves drawing a large sample of inputs from $G(\mathbf{x})$, running the model at each sampled input, and basing inferences about Y on the sample of resulting outputs. We are interested in the case when the computer model is computationally expensive, to the extent that Monte Carlo methods are not practical, due to the computing times required. We consider a Bayesian approach which uses the Gaussian process prior for unknown functions to learn about the computer code itself. We run the code at a particular set of inputs, and then consider our beliefs about the output of the code at further untested inputs. This enables us to make accurate inferences about Y using a small number of runs of the code.

In this thesis, we make inferences about two summaries of Y, the distribution and density functions. We also consider choosing which inputs to run the code at to learn about specific percentiles of Y. Finally, we investigate the use of expert prior knowledge about the code itself to further reduce the number of runs of the code required. We consider a simulation approach involving generating random functions which can used to obtain a variety of inferences about Y, and which also bypasses some of the difficulties encountered when trying to obtain various summaries analytically. We demonstrate that it is possible to obtain inferences about Y that are comparable in accuracy to Monte Carlo estimates, and yet require considerably fewer runs of the computer code. Examples are given regarding computer models in the field of radiological protection, and in the design of a sewer network that is required to meet certain environmental standards.

Contents

1 Introduction **1**

1.1 Computer models and statistics 1

1.2 Uncertainty Analysis 3

1.3 Notation 4

1.4 The classical approach to uncertainty analysis 5

1.5 Uncertainty analysis for computationally expensive computer models 6

2 Inference about functions **8**

2.1 Introduction 8

2.2 Bayesian inference using Gaussian processes 10

2.3 The Gaussian Process model for derivatives of functions 32

2.4 Conclusions 35

3 Uncertainty analysis using simulation **37**

3.1 Introduction 37

3.2 Generating random functions 38

3.3 Example: uncertainty analysis with alternative covariance functions . 51

3.4 Conclusions 51

4 The distribution and percentile functions **52**

4.1 Introduction 52

4.2 Posterior moments of $F_Y(y)$ 53

4.3 Alternative methods for estimating $F_Y(y)$ 59

4.4 Estimating the percentile function 61

4.5 Example: the ^{131}I algorithm 61
4.6 Comparison between the Bayesian and Monte Carlo approaches . . . 70
4.7 Conclusions 71

5 Estimating the density function 73
5.1 Introduction 73
5.2 Posterior moments 75
5.3 Alternative approaches to estimating the density function 77
5.4 Density estimation via the simulation approach 82
5.5 Conclusions 86

6 Optimal designs for estimating percentiles 90
6.1 Introduction 90
6.2 The SIMPOL model 91
6.3 Methodology for estimating the 95th percentile 92
6.4 Choosing the simulation design points 94
6.5 Obtaining a true value of the 95th percentile 95
6.6 Applying the Bayesian approach to the SIMPOL model 96
6.7 Conclusions 104

7 Eliciting Prior Beliefs 106
7.1 Introduction 106
7.2 Prior to posterior analysis 107
7.3 Example: a one dimensional function with a proper prior distribution 108
7.4 Methodology for eliciting prior beliefs 110
7.5 Example: the FARMLAND model 117
7.6 Conclusions 127

8 Discussion 129

Chapter 1

Introduction

1.1 Computer models and statistics

Computer models are used in a variety of scientific fields. In some applications they may be used to predict future events, for example in weather or economic forecasting. When experimentation on real life systems is too costly or impractical, a computer model of the system may be used as a surrogate. An example of this scenario would be investigating the dispersal of a pollutant from a source under a variety of atmospheric conditions. We think of the computer model as returning a number of outputs when provided with a set of inputs that will correspond to the situation being modelled. Typically, these models are deterministic, so that running them repeatedly at the same inputs will always give the same outputs. We describe the process of running a computer model at a variety of different input values as a 'computer experiment'.

There will be some drawbacks in using a computer model as a substitute for reality. Firstly, the model will generally be an approximation to the real process. Consequently, there may be discrepancies between outcomes predicted by the model and outcomes observed in reality. When real-life observations are available, there will be interest in incorporating this information into the computer model, to improve its accuracy. In constructing the model, variables known to be correlated to the output of interest may be included as parameters or inputs in the model. In some

cases it may not be possible to measure the exact values of some of these variables. Thus there will be uncertainty in some of the model parameters which may then be propagated through to the output. A further complication is when the model is computationally expensive, so that obtaining a single output at a specific set of inputs may take a considerable amount of computing time. In this case, if the output is required over a range of different input values, the choice of which actual input values to run the model at is also an issue.

In recent years, these problems associated with the use of computer models have attracted the interest of statisticians, even though the models themselves have no random components. In the simplest case, only the output of the computer model is used; real life observations are not considered. In this case, there is a role for statistics in the experiment when it is only practical to evaluate the output of the model at a limited number of different inputs. After running the model at various inputs, there will then be interest in predicting the output of the model at different inputs that have not been tried. A design issue is also apparent. If we wish to make good predictions about the output of the model, we need to consider what choice of inputs to run the model at initially will lead to the best predictions. The earliest known work in choosing designs for computer experiments is McKay et al. (1979), who developed the Latin Hypercube sampling scheme. This is described in chapter two. A good discussion of the role of statistics in computer experiments is given in Sacks et al., (1989) and both the design and prediction problems are considered. Designs for computer experiments is the subject of Sacks et al. (1989a), and predicting the output of the model from a Bayesian perspective is discussed in Currin et al. (1991) and Mitchell et al. (1993).

A more complex situation is when information from real observations is also included in the computer experiment. In this case, we may wish to use the observations to learn about the uncertain inputs in the model, i.e. calibrate the model to the real observations. It may be desirable to base future predictions both on the observations and the output of the model, particularly in cases when there is little real data available. These are both considered in Kennedy and O'Hagan (1998). A similar concept to the calibration idea is history matching, which involves finding

values of the model inputs such that the outputs are close to those observed in reality. An example of this from a Bayesian perspective is in Craig et al. (1996).

Another point of interest is that of the sensitivity of the model output to the various model inputs, particularly in the case when the model input is high dimensional. In certain cases it may be necessary to first identify a subset of influential inputs before proceeding with the main analysis of interest. See for example the uncertainty analysis of a twenty-nine dimensional model in Haylock (1997). A method for assessing which inputs are influential is given in Sacks et al. (1989), and a Bayesian approach is described in O'Hagan et al. (1999). Sensitivity analysis is also the subject of the book edited by Saltelli et al. (1999).

1.2 Uncertainty Analysis

This thesis concentrates on the problem of uncertainty in the model parameters or inputs, which is known as uncertainty analysis. We explain the concept of uncertainty analysis with the following simple example. Suppose there has been an accidental release of radionuclides from a point source, and we wish to predict the resulting concentration of radionuclides at a particular location near to the source. Mathematical models have been developed to perform such a task, for example, the Gaussian Plume Diffusion model in Clarke (1979). The model assumes that the dispersion of radionuclides in the horizontal and vertical directions can be described by Gaussian distributions, and the simplest version of the model is given by

$$C(x,y,z) = \frac{Q}{2\pi u_{10}\sigma_z\sigma_y} \exp\left[-\frac{1}{2}\left\{\frac{y^2}{\sigma_y^2} + \frac{(z-h)^2}{\sigma_z^2}\right\}\right], \tag{1.1}$$

where C is air concentration of the radionuclide, Q is the total amount released, u_{10} is the wind speed at $10m$ above ground, σ_y and σ_z are the standard deviations of the horizontal and vertical Gaussian distributions respectively, h is the release height, and (x, y, z) are the coordinates along the wind direction, cross wind and above ground respectively. If we know the values of the inputs relating to a particular release, we can then make a prediction of the concentration at any location using this model. However, in practise it is quite likely that we will not know the values

of all the inputs. It may not be possible to obtain a measurement of the intial quantity released, Q, or the wind speed u_{10} may be unknown. If we do not know the true values of all the inputs, we will not know the value of the output of the model evaluated at the true inputs. Consequently, this output, which we consider to be the 'true' output, is a random variable. The aim of uncertainty analysis is to learn about the uncertainty induced in the true output by the uncertainty in the inputs.

Since this model is only an approximation of the real life process, it is unlikely that even if we do know the true values of the all the input parameters, the output of the model will be the same as the observed concentration in reality. In addition, it may not always be meaningful to talk about a 'true' value of an input parameter. For example, we can think of the true value of Q as being the precise quantity of a radionuclide released during a particular accident. However, if we consider the two standard deviation parameters σ_y and σ_z, then if the real behaviour of the plume of radionuclides is not perfectly described by Gaussian distributions, then what do the 'true' values of these parameters represent? These issues are not considered in uncertainty analysis. In the case of predicting some event in the future, the computer model may be the only source of information available, and the uncertainty resulting from unknown inputs needs to be understood. A model may give a very accurate prediction given all the correct inputs, but still be rendered ineffective if uncertainty about a particular input value results in high uncertainty about the output. A decision to invest resources in learning more about model inputs can be guided by an uncertainty analysis.

1.3 Notation

We define the inputs of a computer model to be some vector $\mathbf{x}$, and the output of the model to be a scalar y. In many cases, the computer model will return a vector of outputs, but here we confine our attention to a scalar output. We then represent the computer code by some function $\eta(.)$, so that the relationship between the inputs and output is given by

$$y = \eta(\mathbf{x}). \tag{1.2}$$

We now define $\mathbf{X}$ to be the true values of the unknown inputs, and Y to be the output of the model when run at the true inputs, so that

$$Y = \eta(\mathbf{X}). \tag{1.3}$$

Since $\mathbf{X}$ is unknown, we must consider our beliefs about $\mathbf{X}$. Knowledge about $\mathbf{X}$ is represented by the probability distribution $G(\mathbf{x})$. The distribution $G(\mathbf{x})$ will be derived from someone with expert knowledge about $\mathbf{X}$. The elicitation of $G(\mathbf{x})$ is not considered here. We will denote the sample space of $\mathbf{X}$ by $\mathcal{X}$. We can now give a definition of uncertainty analysis as follows:

> If, for some computer code represented by a function $\eta(.)$, the true input $\mathbf{X}$ is unknown and has distribution $G(\mathbf{x})$, what is the distribution of $Y = \eta(\mathbf{X})$?

The distribution of Y is known as the uncertainty distribution.

1.4 The classical approach to uncertainty analysis

The above question can be answered using Monte Carlo methods. We first draw a sample of random inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $G(\mathbf{x})$. Then for each input we run the code, and so we obtain a sample of random outputs $y_1 = \eta(\mathbf{x}_1), \ldots, y_n = \eta(\mathbf{x}_n)$. This sample can then be used to make inferences about Y. In principle, we can obtain an inference about Y to any desired accuracy by making the sample of outputs sufficiently large. A more efficient approach involves the use of Latin Hypercube Sampling, which is designed to ensure that the whole sample space of $\mathbf{X}$ is represented, without having to use a very large sample size. A full description of this is given in chapter two. Examples of classical uncertainty analyses are in Crick et al. (1988) and Helton et al. (1991).

1.5 Uncertainty analysis for computationally expensive computer models

In this thesis, we are interested in the specific case when the computer model is computationally expensive, so that obtaining one single output of the model takes a significant amount of time. In this case, the Monte Carlo approach described may not be suitable, since it may not be practical to run the code a sufficient number of times to obtain an accurate inference about Y. Our objective is to learn as much as possible about Y based on a small number of runs of the code. We attempt to do this from a Bayesian perspective. At the heart of this approach is the idea that the function $\eta(.)$ is itself a random variable. Learning the value of the output $y = \eta(\mathbf{x})$ for any input $\mathbf{x}$ is a non-trivial exercise. We will only be able to evaluate $\eta(.)$ at a small number of distinct inputs, and for all untested inputs $\mathbf{x}$ we will have to consider our beliefs about the value of $\eta(\mathbf{x})$. It is by considering all the information that we gain about $\eta(.)$ once $\eta(\mathbf{x})$ has been evaluated that will allow us to make inference about Y more efficiently. When using Monte Carlo methods, only the value of $\eta(\mathbf{x})$ is used from the run of the code; all other information is ignored.

1.5.1 Existing methods in Bayesian uncertainty analysis

The Bayesian approach to uncertainty analysis was pioneered by Haylock and O'Hagan (1996). Two summaries of the distribution of Y were considered, its mean and variance. An example was given involving a particular computer model with two unknown inputs where they were able to achieve an estimate of Y using ten runs of the code that was comparable in accuracy to a Monte Carlo estimate based on one thousand runs of the code. In Haylock (1997), the methodology was applied to a more complex model where fourteen inputs were considered uncertain.

1.5.2 Application: Radiological Protection

This work is supported by the National Radiological Protection Board (NRPB). The NRPB was set up following the Radiological Protection Act, 1970. Its purpose is

to advise the UK government on matters regarding protecting the public from all forms of radiation. The use of computer models in this area are common, and in many cases there will be uncertainty in some of the model inputs.

1.5.3 Overview of the remaining chapters

The central theory used in this thesis involves making inferences about functions, and this is explored in chapter two. Given a model for the function $\eta(.)$, it can be useful in many situations to make random draws from the distribution of $\eta(.)$, and in chapter three, we describe a method for simulating random functions, and consider the various practical issues that are concerned. In chapters four and five, we address the uncertainty analysis problem, and make inferences about the distribution and density functions of Y. We attempt to derive these summaries analytically, and encounter various difficulties which can be resolved using the simulation method described in chapter three. In chapter six, we consider the issue of choosing suitable design points to run the code at with the purpose of obtaining an estimate of the 95th percentile of Y. The work in this chapter was motivated by a specific problem brought to us by the Water Research Centre (WRc). In chapter seven, we consider the use of proper prior information about the function $\eta(.)$, and give an example involving a model used by the NRPB to predict concentrations of radionuclides in grain following accidental releases. A discussion of the results obtained and of areas for future research is given in chapter eight.

Chapter 2

Inference about functions

2.1 Introduction

In this chapter we review methods for making inferences about unknown functions, focusing on the Gaussian process model, used in the context of uncertainty analysis by Haylock and O'Hagan (1996). We investigate two complications that may be encoutered in the use of this model, involving the specification of the prior mean of the unknown function, and the unknown parameters in the correlation function describing the correlations between outputs at different inputs. We review various techniques for choosing suitable design points, and consider an analytic approach conditional on some assumptions about the input distribution $G(\mathbf{x})$ and the correlation function. We also describe the extensions to making inferences about derivatives of an unknown function, as given in O'Hagan (1994). Finally, we explore the use of derivatives in quadrature involving unknown functions.

We have a function denoted by $y = \eta(\mathbf{x})$, and we are restricted to evaluating $\eta(.)$ at a limited number of distinct values of $\mathbf{x}$. We run the code at the inputs $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and obtain data $(y_1 = \eta(\mathbf{x}_1), \ldots, y_n = \eta(\mathbf{x}_n))$. Based on these data, we then want to make inferences about $\eta(\mathbf{x})$ for other values of $\mathbf{x} \in \mathcal{X}$. This will involve some form of interpolation of the function. It is unlikely that the chosen interpolant will predict the output exactly. Consequently, we need a means of allowing for the uncertainty in the interpolation. The computation required for our method of

inference about $\eta(.)$ also needs to be taken into consideration. The computing time required should not exceed the time needed to obtain accurate inferences about Y using Monte Carlo methods. A review of various approaches to interpolation is given in Diaconis (1988). Of particular interest is an account of work by Poincaré (1898), who considered the problem from a Bayesian perspective.

We proceed using a stochastic process approach. We begin with the following model for $\eta(.)$:

$$\eta(\mathbf{x}) = \sum_{i=1}^{n} \beta_i h_i(\mathbf{x}) + Z(\mathbf{x}), \tag{2.1}$$

where for each value of i, $h_i(\mathbf{x})$ is a known function of $\mathbf{x}$ and β_i is an unknown coefficient. The function $Z(.)$ is a stochastic process with mean zero, and covariance between $Z(\mathbf{x})$ and $Z(\mathbf{x}')$ given by some function $C(\mathbf{x}, \mathbf{x}')$. Thus we begin with some suitable regression model and then suppose that the function $\eta(.)$ deviates from this according to a stationary stochastic process. This provides a framework for interpolation within both the Bayesian and frequentist procedures, though there are problems with the interpretation of this model from a frequentist viewpoint, assuming the computer model is not actually generated by a stochastic process. We adopt a Bayesian approach and treat $Z(.)$ as a Gaussian process. This is discussed fully in the next section.

Kimeldorf and Wahba (1970) used this model in a Bayesian framework for predicting the output of a function. They chose a correlation function $C(\mathbf{x}, \mathbf{x}')$ such that the interpolant was a spline function. Sacks et al (1989) proceed using an idea based on kriging (Matheron, 1963). Given observations $\mathbf{y}^T = (y_1 = \eta(\mathbf{x}_1), \ldots, y_n = \eta(\mathbf{x}_n))$, they then consider a linear predictor

$$\hat{\eta}(\mathbf{x}) = c'(\mathbf{x})\mathbf{y}, \tag{2.2}$$

to predict the output at one untried input $\mathbf{x}$. A frequentist approach is adopted and the data $\mathbf{y}$ are treated as being random. Consequently, the predictor $\hat{\eta}(\mathbf{x})$ is then thought of as being a function of some random observations $\mathbf{Y}$. The best linear unbiased predictor is found by choosing $c(\mathbf{x})$ that will minimise

$$E_{\mathbf{Y}}\{c'(\mathbf{x})\mathbf{Y} - \eta(\mathbf{x})\}^2, \tag{2.3}$$

subject to the constraint that

$$E_{\mathbf{Y}}\{c'(\mathbf{x})\mathbf{Y}\} = E\{\eta(\mathbf{x})\}, \tag{2.4}$$

i.e. the estimator must be unbiased.

2.2 Bayesian inference using Gaussian processes

The Gaussian process model for Z(.) has been used in various applications. O'Hagan (1978) and Neal (1999) used a Gaussian process prior for functions in regression. Currin et al. (1991) used Gaussian processes for making predictions about outputs of computer models at untested inputs. O'Hagan (1991) was concerned with integrals of computationally expensive functions. Neal (1999) also used Gaussian process priors for classification problems.

In the Bayesian approach, we begin by treating $\eta(\mathbf{x})$ as a random variable. We consider $\eta(.)$ to be random simply in the sense that it is unknown, as opposed to being the product of some random process. The computer model can tell us the value of $\eta(\mathbf{x})$ exactly, but until the code is run, this quantity is unknown, and so we can describe our beliefs about $\eta(\mathbf{x})$ through a probability distribution. Thus we first need to describe our prior beliefs about $\eta(.)$.

If we have no prior knowledge at all about $\eta(.)$, then we might consider stating that $E[\eta(\mathbf{x})] = 0 \quad \forall \mathbf{x}$, by symmetry. Alternatively, we may have some knowledge about the general form of $\eta(.)$. For instance, we might believe that $\eta(\mathbf{x})$ is approximately linear in $\mathbf{x}$. In this case, we would write

$$E[\eta(\mathbf{x})] = \beta_0 + \beta_1\mathbf{x}. \tag{2.5}$$

In general, we state that

$$E[\eta(x)|\boldsymbol{\beta}] = \mathbf{h}(\mathbf{x})^T\boldsymbol{\beta}. \tag{2.6}$$

The vector $\mathbf{h}(.)$ consists of q known regression functions of $\mathbf{x}$, and is chosen to incorporate the beliefs that we have about $\eta(.)$. The vector $\boldsymbol{\beta}$ will consist of q unknown coefficients. We will proceed using this form for the prior mean, as in

practice we expect that it will always be possible to propose an appropriate form for $\mathbf{h}(.)$.

We now need to consider how we expect the true function $\eta(.)$ to deviate from $\mathbf{h}(.)^T\boldsymbol{\beta}$. We restrict our attention to functions $\eta(.)$ that are relatively smooth. Consequently, we expect there to be a high correlation between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ if $\mathbf{x}$ and $\mathbf{x}'$ are sufficiently close. Furthermore, the correlation should decrease as the distance between $\mathbf{x}$ and $\mathbf{x}'$ increases. We define the covariance between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ to be

$$Cov(\eta(\mathbf{x}), \eta(\mathbf{x}')|\sigma^2) = \sigma^2 c(\mathbf{x}, \mathbf{x}'), \tag{2.7}$$

for some function $c(.,.)$, which will decrease as $|\mathbf{x}-\mathbf{x}'|$ increases, and satisfy $c(\mathbf{x}, \mathbf{x}) = 1$ for all $\mathbf{x}$, so that the variance of $\eta(\mathbf{x})$ conditional on σ^2 is σ^2. If we think of $c(.,.)$ as being a function of $|\mathbf{x} - \mathbf{x}'|$, then Bochner's theorem states that for $c(.,.)$ to be a valid correlation function, $c(|\mathbf{x}-\mathbf{x}'|)$ must be the characteristic function of a random variable whose distribution function is symmetric about the origin. (See for example Feller, 1966).

Finally we consider the distribution of $\eta(\mathbf{x})$ conditional on σ^2. A suitable and mathematically convenient choice is the Gaussian process model for $\eta(.)$. A Gaussian process can be thought of as an infinite collection of random variables with the property that any subset of these variables will have a multivariate normal distribution. Thus for any set of inputs $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, we model the corresponding outputs $\{\eta(\mathbf{x}_1), \ldots, \eta(\mathbf{x}_n)\}$ as having a multivariate normal distribution. Note that n can be any positive integer.

The next part of the prior model for $\eta(.)$ is the prior distribution for $\boldsymbol{\beta}$ and σ^2. We generally use a weak prior distribution:

$$p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}. \tag{2.8}$$

However, it is possible to include expert knowledge about $\eta(.)$ through the prior distribution for $\boldsymbol{\beta}$ and σ^2. This is discussed in chapter seven.

We now need to choose a form for the function $c(.,.)$. For any set of distinct inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the matrix A whose $i,j-$th element is $c(\mathbf{x}_i, \mathbf{x}_j)$ must be positive definite. Sacks et al. (1989) consider correlation functions which are products of

one dimensional correlations, specifically, functions of the form

$$c(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{r} \exp(-\theta_i |x_i - x_i'|^p), \tag{2.9}$$

assuming that $\mathbf{x} = (x_1, \ldots, x_r)^T$. Currin et al. (1991) consider a non-negative linear correlation function:

$$c(\mathbf{x}, \mathbf{x}') = \begin{cases} 1 - \frac{1}{\theta}|d| & |d| < \theta \\ 0 & |d| \geq \theta \end{cases} \tag{2.10}$$

where $|d|$ is the distance between $\mathbf{x}$ and $\mathbf{x}'$, and θ is positive. This results in an interpolant which is a linear spline. Another function suggested in Currin et al. (1991) is the non-negative cubic correlation function

$$c(\mathbf{x}, \mathbf{x}') = \begin{cases} 1 - 6\left(\frac{|d|}{\theta}\right)^3 & |d| < \frac{\theta}{2} \\ 2\left(1 - \frac{|d|}{\theta}\right)^3 & \frac{\theta}{2} \leq |d| < \theta \\ 0 & |d| \geq \theta \end{cases} \tag{2.11}$$

with $|d|$ and θ defined as before. In this case the interpolant is a cubic spline. Here, we confine our attention to the correlation function

$$c(\mathbf{x}, \mathbf{x}') = \exp\{-(\mathbf{x} - \mathbf{x}')^T B(\mathbf{x} - \mathbf{x}')\}, \tag{2.12}$$

where B is a diagonal matrix of smoothing parameters, i.e. the form used by Sacks et al. (1989) with p fixed at two. This will be appropriate when $\eta(.)$ is a smooth function, for reasons that will become apparent when we discuss derivatives of unknown functions in section 2.3. This form is convenient for certain computations, such as selecting good design points, as will be seen in section 2.2.3.

The $i, i-$th element of B describes how rough $\eta(.)$ is in the $i-$th dimension of its input. The smoother $\eta(.)$ is, the higher the correlation between two points $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ will be. The correlation between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ depends on the distance between $\mathbf{x}$ and $\mathbf{x}'$. The matrix B has the effect of rescaling the distance between $\mathbf{x}$ and $\mathbf{x}'$. Thus B determines how close two inputs $\mathbf{x}$ and $\mathbf{x}'$ need to be such that the correlation between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ takes a particular value. It is not practical to give B a fully analytical Bayesian treatment. Various procedures for dealing with B are discussed in section 2.2.1.

We now 'observe' $\eta(.)$ at n design points, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to obtain data

$$\mathbf{y}^T = (y_1 = \eta(\mathbf{x}_1), \ldots, y_n = \eta(\mathbf{x}_n)). \tag{2.13}$$

To update the distribution of $\eta(.)$, we use the following property of multivariate normal distributions:

Let $\mathbf{z}$, an $n \times 1$ vector, have a multivariate normal distribution, with mean $\boldsymbol{\mu}$ and variance-covariance matrix Σ. Now partition $\mathbf{z}$ into two vectors $\mathbf{z}_1$ and $\mathbf{z_2}$ of dimensions $p \times 1$ and $(n-p) \times 1$ respectively. Partition $\boldsymbol{\mu}$ and Σ accordingly into $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, so that $E(\mathbf{z}_i) = \boldsymbol{\mu_i}$, and $Cov(\mathbf{z}_i, \mathbf{z}_j) = \Sigma_{ij}$. Then $\mathbf{z}_1 | \mathbf{z}_2 = \mathbf{f}$ also has a multivariate normal distribution with mean $\boldsymbol{\mu_1} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{f} - \boldsymbol{\mu_2})$ and variance-covariance matrix $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. A proof of this result is given in Krzanowski (1989).

Noting that

$$\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2 \sim N(H\boldsymbol{\beta}, \sigma^2 A), \tag{2.14}$$

where

$$H^T = (\mathbf{h}(\mathbf{x}_1), \ldots, \mathbf{h}(\mathbf{x}_n)), \tag{2.15}$$

$$A = \begin{pmatrix} 1 & c(\mathbf{x}_1, \mathbf{x}_2) & \cdots & c(\mathbf{x}_1, \mathbf{x}_n) \\ c(\mathbf{x}, \mathbf{x}_1) & 1 & & \vdots \\ \vdots & & \ddots & \\ c(\mathbf{x}_n, \mathbf{x}_1) & \cdots & & 1 \end{pmatrix}, \tag{2.16}$$

it then follows that

$$\eta(.) \mid \boldsymbol{\beta}, \sigma^2, \mathbf{y} \sim N(m^*(.),\ \sigma^2 c^*(.,.)), \tag{2.17}$$

where

$$m^*(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T\boldsymbol{\beta} + \mathbf{t}(\mathbf{x})^T A^{-1}(\mathbf{y} - H\boldsymbol{\beta}), \tag{2.18}$$

$$c^*(\mathbf{x}, \mathbf{x}') = c(\mathbf{x}, \mathbf{x}') - \mathbf{t}(\mathbf{x})^T A^{-1}\mathbf{t}(\mathbf{x}'), \tag{2.19}$$

$$\mathbf{t}(\mathbf{x})^T = (c(\mathbf{x}, \mathbf{x}_1), \ldots, c(\mathbf{x}, \mathbf{x}_n)), \tag{2.20}$$

$$\mathbf{y}^T = (\eta(\mathbf{x}_1), \ldots, \eta(\mathbf{x}_n)). \tag{2.21}$$

The next stage is to derive the distribution of $\eta(.)|\mathbf{y}$ unconditional on $\boldsymbol{\beta}$ and σ^2. If we consider the likelihood function of $\boldsymbol{\beta}$ and σ^2, we note that

$$f(\mathbf{y}|\boldsymbol{\beta},\sigma^2) = (2\pi\sigma^2)^{(-\frac{n}{2})} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-H\boldsymbol{\beta})^T A^{-1}(\mathbf{y}-H\boldsymbol{\beta})\right\}, \tag{2.22}$$

and that

$$(\mathbf{y}-H\boldsymbol{\beta})^T A^{-1}(\mathbf{y}-H\boldsymbol{\beta}) = (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T H^T A^{-1} H(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + (n-q-2)\hat{\sigma}^2, \tag{2.23}$$

where

$$\hat{\boldsymbol{\beta}} = (H^T A^{-1} H)^{-1} H^T A^{-1}\mathbf{y}, \tag{2.24}$$

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T(A^{-1} - A^{-1}H(H^T A^{-1}H)^{-1}H^T A^{-1})\mathbf{y}}{n-q-2}. \tag{2.25}$$

Then combining (2.8) with (2.14) using Bayes' theorem we see that $\boldsymbol{\beta}$ and σ^2 conditional on $\mathbf{y}$ have a normal inverse gamma distribution:

$$f(\boldsymbol{\beta},\sigma^2|\mathbf{y}) \propto \sigma^{2^{-\frac{n+2}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T H^T A^{-1} H(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + (n-q-2)\hat{\sigma}^2\right\}. \tag{2.26}$$

It can then be seen that

$$\boldsymbol{\beta}\,|\sigma^2,\mathbf{y} \sim N\left(\hat{\boldsymbol{\beta}}, \sigma^2(H^T A^{-1}H)^{-1}\right), \tag{2.27}$$

and

$$\sigma^2 \mid \mathbf{y} \sim (n-q-2)\hat{\sigma}^2\,\chi^{-2}_{n-q}, \tag{2.28}$$

To obtain the distribution of $\eta(.)|\mathbf{y},\sigma^2$, we combine (2.17) and (2.27), and then integrate out $\boldsymbol{\beta}$ to obtain

$$\eta(.)|\mathbf{y},\sigma^2 \sim N(m^{**}(.), \sigma^2 c^{**}(.,.)), \tag{2.29}$$

where

$$m^{**}(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T\hat{\boldsymbol{\beta}} + \mathbf{t}(\mathbf{x})^T A^{-1}(\mathbf{y}-H\hat{\boldsymbol{\beta}}), \tag{2.30}$$

$$c^{**}(\mathbf{x},\mathbf{x}) = c^*(\mathbf{x},\mathbf{x}) + (\mathbf{h}(\mathbf{x})^T - \mathbf{t}(\mathbf{x})^T A^{-1}H)(H^T A^{-1}H)^{-1} \tag{2.31}$$

$$\times(\mathbf{h}(\mathbf{x})^T - \mathbf{t}(\mathbf{x})^T A^{-1}H)^T. \tag{2.32}$$

Finally, by combining (2.28) with (2.29), and integrating out σ^2, we obtain the result that

$$\frac{\eta(\mathbf{x}) - m^{**}(\mathbf{x})}{\hat{\sigma}\sqrt{c^{**}(\mathbf{x},\mathbf{x})}} \sim t_{n-q}, \tag{2.33}$$

where t_{n-q} is a student t random variable with $n-q$ degrees of freedom.

We now have a quick approximation of $\eta(\mathbf{x})$ for any $\mathbf{x}$, since $m^{**}(\mathbf{x})$ does not include any terms involving $\eta(x)$. Note that unless a proper prior distribution is used for $\boldsymbol{\beta}$, the estimator $m^{**}(\mathbf{x})$ is the same as the estimator derived by Sacks et al. (1989). Thus from a frequentist perspective, $m^{**}(\mathbf{x})$ is the best linear unbiased predictor of $\eta(\mathbf{x})$. Examining (2.30), we can see that $m^{**}(\mathbf{x})$ consists of two components. The first component, $\mathbf{h}(.)^T\hat{\boldsymbol{\beta}}$ relates to our prior expectation of $\eta(.)$, which conditional on $\boldsymbol{\beta}$ is $\mathbf{h}(.)^T\boldsymbol{\beta}$. The expected value of $\boldsymbol{\beta}$ has been updated in light of the data $\mathbf{y}$. The second component, $\mathbf{t}(\mathbf{x})^T A^{-1}(\mathbf{y} - H\hat{\boldsymbol{\beta}})$ adjusts the posterior mean so that it passes through all the observed outputs; if we have observed $\eta(\mathbf{x}_i) = y_i$, then we have $m^{**}(\mathbf{x}_i) = y_i$. How smoothly $m^{**}(\mathbf{x})$ departs from $\mathbf{h}(\mathbf{x})^T\hat{\boldsymbol{\beta}}$ towards the observed output y_i for $\mathbf{x}$ close to $\mathbf{x}_i$ will depend on B.

The posterior covariance of $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ is given by $\hat{\sigma}^2 c^{**}(\mathbf{x},\mathbf{x}')$. If $\eta(\mathbf{x}_i) = y_i$ is known, then $c^{**}(\mathbf{x}_i,\mathbf{x}) = 0$, for all $\mathbf{x}$. For simplicity, we will write $c^{**}(\mathbf{x},\mathbf{x}) = c^{**}(\mathbf{x})$.

2.2.1 Estimating the smoothness parameters

The covariance function $c(.,.)$ contains a diagonal matrix B of parameters which describe how rough the function $\eta(.)$ is in each dimension of the input $\mathbf{x}$. Since the function $\eta(.)$ is unknown, the values of the elements in B will be unknown. Unfortunately, unlike the parameters $\boldsymbol{\beta}$ and σ^2, there is no analytical way of dealing with the uncertainty in B. The simplest option is to keep B fixed. Any single value of B will imply a particular distribution for the smoothness of $\eta(.)$. This will be apparent from generating realisations from the distribution of $\eta(.)$ conditional on B, and we illustrate this in chapter seven. Prior knowledge about the smoothness of $\eta(.)$ may suggest an appropriate value of B, and this is also discussed in chapter seven. Alternatively, we may wish to estimate B from the data. We consider two techniques.

Estimating B from the posterior mode

Haylock (1997) considers estimating B using the posterior mode. The density of $\mathbf{y}$ conditional on B, $\boldsymbol{\beta}$ and σ^2 is given by

$$f(\mathbf{y}|\boldsymbol{\beta},\sigma^2,B) = \frac{|A|^{-\frac{1}{2}}}{(\sigma^2)^{\frac{1}{2}n}(2\pi)^{\frac{n}{2}}}\exp\left[-(\mathbf{y}-H\boldsymbol{\beta})^T\frac{A^{-1}}{2\sigma^2}(\mathbf{y}-H\boldsymbol{\beta})\right], \tag{2.34}$$

which is the likelihood function for B, $\boldsymbol{\beta}$ and σ^2. If non-informative priors are used for $\boldsymbol{\beta}$ and σ^2, and an improper uniform prior is used for each element of B, then the posterior density of B, $\boldsymbol{\beta}$ and σ^2 is given by

$$f(\boldsymbol{\beta},\sigma^2,B|\mathbf{y}) = \frac{|A|^{-\frac{1}{2}}}{(\sigma^2)^{\frac{1}{2}(n+2)}(2\pi)^{\frac{p}{2}}}\exp\left[-(\mathbf{y}-H\boldsymbol{\beta})^T\frac{A^{-1}}{2\sigma^2}(\mathbf{y}-H\boldsymbol{\beta})\right]. \tag{2.35}$$

Integrating out $\boldsymbol{\beta}$ gives us

$$f(\sigma^2,B|\mathbf{y}) \propto \frac{|A|^{-\frac{1}{2}}|H^TA^{-1}H|^{-\frac{1}{2}}}{(\sigma^2)^{\frac{1}{2}(n+2-q)}}\exp\left[-(\mathbf{y}-H\hat{\boldsymbol{\beta}})^T\frac{A^{-1}}{2\sigma^2}(\mathbf{y}-H\hat{\boldsymbol{\beta}})\right], \tag{2.36}$$

and integrating out σ^2 gives

$$f(B|\mathbf{y}) \propto (\hat{\sigma}^2)^{-\frac{(n-q)}{2}}|A|^{-\frac{1}{2}}|H^TA^{-1}H|^{-\frac{1}{2}}, \tag{2.37}$$

recognising that (2.36) is proportional to an inverse gamma density function.

In certain cases there may be problems in estimating B by its posterior mode. We illustrate these with a simple one-dimensional example.

Consider the function

$$\eta(x) = 5 + x + \cos x, \tag{2.38}$$

and suppose that the true input X has a $N(0,4)$ distribution. We set $\mathbf{h}(x)^T = (1 \quad x)$ and evaluate $\eta(x)$ at five inputs: $(-4.334, -2.054, 0, 2.054, 4.334)$. The data are shown in figure 2.1.

We now treat $\eta(.)$ as an unknown function, and using these five observations only, we now wish to derive the posterior distribution of $\eta(.)$. We consider the correlation function

$$c(x,x') = \exp\{-b(x-x')^2\}. \tag{2.39}$$

We can think of two contrasting types of behaviour of $\eta(.)$ corresponding to two different values of b. Firstly, $\eta(.)$ may be highly correlated at values of x close to

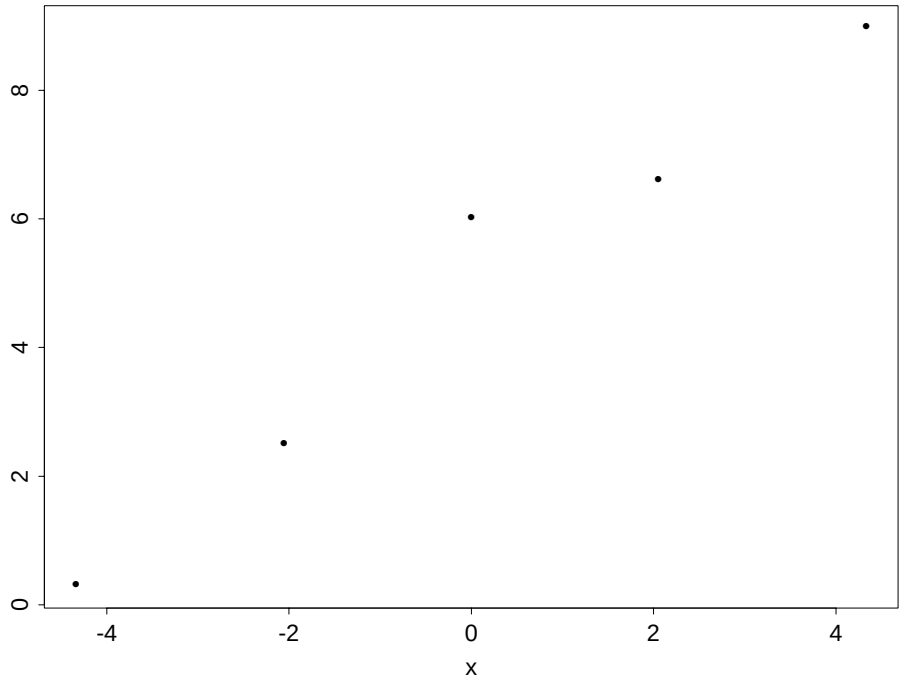

Figure 2.1: $\eta(x)$ evaluated at five inputs

each other, and so a small value of b is appropriate. The posterior mean $m^{**}(.)$ will deviate smoothly from $\mathbf{h}(x)^T\hat{\boldsymbol{\beta}}$ to the five actual observed outputs. Alternatively, there could be little or no correlation between outputs at neighbouring inputs. A linear form was chosen for $\mathbf{h}(.)$, and so $m^{**}(x)$ will follow a straight line with intercept and gradient given by the posterior mean of $\boldsymbol{\beta}$. The posterior mean of $\eta(.)$ will deviate sharply towards each of the observed outputs. An appropriate value of σ^2 will account for these deviations. In figures 2.2 and 2.3, we plot the posterior mean of $\eta(x)$ in the cases $b = 1.5$ and $b = 1000$. The true function is plotted as the dotted line. We can see that the two different values of b have resulted in noticeably different behaviour of the posterior mean in each case. However, with no prior information about b or σ^2, it is difficult to distinguish between these two cases. At $b = 1.5$ we have $f(b|\mathbf{y}) = 0.03207k$, and at $b = 1000$ we have $f(b|\mathbf{y}) = 0.03211k$ for some constant k. In fact, the posterior distribution for b is improper in this case, since (2.37) tends to a non zero constant as b tends to infinity.

Haylock (1997) considers using an arbitrary proper prior distibution for B, in

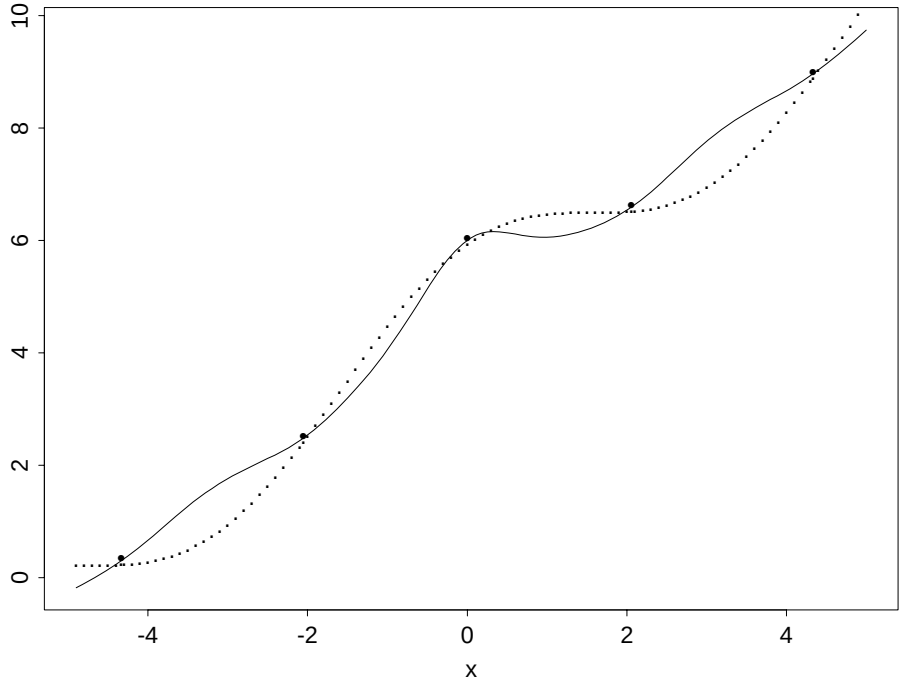

Figure 2.2: The posterior mean of $\eta(\mathbf{x})$ with $b = 1.5$

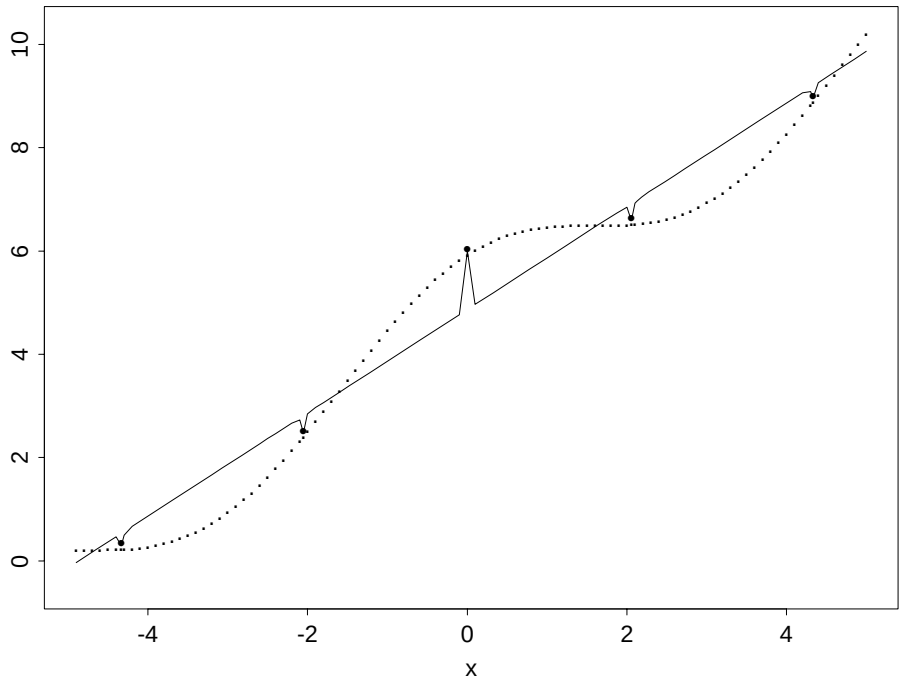

Figure 2.3: The posterior mean of $\eta(\mathbf{x})$ with $b = 1000$

the form of a lognormal distribution. Less support is given to extreme values of B, though the actual posterior mode may be sensitive to the choice of prior. In chapter seven we discuss how to derive a prior for B using expert knowledge about $\eta(.)$.

Estimating B using cross validation

Another option is to use a cross validation method to choose B. Given the data $\mathbf{y}$, we omit one observation $y_i = \eta(\mathbf{x}_i)$ to obtain a vector of observations denoted by $\mathbf{y}_{-i}$. For a given value of B we derive the posterior distribution of $\eta(.)$ given $\mathbf{y}_{-i}$, and determine d_i, the absolute distance between the posterior mean of $\eta(\mathbf{x}_i)$ and the known true value $y_i = \eta(\mathbf{x}_i)$. We carry out this process for $i = 1, \ldots, n$. and find B that minimises $\sum_{i=1}^{n} d_i$. For functions that deviate smoothly from the regression function, this procedure should lead to a sensible choice of B. In the one dimensional example, we can see in figures 2.2 and 2.3 that there is some difference in the predictive ability of $m^{**}(x)$ in the two cases $b = 1.5$ and $b = 1000$. Using the cross validation, the best estimate of b is 0.262, with $\sum_{i=1}^{5} d_i = 6.065$. If we set $b = 1000$, then we obtain $\sum_{i=1}^{5} d_i = 7.111$. In figure 2.4 we plot the function $m^{**}(x)$ and the true function (as a dotted line) with $b = 0.262$. In higher dimensional problems we have found this approach to work better than using the posterior mode. Since we are confining out attention to functions that are fairly smooth, we believe that this approach should usually give an appropriate value of B to use.

Note however that this method relies on the fact that $m^{**}(.)$ will be sensitive to the exact value of B. From this it is immediately obvious that by fixing B at a posterior estimate, we are not accounting for all the uncertainty that we have about $\eta(.)$.

2.2.2 Choice of $h(.)$

We use the one dimensional example to examine further the choice of $h(.)$. We consider two possible forms; $\mathbf{h}(x)^T = (1 \quad x)$ and $\mathbf{h}(x) = (1)$. We write $M1$ to refer to the prior model $E\{\eta(x)\} = \beta_{11}$, and $M2$ to denote the prior model $E\{\eta(x)\} = \beta_{21} + \beta_{22}x$. In figure (2.5) we show the posterior mean and a 95% interval for $\eta(\mathbf{x})$

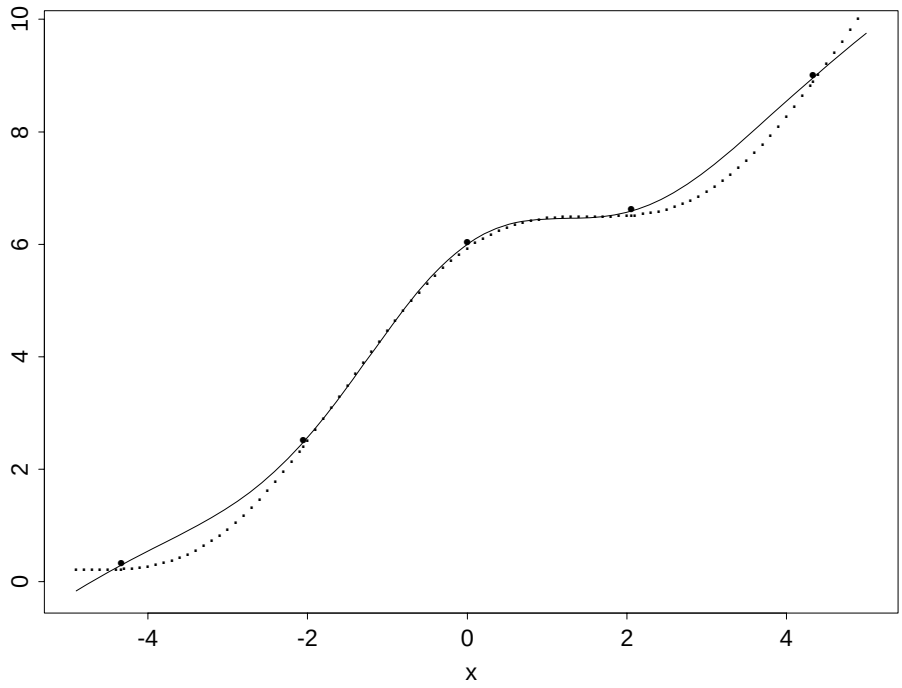

Figure 2.4: The posterior mean of $\eta(\mathbf{x})$ with b estimated using cross validation

after the five observations. Plot (a) refers to $M2$ and plot(b) refers to $M1$. With $h(x)^T = (1 \quad x)$ we have already noted that we have to use the cross validation method to find b, and so we use $b = 0.262$ in plot (a). With $\mathbf{h}(x) = (1)$, we find that the posterior distribution of b has a clear mode at $b = 0.061$, and we use this value in plot (b). Given that the true function is $\eta(x) = 5 + x + \cos x$, one might suppose that the choice $\mathbf{h}(x)^T = (1 \quad x)$ would give better results, but within the range $(-4.334, 4.334)$ where we have evaluated the function $\eta(.)$, this is visibly not the case.

The obvious explanation for the difference in this example is that with $\mathbf{h}(x) = (1)$, the resulting estimate of the smoothing is significantly smaller than 0.262, which has lead to a smaller posterior variance. However, even if we set $b = 0.061$ in the case $\mathbf{h}(x)^T = (1 \quad x)$, the results are still slightly better if we set $\mathbf{h}(x) = (1)$. In figure 2.6 we show the posterior mean and standard deviations of $\eta(x)$ for the two choices of $h(.)$ using $b = 0.061$ in each case. The solid line represents $\mathbf{h}(x)^T = (1 \quad x)$ and the dotted line represents $\mathbf{h}(x) = (1)$. Clearly, for x well outside the interval

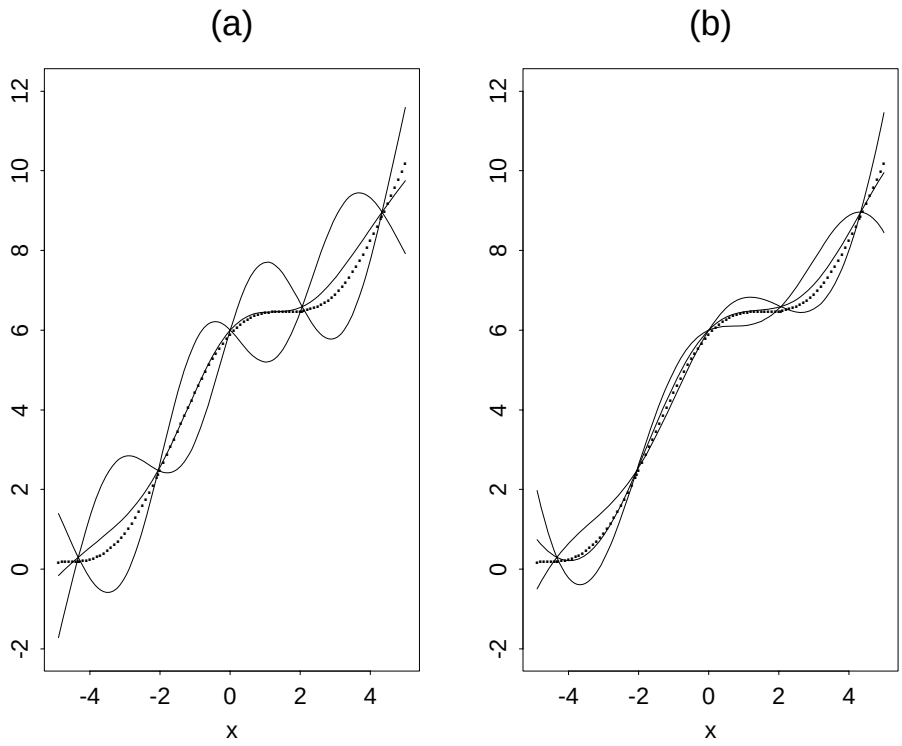

Figure 2.5: The posterior distribution of $\eta(x)$ for different choices of $h(x)$

$(-4.334, 4.334)$, the posterior mean using $M1$ will just be $\hat{\beta}_{11}$, and will become further away from $\eta(x)$ as $|x|$ increases. There is also a dependence on how well $\mathbf{h}(x)^T\hat{\boldsymbol{\beta}}$ fits the data. If for example, the design points had been chosen at values of x where $\cos x = 0$, then the posterior mean of σ^2 under $M1$ would be zero. This would result in the posterior variance of $\eta(x)$ being zero everywhere, though the posterior mean of $\eta(.)$ would be a straight line and not the true function.

Examining the function $m^{**}(\mathbf{x})$, we note that depending on b, the inclusion of a linear term in $h(.)$ may have little effect on $m^{**}(\mathbf{x})$ within the range of the design points. For design points $x_1, \ldots, x_n$ we can re-write $m^{**}(x)$ as

$$m^{**}(x) = \mathbf{h}(x)^T\hat{\boldsymbol{\beta}} + \sum_{i=1}^{n} r_i(x)\{\eta(x_i) - \mathbf{h}(x_i)^T\hat{\boldsymbol{\beta}}\}, \tag{2.40}$$

where $r_i(x)$ is the i-th element of $\mathbf{t}(x)^T A^{-1}$, so that (in the case of a weak prior distribution) the posterior mean is equal to the least squares fit of $\mathbf{h}(x)^T\boldsymbol{\beta}$, modified by a linear combination of the residuals. For simplicity, suppose that the design points are evenly spaced out and arranged in ascending order, and that if $x \in$

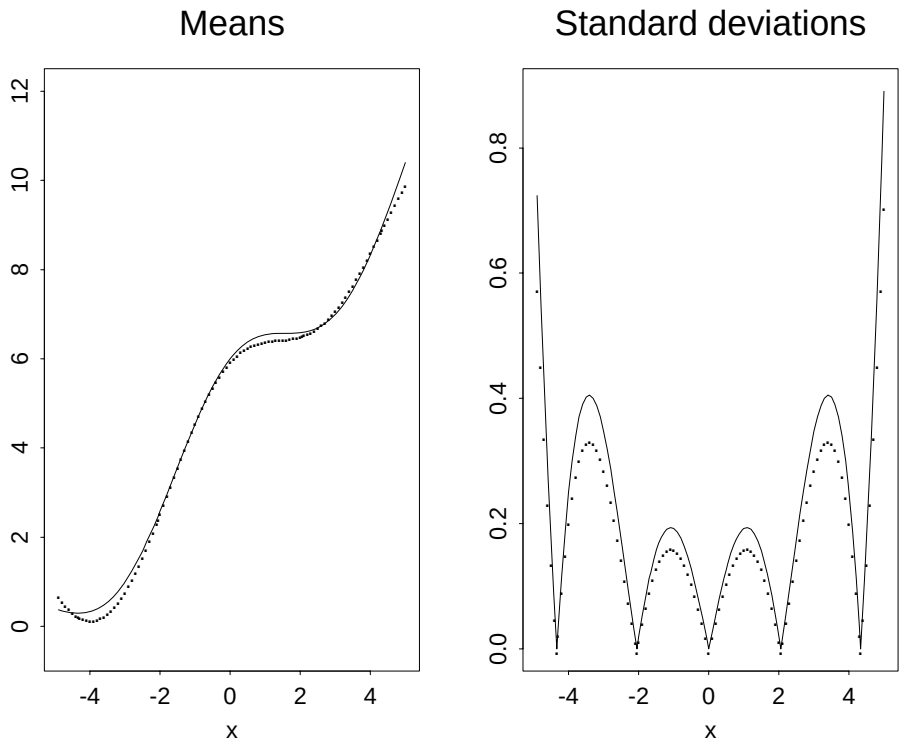

Figure 2.6: The posterior means and standard deviations of $\eta(x)$ for different choices of $h(x)$ with $b = 0.061$

(x_j, x_{j+1}) for $j = 1, \ldots, n-1$, then $r_i(x) = 0$ for $i \neq j$ or $i \neq j+1$. Now consider the posterior mean under $M1$, which we denote by $m_1^{**}(\mathbf{x})$. We have for $x \in (x_j, x_{j+1})$

$$m_1^{**}(x) = \hat{\beta}_{11} + r_j(x)\{\eta(x_j) - \hat{\beta}_{11}\} + r_{j+1}(x)\{\eta(x_{j+1}) - \hat{\beta}_{11}\}. \tag{2.41}$$

Defining $m_2^{**}(x)$ to be the posterior mean under $M2$, we have for $x \in (x_j, x_{j+1})$

$$m_2^{**}(x) = \hat{\beta}_{21} + \hat{\beta}_{22}x + r_j(x)\{\eta(x_j) - \hat{\beta}_{21} - \hat{\beta}_{22}x_j\} + r_{j+1}(x)\{\eta(x_{j+1}) - \hat{\beta}_{11} - \hat{\beta}_{22}x_{j+1}\}. \tag{2.42}$$

Since $t(x_j)$ is the $j-$th row of A, we have $r_i(x_j) = 1$ if $i = j$ and $r_i(x_j) = 0$ if $i \neq j$. Now suppose we allow $r_j(x)$ and $r_{j+1}(x)$ to be any two functions that satisfy this condition. Then if

$$r_j(x) = \frac{x - x_{j+1}}{x_j - x_{j+1}}, \tag{2.43}$$

and

$$r_{j+1}(x) = \frac{x - x_j}{x_{j+1} - x_j}, \tag{2.44}$$

then for $x \in (x_j, x_{j+1})$ it follows that (2.41) and (2.42) are equal.

In figure 2.7 we plot the five functions $r_i(x)$ for $i = 1, \ldots, 5$ in the one dimensional example as dotted lines. If, for $x \in (-4.334, 4.334)$, these functions are either zero, or equal to the staight lines shown on the plot, then including a linear term in x will have no effect on the posterior mean. From this graph we would expect the posterior means under $M1$ and $M2$ to be particularly close to each other for $x \in (-2.054, 2.054)$, and this can be seen in figure 2.6.

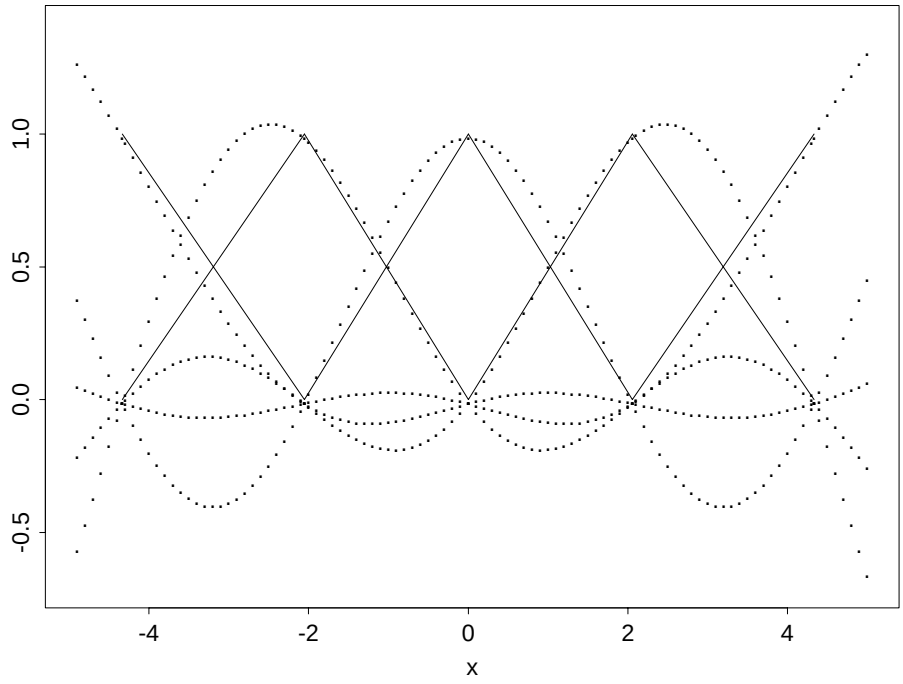

Figure 2.7: The functions $r_i(x)$ for $i = 1, \ldots, 5$, shown as dotted lines

The larger standard deviation resulting from $h(x) = (1 \quad x)^T$ is more suprising. Firstly, note that under $M2$, we have introduced an additional parameter β_{22} into the model, and there will be additional uncertainty in the distribution of $\eta(x)$ resulting from uncertainty about β_{22}. Under $M1$ we would expect the residuals $H\hat{\boldsymbol{\beta}} - \mathbf{y}$ to be larger, and so the posterior mean of σ^2 should also be larger. In this example we have five observations, and so σ^2 has four degrees of freedom under $M1$ and three degrees of freedom under $M2$. It is this difference in the degrees of freedom which results in $E(\sigma^2|\mathbf{y})$ being larger under $M2$, even though the residuals themselves are

smaller. If further observations are taken, the difference in degrees of freedom will have less effect. We now observe $\eta(.)$ at a further five inputs, and plot the posterior means and standard deviations in figure 2.8. As before, the dotted line represents $M1$ and the solid line represents $M2$. We can now see that the standard deviation is smaller with $M2$.

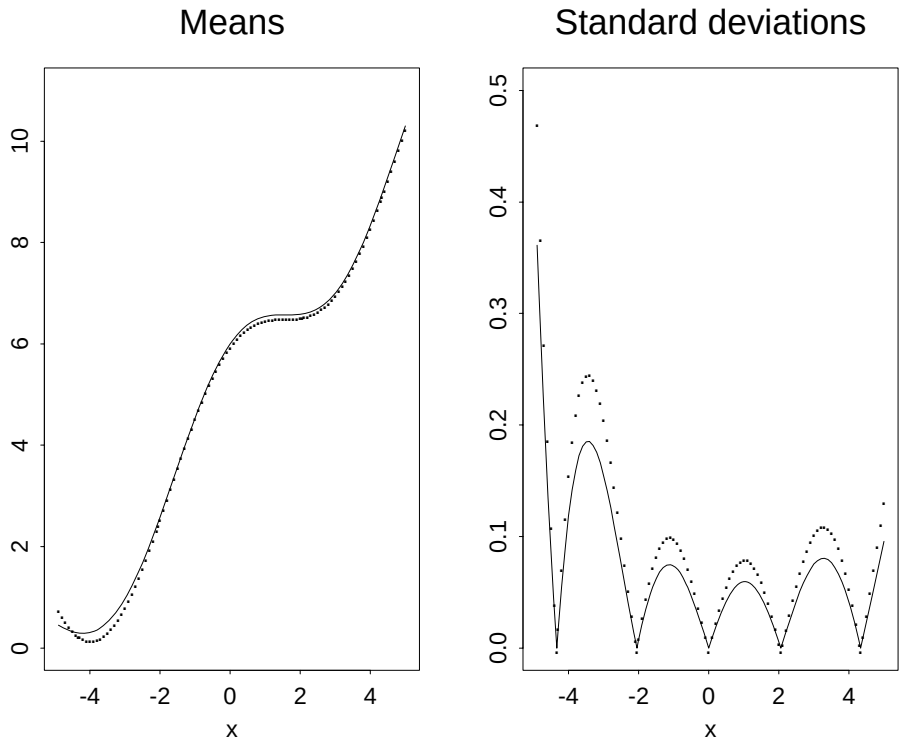

Figure 2.8: The posterior means and standard deviations of $\eta(x)$ for different choices of $h(x)$ with ten observations

This is an area for further investigation. In some cases we may have strong prior beliefs to guide the choice of $\mathbf{h}(.)$. However, we have seen that the posterior variance of $\eta(.)$ can be sensitive to $\mathbf{h}(.)$, and so prior uncertainty about an appropriate form for $\mathbf{h}(.)$ can induce uncertainty in the posterior distribution of $\eta(.)$. In the absence of prior knowledge about $\mathbf{h}(.)$ we should attempt to account for this resulting uncertainty in the posterior.

2.2.3 Choice of design points

Criterion based design

We need to consider what values of the inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ we should run the model at. One approach is to define some criterion that describes what a good design is, and then find the design that best satisfies that criterion. In this context, we are trying to estimate the value of the output of the code when run at the true input $\mathbf{X}$, where $\mathbf{X}$ is unknown and has distribution $G(\mathbf{x})$. Haylock (1997) considers the following loss function for a design D:

$$L\{D, \mathbf{y}, \mathbf{X}, \eta(\mathbf{X})\} = \{m^{**}(\mathbf{X}) - \eta(\mathbf{X})\}^2, \tag{2.45}$$

i.e. we choose design points $D = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, obtain data $y_1 = \eta(\mathbf{x}_1)$,. . .,$y_n = \eta(\mathbf{x}_n)$, and compare the distance between $\eta(\mathbf{X})$ and the posterior mean of $\eta(\mathbf{X})$. Since $\mathbf{X}, \eta(\mathbf{X})$ and $\mathbf{y}$ are unknown, a design is chosen to minimise the expected loss. We first take the expectation of (2.45) with respect to $\mathbf{X}$:

$$L\{D, \mathbf{y}, \eta(\mathbf{X})\} = \int_{\mathcal{X}} \{m^{**}(\mathbf{x}) - \eta(\mathbf{x})\}^2 \, dG(\mathbf{x}). \tag{2.46}$$

We have used the notation $E_A[L\{A, B\}] = L\{B\}$. We now take the expectation of (2.46) with respect to $\eta(.)$:

$$\begin{aligned} L\{D, \mathbf{y}\} &= \int_{\mathcal{X}} E[\{m^{**}(\mathbf{x}) - \eta(\mathbf{x})\}^2 | D, \mathbf{y}] \, dG(\mathbf{x}) && (2.47) \\ &= \int_{\mathcal{X}} Var\{\eta(\mathbf{x}) | D, \mathbf{y}\} dG(\mathbf{x}) && (2.48) \\ &= E(\sigma^2 | D, \mathbf{y}) \int_{\mathcal{X}} c^{**}(\mathbf{x}) dG(\mathbf{x}). && (2.49) \end{aligned}$$

The next step is to remove the conditioning on $\mathbf{y}$. Note that $c^{**}(\mathbf{x})$ is not a function of $\mathbf{y}$, and

$$E_{\mathbf{y}}\{E(\sigma^2 | D, \mathbf{y}) | D\} = E(\sigma^2 | D) \tag{2.50}$$

Finally, Haylock(1997) states that

$$E(\sigma^2 | D) \propto E(\sigma^2), \tag{2.51}$$

since knowledge of the position of the design points alone gives no information about the value of σ^2. Hence

$$L(D) \propto \int_{\mathcal{X}} c^{**}(\mathbf{x}) \, dG(\mathbf{x}). \tag{2.52}$$

This can be minimised by finding D that will minimise

$$\int_{\mathcal{X}} \left\{(\mathbf{h}(\mathbf{x})^T - \mathbf{t}(\mathbf{x})^T A^{-1}H)(H^T A^{-1}H)^{-1}(\mathbf{h}(\mathbf{x})^T - \mathbf{t}(\mathbf{x})^T A^{-1}H)^T - \mathbf{t}(\mathbf{x})^T A^{-1}\mathbf{t}(\mathbf{x})\right\} dG(\mathbf{x}). \quad (2.53)$$

In certain cases this integral can be solved analytically. We consider a d dimensional input $\mathbf{x} = (x_1, \ldots, x_d)^T$, where the elements of the true input $\mathbf{X}$ are independent and have normal distributions. For simplicity we consider standardised inputs, so that $X_i \sim N(0,1)$. We denote the sample space of X_i to be $\mathcal{X}_i$. We use the covariance function $c(\mathbf{x}, \mathbf{x}') = \exp\{(\mathbf{x} - \mathbf{x}')^T B(\mathbf{x} - \mathbf{x}')\}$, where B is a diagonal matrix whose i,i-th element is b_i. We set $h(x)^T = (1\ x_1\ x_2\ \ldots\ x_d)$. The design points are denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i = (x_{i_1}, \ldots, x_{i_d})^T$. We first note that

$$\begin{aligned} \int_{\mathcal{X}} \mathbf{t}(x)^T A^{-1}\mathbf{t}(x) dG(x) &= tr\left\{A^{-1}\int_{\mathcal{X}} \mathbf{t}(\mathbf{x})\mathbf{t}(\mathbf{x})^T dG(\mathbf{x})\right\} & (2.54)\\ &= tr(A^{-1}P) & (2.55)\end{aligned}$$

where the $j,k-$th element of P is given by

$$\begin{aligned} P[j,k] &= \prod_{k=1}^{d} \int_{\mathcal{X}_k} \frac{1}{\sqrt{2\pi}} \exp\left[-b_k\left\{(x_k - x_{i_k})^2 + (x_k - x_{j_k})^2\right\} - \frac{x_k^2}{2}\right] dx_k & (2.56)\\ &= \prod_{k=1}^{d} \int_{\mathcal{X}_k} \frac{1}{\sqrt{2\pi}} \exp\left\{\left(2b + \frac{1}{2}\right)\left(x_k - \frac{b(x_{i_k} + x_{j_k})}{2b + \frac{1}{2}}^2\right)\right. \\ &\quad \left. -b(x_{i_k}^2 + x_{j_k}^2) + \frac{b^2(x_{i_k} + x_{j_k})^2}{2b + \frac{1}{2}}\right\} dx_k & (2.57)\\ &= \prod_{k=1}^{d} \frac{1}{4b+1} \exp\left\{-b(x_{i_k}^2 + x_{j_k}^2) + \frac{b^2(x_{i_k} + x_{j_k})^2}{2b + \frac{1}{2}}\right\}. & (2.58)\end{aligned}$$

We also have

$$\begin{aligned} &(\mathbf{h}(\mathbf{x})^T - \mathbf{t}(\mathbf{x})^T A^{-1}H)(H^T A^{-1}H)^{-1}(\mathbf{h}(\mathbf{x})^T - \mathbf{t}(\mathbf{x})^T A^{-1}H)^T \\ &= tr(H^T A^{-1}H)^{-1}(\mathbf{h}(\mathbf{x})^T - \mathbf{t}(\mathbf{x})^T A^{-1}H)^T(\mathbf{h}(\mathbf{x})^T - \mathbf{t}(\mathbf{x})^T A^{-1}H). \quad (2.59)\end{aligned}$$

Now we consider $\int_{\mathcal{X}} \mathbf{h}(x)\mathbf{t}(x)^T dG(x)$. We need to evaluate $\int_{\mathcal{X}} c(\mathbf{x}, \mathbf{x}') dG(\mathbf{x})$ and $\int_{\mathcal{X}} x_i c(\mathbf{x}, \mathbf{x}') dG(\mathbf{x})$. Firstly,

$$\int_{\mathcal{X}} c(\mathbf{x}, \mathbf{x}_j) dG(\mathbf{x}) = \prod_{k=1}^{d} \int_{\mathcal{X}_k} \frac{1}{\sqrt{2\pi}} \exp\{-b_k(x_k - x_{j_k})^2\} \exp(-\frac{1}{2}x_k^2) dx_k$$

$$= \prod_{k=1}^{d} \int_{\mathcal{X}_k} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\left(b_k + \frac{1}{2}\right)\left(x_k - \frac{b_k x_{j_k}}{b_k + \frac{1}{2}}\right)^2 - b x_{j_k}^2 + \frac{b_k^2 x_{j_k}^2}{b_k + \frac{1}{2}} \right\} dx_k$$

$$= \prod_{k=1}^{d} \frac{1}{\sqrt{2b_k + 1}} \exp\left\{ -b_k x_{j_k}^2 \left(\frac{1}{2b_k + 1}\right) \right\}, \tag{2.60}$$

and

$$\int_{\mathcal{X}} x_i c(\mathbf{x}, \mathbf{x}_j) dG(\mathbf{x}) = \prod_{k=1}^{d} \int_{\mathcal{X}k} x_i \frac{1}{\sqrt{2\pi}} \exp\{-b_k (x_k - x_{j_k})^2\} \exp(-\frac{1}{2} x_k^2) dx_k$$

$$= \frac{2b_i x_{j_i}}{2b_i + 1} \prod_{k=1}^{d} \frac{1}{\sqrt{2b_k + 1}} \exp\left\{ -b_k x_{j_k}^2 \left(\frac{1}{2b_k + 1}\right) \right\}. \tag{2.61}$$

Finally, we have

$$\int_{\mathcal{X}} \mathbf{h}(x)\mathbf{h}(x)^T dG(x) = I_{d+1}, \tag{2.62}$$

since

$$\int_{\mathcal{X}} x_i^2 dG(\mathbf{x}) = 1, \tag{2.63}$$

and

$$\int_{\mathcal{X}} x_i x_j dG(\mathbf{x}) = 0, \tag{2.64}$$

as X_i and X_j are independent. Combining these results we obtain

$$\int_{\mathcal{X}} c^{**}(x) dG(x) = 1 + tr(R), \tag{2.65}$$

where

$$R = A^{-1}P + (H^T A^{-1} H)^{-1} \left\{ H^T A^{-1} P A^{-1} H + S A^{-1} H + H^T A^{-1} S^T + I_{d+1} \right\}, \tag{2.66}$$

and the i, j-th element of S is given by (2.60) for $i = 1$ and (2.61) for $i = 2, \ldots, d+1$.

Clearly, we will need to assign a value to the smoothing parameter, even though we have no data. A consequence of this is that no design can be thought of as 'optimal'. To illustrate this, we consider a one dimensional design of the form $(-x, 0, x)$ where the input has a standard normal distribution. For different values of the smoothing parameter b, we find x that minimises (2.52). The results are given in the following table: Thus it can be seen that the 'optimal' design varies to a small degree according to the assumed value of b. Sacks et al. (1989a) perform a robustness study on b. They examine the effect of assuming a value b_1 on the design

Table 2.1: Optimal designs for various values of b

b	0.25	0.5	1	2	5
x	1.566	1.467	1.357	1.262	1.299

when the true value of b is b_2. For an assumed value b_1, they first find the optimal design, D_{b_1}. The criterion they use is the integrated mean squared error for their estimator $\hat{\eta}(x)$, as a function of the design D and the smoothing parameter b:

$$J(D,b) = \int_{\mathcal{X}} E[\hat{\eta}(x) - \eta(x)]^2 \, dx. \tag{2.67}$$

Note that their interest is simply in estimating the value of $\eta(x)$ at untried inputs, within some range of inputs $\mathcal{X}$, and so no distribution for x is assumed. However, obvious parallels can be seen between (2.67) and (2.52). Once the design has been selected, they then evaluate (2.67) assuming that $b = b_2$, obtaining $J(D_{b_1}, b_2)$. They then determine the optimum design according to their criterion when the true value b_2 is assumed, with this design denoted by D_{b_2}. The ratio $J(D_{b_2}, b_2)/J(D_{b_1}, b_2)$ is computed. A value of the ratio close to 1 indicates that the assumed value b_1 is robust if the true value is b_2. This was carried out for values of b_1 and b_2 ranging up to 100. It was found that assuming a value of 1 gave the most robust design.

We repeat the study here, using the criterion (2.53) in place of (2.67). We assume $b = b_1$ and find a design of the form $(-x, 0, x)$ to minimise (2.53). The results are summarised in table 2.2. This suggests that assuming a value of b in the range (0.5,1) will give a robust design. Observe that when the true value of b is 100, assuming a small value of b such as 0.25 results in a better performance than when assuming a value slighter closer to 100 such as 5. When b is small, the optimum design points are positioned further apart due to the higher correlations between them. When b is very large, the optimum design points are also positioned further apart, for the following reason.

When b is sufficiently large, we have $c(x, x') \simeq 0$ for $x \neq x'$. Recall that we are finding design points to minimise

$$\int_{\mathcal{X}} c^{**}(x) dG(x) \propto \int_{\mathcal{X}} Var\{\eta(x)|\mathbf{y}, \sigma^2\} dG(x), \tag{2.68}$$

Table 2.2: Robustness study for choice of b

		b_1					
		0.25	0.5	1	2	5	100
	0.25	1	0.974	0.892	0.796	0.761	0.135
	0.5	0.976	1	0.967	0.890	0.856	0.246
b_2	1	0.921	0.975	1	0.978	0.959	0.393
	2	0.896	0.944	0.986	1	0.998	0.571
	5	0.945	0.968	0.989	0.999	1	0.821
	100	0.906	0.895	0.870	0.862	0.855	1

where

$$\begin{aligned} Var\{\eta(x)|\mathbf{y},\sigma^2\} &= (\mathbf{h}(x)^T - \mathbf{t}(x)^T A^{-1}H)Var(\boldsymbol{\beta}|\mathbf{y},\sigma^2)(\mathbf{h}(x)^T - \mathbf{t}(x)^T A^{-1}H)^T \\ &\quad +\sigma^2 \mathbf{t}(x)^T A^{-1}\mathbf{t}(x). \end{aligned} \tag{2.69}$$

Now if b is sufficiently large, then A is the identity matrix and $\mathbf{t}(x)$ is a vector of zeros if x is not a design point. Then $Var\{\eta(x)|\mathbf{y},\sigma^2\}$ reduces to $\mathbf{h}(x)^T Var(\boldsymbol{\beta}|\mathbf{y},\sigma^2)\mathbf{h}(x)$ almost everywhere, and minimising (2.53) can be achieved by minimising the individual elements of $Var(\boldsymbol{\beta}|\mathbf{y},\sigma^2)$. Since

$$Var(\boldsymbol{\beta}|\mathbf{y},\sigma^2) = \sigma^2(H^T A^{-1}H)^{-1} \tag{2.70}$$

$$= \sigma^2 \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1}, \tag{2.71}$$

when b is large and $\mathbf{h}(x)^T = (1 \quad x)$, the posterior variance of $\boldsymbol{\beta}$ can be made small by choosing symetrical design points about zero, and setting the absolute value of each design point to be large. Thus when the optimal design is of the form $(-x, 0, x)$ we have $x \to \infty$ as $b \to \infty$. A second complication is that finding design points that will maximise (2.53) can itself be a computationally intensive procedure, as reported in Haylock (1997).

Product designs and Kronecker products

A general simplification using Kronecker products is given in O'Hagan (1991), in the case of product designs. First note that for any two matrices M_1 and M_2, if M_1^{-1} and M_2^{-1} both exist then

$$(M_1 \bigotimes M_2)^{-1} = M_1^{-1} \bigotimes M_2^{-1}, \tag{2.72}$$

where $M_1 \otimes M_2$ is the Kronecker product of M_1 and M_2. (See Lancaster and Tismenetsky, 1985).

Now consider a d dimensional input $\mathbf{x} = (x_1, \ldots, x_d)$, and suppose that the correlation function satisfies

$$c(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{d} c_i(x_i, x_i'), \tag{2.73}$$

so that the correlation between two d dimensional inputs can be expressed as the product of d correlations between one dimensional inputs. Now suppose that we choose the design points in a product form. That is, we choose n_i points in dimension i of the input space, and then use each possible combination of points to obtain $n_1 \times n_2 \times \ldots \times n_d$ design points. Then if A_i is the variance covariance matrix of the points $(x_{i_1}, \ldots, x_{i_{n_i}})$, we have the result

$$A^{-1} = A_1^{-1} \bigotimes A_2^{-1} \bigotimes \ldots \bigotimes A_d^{-1} \tag{2.74}$$

Thus the task of inverting the potentially large matrix A can be broken down to inverting several smaller matrices.

Randomised design using Latin hypercube sampling

Without prior information about B, we might simply consider a design that represents the sample space $\mathcal{X}$ without taking into account correlations between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$. McKay et al (1979) propose the use of Latin hypercube sampling. Suppose $\mathbf{x} = (x_1, \ldots, x_d)$ and we wish to draw n random values of $\mathbf{x}$. For $i = 1, \ldots, d$ we divide the sample space of x_i into n regions of equal marginal probability. We then draw one random value of x_i from each region. Then to obtain one random value of $\mathbf{x}$, we sample without replacement from the values $x_{i_1}, \ldots, x_{i_n}$ for $i = 1, \ldots, d$. This

ensures that each dimension in the input space is fully represented. A refinement to this procedure is given in Mitchell and Morris (1995). Generate one Latin hypercube sample, find the two inputs that are closest together, and then record the distance s, say, between them. Then generate a large number of samples, and choose the one that has the largest value of s. This should give design points that are well spaced out, while still ensuring that each dimension has its sample space well represented. An advantage of this approach are that when finding a distribution to represent the expert's beliefs about $\mathbf{X}$, any distribution can be used for $G(\mathbf{x})$, as long as it can be sampled from. This procedure is also computationally cheap.

Sequential design

We might also consider using both the previously mentioned ideas sequentially. An initial set of design points could be chosen using the Latin hypercube scheme. The data obtained from running the code at these inputs could then be used to estimate the smoothing parameters. Given an estimate of B, we could then choose further design points using the criterion based approach.

Alternative approaches

Currin et al. (1991) used the concept of entropy in selecting a design. Entropy is a means of quantifying the "amount of information", or alternatively, the "amount of uncertainty" in an experiment, as defined by Shannon (1948). Lindley (1956) advocated choosing a design that would maximise the expected reduction in posterior entropy. Currin et al. (1991) were concerned with predicting $\eta(.)$ over some finite region of inputs S. Thus a design D is needed to maximise the reduction in entropy of $\eta(.)$ over the region $S - D$. Shewry and Wynn (1987) showed that this is equivalent to finding a design D such that the entropy of $\eta(.)$ in D is maximised. This can be done by maximising the determinant of the correlation matrix A. However, this procedure will also require a value of B to be assumed.

O'Hagan (1991) used the gaussian process model to estimate integrals of expensive functions $\eta(.)$. Designs are chosen to minimise the variance of the integral. This can be useful in uncertainty analysis if we wish to estimate the expected value of Y,

given by

$$\int_{\mathcal{X}} \eta(\mathbf{x}) dG(\mathbf{x}). \tag{2.75}$$

Suggested designs are tabulated in O'Hagan (1991).

2.3 The Gaussian Process model for derivatives of functions

O'Hagan (1992) shows how Gaussian processes can also be used to model the derivatives of an unknown function $\eta(.)$. If $\eta(\mathbf{x})$ has a normal distribution, and we define $\mathbf{x} = (x_1, \ldots, x_k)$, then

$$\frac{\partial}{\partial x_i}\eta(\mathbf{x}) = \lim_{u \to 0} \frac{1}{u}\eta(\mathbf{x} + (0, \ldots, 0, u, 0, \ldots, 0)) - \eta(\mathbf{x}) \tag{2.76}$$

also has a normal distribution, providing the limit exists. Consequently all derivatives of $\eta(.)$ have a joint distribution described by a Gaussian process. If

$$\begin{aligned} E\{\eta(\mathbf{x})\} &= \mathbf{h}^T(\mathbf{x})\boldsymbol{\beta} & (2.77)\\ Cov\{\eta(\mathbf{x}), \eta(\mathbf{x}')\} &= c(\mathbf{x}, \mathbf{x}'), & (2.78) \end{aligned}$$

then

$$\begin{aligned} E\left\{\frac{\partial^n}{\partial x_i^n}\eta(\mathbf{x})\right\} &= \frac{\partial^n}{\partial x_i^n}\mathbf{h}^T(\mathbf{x})\boldsymbol{\beta} & (2.79)\\ Cov\left\{\frac{\partial^n}{\partial x_i^n}\eta(\mathbf{x}), \frac{\partial^m}{\partial x_j'^m}\eta(\mathbf{x}')\right\} &= \frac{\partial^{n+m}}{\partial x_i^n \partial x_j'^m}c(\mathbf{x}, \mathbf{x}'), & (2.80) \end{aligned}$$

provided both $\mathbf{h}(.)$ and $c(.,.)$ are differentiable. We can then derive the posterior distribution of $\frac{\partial^r}{\partial x_i^r}\eta(\mathbf{x})|\mathbf{y}$ to obtain

$$\frac{\frac{\partial^r}{\partial x_i^r}\eta(\mathbf{x}) - m^{**}_{r,x_i}(\mathbf{x})}{\hat{\sigma}\sqrt{c^{**}_{r,x_i}(\mathbf{x}, \mathbf{x})}} \sim t_{n-q}, \tag{2.81}$$

where

$$\begin{aligned} m^{**}_{r,x_i}(\mathbf{x}) &= \frac{\partial^r}{\partial x_i^r}\mathbf{h}(\mathbf{x})^T\hat{\boldsymbol{\beta}} + \frac{\partial^r}{\partial x_i^r}\mathbf{t}(\mathbf{x})^T A^{-1}(\mathbf{y} - H\hat{\boldsymbol{\beta}}), & (2.82)\\ c^{**}_{r,x_i}(\mathbf{x}, \mathbf{x}) &= 1 - \left\{\frac{\partial^r}{\partial x_i^r}\mathbf{t}(\mathbf{x})^T\right\} A^{-1} \left\{\frac{\partial^r}{\partial x_i^r}\mathbf{t}(\mathbf{x})\right\} \end{aligned}$$

$$\times \left[\left\{ \frac{\partial^r}{\partial x_i^r} \mathbf{h}(\mathbf{x})^T \right\} - \left\{ \frac{\partial^r}{\partial x_i^r} \mathbf{t}(\mathbf{x})^T \right\} A^{-1} H \right] (H^T A^{-1} H)^{-1}$$
$$\times \left[\left\{ \frac{\partial^r}{\partial x_i^r} \mathbf{h}(\mathbf{x})^T \right\} - \left\{ \frac{\partial^r}{\partial x_i^r} \mathbf{t}(\mathbf{x})^T \right\} A^{-1} H \right]^T \quad (2.83)$$

Note that if the correlation function is of the form

$$c(\mathbf{x}, \mathbf{x}') = \prod \exp(-\theta_i |x_i - x_i'|^p), \quad (2.84)$$

with $p = 2$, and if $\mathbf{h}(.)$ is infinitely differentiable then we can see that the posterior mean and variance functions are also infinitely differentiable, and the distributions of all the derivatives of $\eta(.)$ are well defined.

2.3.1 The use of derivatives in Bayes-Hermite quadrature

O'Hagan (1991) uses the Gaussian process model to make inferences about the integral of an unknown function

$$k = \int \eta(\mathbf{x})\, d\mathbf{x}. \quad (2.85)$$

This is of interest in the uncertainty analysis context when considering summaries of Y such as its mean:

$$E\{Y = \eta(\mathbf{X})\} = \int_{\mathcal{X}} \eta(\mathbf{x}) dG(\mathbf{x}), \quad (2.86)$$

as can be seen in Haylock and O'Hagan (1996). We now give a simple example to explore the value of observing derivatives in quadrature.

Suppose we wish to estimate the integral $k = \int \eta(x)\, dG(x)$, for a one dimensional $\eta(x)$. We set $h(x) = (1)$ and $c(x, x') = \exp\left\{-\frac{1}{2}(x - x')^2\right\}$. We assume that $x \sim N(0, 1)$. For this choice of $h(.)$ and a weak prior ditribution of β, σ^2, we need a minimum of four observations. We compare the variance of the estimate of k when observing $\mathbf{y}_1^T = \{\eta(-2), \eta(2), \eta(-x), \eta(x)\}$ with the variance when observing $\mathbf{y}_2^T = \{\eta(-2), \eta(2), \eta(x), \eta'(x)\}$. In O'Hagan (1991) the variance of $k \mid \mathbf{y}$ is given by

$$var(k \mid \mathbf{y}) = U - TA^{-1}T^T + (R - TA^{-1}H)(H^T A^{-1} H)(R - TA^{-1}H)^T$$

where

$$\begin{aligned} U &= \int_X \int_X c(x,x')dG(x)\,dG(x') \\ T &= \int_X t(x)^T dG(x) \\ R &= (1) \end{aligned}$$

In figure 2.9 is a plot of $v1 = var(k \mid \mathbf{y}_1)$ and $v2 = var(k \mid \mathbf{y}_2)$ for different x, with the dotted line indicating the variance when the derivative is observed. The

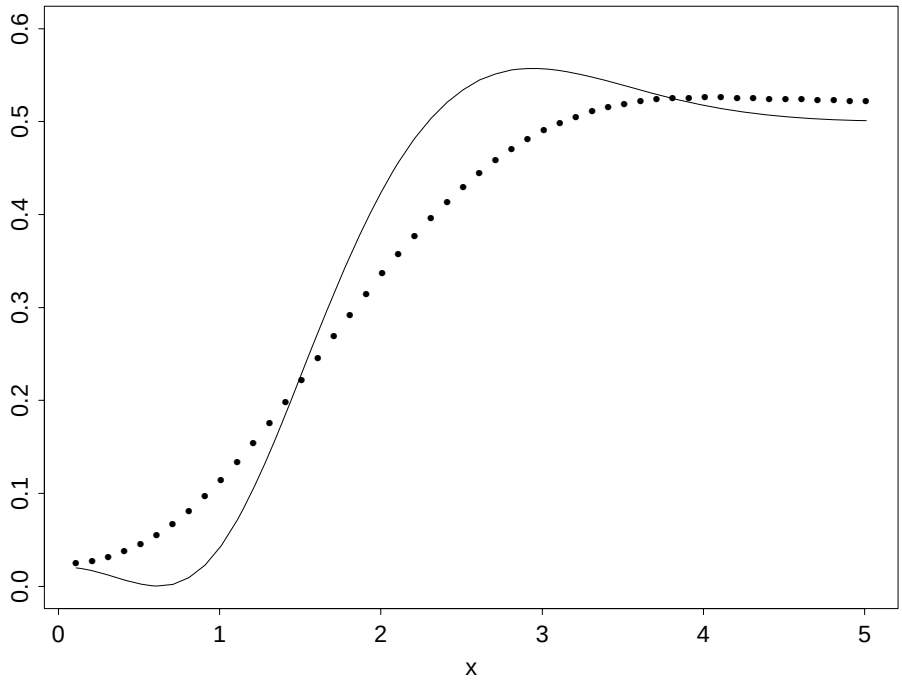

Figure 2.9: Investigating observing derivatives in quadrature

graph suggests that in the limit as x tends to zero, observing $\eta(x)$ and $\eta(-x)$ is equivalent to observing $\eta(x)$ and $\eta'(x)$. We observe that for certain values of x, we obtain a smaller variance if $\eta'(x)$ is observed rather than $\eta(-x)$. Note that the correlation between $\eta'(x)$ and $\eta(y)$is given by

$$\sqrt{2b}(y-x)\exp\{-b(x-y)^2\}, \tag{2.87}$$

and so if $|\sqrt{2b}(y-x)| > 1$ then this will be higher than the correlation between $\eta(x)$ and $\eta(y)$. Returning to the example, consider the case $\mathbf{y}_1^T = \{\eta(-2), \eta(2), \eta(-1.8), \eta(1.8)\}$

and $\mathbf{y}_2^T = \{\eta(-2), \eta(2), \eta(1.8), \eta'(1.8)\}$. In figure 2.10 we plot the standard deviation of $\eta(x)|\mathbf{y}_1$ as the solid line, and the standard deviation of $\eta(x)|\mathbf{y}_2$ as the dotted line, both conditional on σ^2. Given the observations at $\eta(-2)$, $\eta(1.8)$, and $\eta(2)$, the

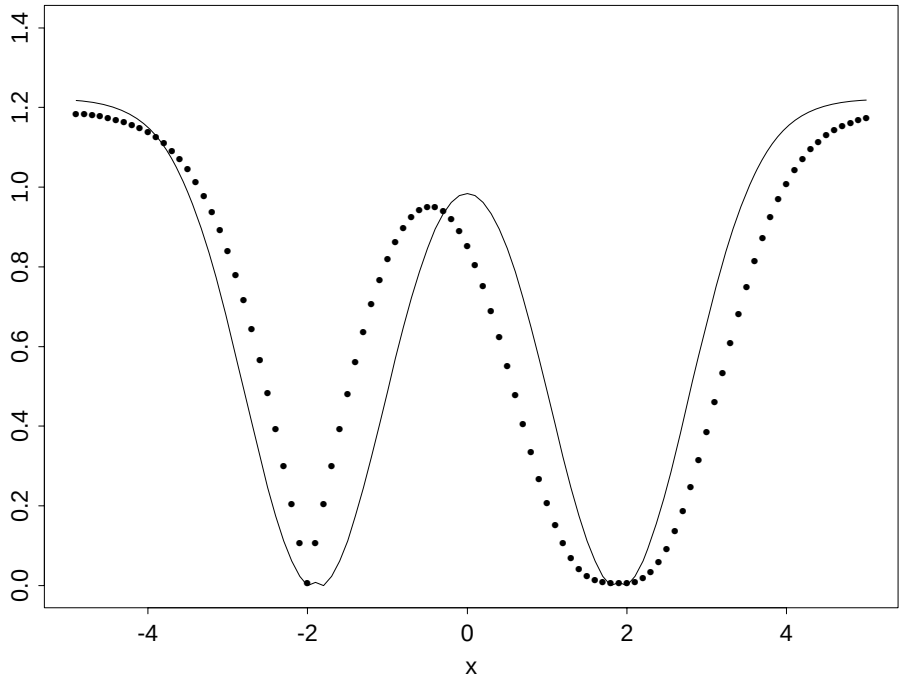

Figure 2.10: The standard deviation of $\eta(x)$ given two sets of data

interest here is in the difference in the variances of $\eta(x)$ after learning $\eta(-1.8)$ or $\eta'(1.8)$. Recalling that $X \sim N(0, 1)$, we can see that the observation at $x = -1.8$ has lead to a greater reduction in uncertainty about $\eta(X)$ in one of the tails of X, whereas the observation of the derivative at $x = 1.8$ has reduced the uncertainty about $\eta(X)$ around the mean of X.

2.4 Conclusions

In this chapter we have reviewed a means of making inference about functions using Gaussian processes, which forms the basis behind the Bayesian approach to uncertainty analysis. We have noted that care will be needed when choosing smoothing parameters in the correlation function. We have also seen how the methodology ex-

tends naturally to inference about derivatives, and that derivative information can usefully supplement runs from the code when learning about the function.

Chapter 3

Uncertainty analysis using simulation

3.1 Introduction

In the previous chapter we developed a model for making inferences about an unknown function $\eta(.)$. Given a distribution for $\mathbf{X}$ and a distribution for $\eta(.)$, we can now consider making inferences about $Y = \eta(\mathbf{X})$. For certain distributions $G(\mathbf{x})$ and certain forms of the correlation function $c(\mathbf{x}, \mathbf{x}')$, some inferences can be made about Y analytically. For example, Haylock (1997) estimates the mean and variance of Y, using a correlation function of the form in (2.12), and independent normally distributed input parameters (or inputs that could be transformed such that the transformed variable had a normal distribution). However, other inferences are not tractable. In chapters four and five, we consider estimating the distribution and density functions of Y, and encounter various difficulties when attempting to obtain estimates analytically. In this chapter, we present a simulation technique, that can be used to make a whole variety of inferences about Y. It will also permit some of the assumptions about $G(\mathbf{x})$ and $c(\mathbf{x}, \mathbf{x}')$ to be relaxed. However, it will only be a practical method when $\mathbf{x}$ has a low number of dimensions.

To recap, we discuss the two levels of uncertainty in the Bayesian approach to uncertainty analysis. Firstly, the value of the true input, $\mathbf{X}$ is unknown. We only

know the distribution of $\mathbf{X}$, given by $G(\mathbf{x})$. Consequently, $Y = \eta(\mathbf{X})$ is a random variable. Now suppose that we wish to make some inference about Y, which we will denote by $S\{\eta(.), G(\mathbf{x})\}$. If we knew the value of $\eta(\mathbf{x})$ for every $\mathbf{x}$, we could determine $S\{\eta(.), G(\mathbf{x})\}$ in principle to an arbitrary precision, perhaps using Monte Carlo methods, or even by analytic means if possible. In the Bayesian approach, we bring in a second level of uncertainty, by treating $\eta(.)$ as an unknown function. For different functions $\eta(.)$, the values of $S\{\eta(.), G(\mathbf{x})\}$ may be different. The result of this is that any summary $S\{\eta(.), G(\mathbf{x})\}$ we might wish to obtain about Y will also be a random variable, and so we will then have to make inferences about $S\{\eta(.), G(\mathbf{x})\}$. Thus for example, Haylock (1997) derives the distribution of $E(Y)$, and the mean and variance of $Var(Y)$.

Having derived the posterior distribution of $\eta(.)$, we can then make inferences about $S\{\eta(.), G(\mathbf{x})\}$ as follows. We first generate a random function from the distribution of $\eta(.)$, which we denote by $\eta_{(i)}(.)$. Now define $Y_{(i)} = \eta_{(i)}(\mathbf{X})$. We can use Monte Carlo sampling to determine $S\{\eta_{(i)}(.), G(\mathbf{x})\}$, assuming it is trivial to evaluate $\eta_{(i)}(\mathbf{x})$ for any $\mathbf{x}$. Repeating this procedure will give us a sample $S\{\eta_{(1)}(.), G(\mathbf{x})\}, S\{\eta_{(2)}(.), G(\mathbf{x})\}, \ldots, S\{\eta_{(N)}(.), G(\mathbf{x})\}$, and this sample can then be used to make inferences about $S\{\eta(.), G(\mathbf{x})\}$. This gives us motivation for finding a technique for generating random functions from the distribution of $\eta(.)$. A requirement is that any simulated function $\eta_{(i)}(.)$ must be computationally cheap, otherwise this process will be slower than using Monte Carlo on the original function $\eta(.)$ directly.

3.2 Generating random functions

In equation (2.33), we gave the posterior distribution of $\eta(\mathbf{x})$ given data $\mathbf{y}$. For any set of inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we know the posterior distribution of the corresponding set of outputs $y_1 = \eta(\mathbf{x}_1), \ldots, y_n = \eta(\mathbf{x}_n)$. The objective is to draw a random function, which we will denote by $\eta_{(i)}(.)$. We cannot obtain an exact realisation of $\eta\,(.)$, since in practice the set $\mathcal{X}$ of possible values of $\mathbf{x}$ is infinite and to sample $\eta(.)$ means to sample $\eta(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. We first choose a set of new design points $(\mathbf{x}'_1, \ldots, \mathbf{x}'_{n'})$,

all distinct from the original design points $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ used to obtain the data $\mathbf{y}$. We call these new points the simulation design points. The vector of outputs $(y'_1 = \eta(\mathbf{x}'_1), \ldots, y'_{n'} = \eta(\mathbf{x}'_{n'}))^T$ will have a joint t distribution, as given in (2.33), and so we can generate values for these outputs, which we will denote by $\mathbf{y}_{(i)}$. We now know the value of the function $\eta_{(i)}(.)$ at $n + n'$ inputs, but everywhere else the function is still a random variable. The function $\eta_{(i)}(.)$ also has a t distribution, and we denote its mean and variance by $m^{**}_{(i)}(.)$ and $\hat{\sigma^2}_{(i)}c^{**}_{(i)}(.)$ respectively. However, for suitably chosen $\mathbf{x}'_1, \ldots, \mathbf{x}'_{n'}$, and for sufficiently large n', the posterior variance $\hat{\sigma}^2_{(i)}c^{**}_{(i)}(.)$ will be small for all values of $\mathbf{x}$ of interest. Consequently, $m^{**}_{(i)}(.)$ will be a close approximation of $\eta_{(i)}(.)$, and so we can think of $m^{**}_{(i)}(.)$ as being an approximate draw from the distribution of $\eta(.)$.

We illustrate this idea with a simple one-dimensional example. We again use the function

$$\eta(x) = 5 + x + \cos x \tag{3.1}$$

where $x \sim N(0, 4)$. For the prior distribution we set $h(x)^T = (1 \quad x)$. For the smoothing parameter we set $b = 0.5$. We then evaluate $\eta(x)$ at five points $(-4.334, -2.054, 0, 2.054, 4.334)$ and derive the posterior distribution of $\eta(.)$. The five outputs are shown in figure 3.1. We sample $\eta(.)$ at a further twelve points, and twelve new generated observations are shown in figure 3.2 as the squares. To obtain a realisation, we now interpolate these points by the posterior mean given all seventeen observations, and this is shown in figure 3.3. In figure 3.4 we have five realisations of $\eta(.)$, shown by the solid lines, and a ninety-five percent pointwise interval for $\eta(x)$ given the five original observations, shown by the dotted lines. Note that each realisation must pass through all the observed outputs. In figure 3.5 we have one realisation $\eta_{(i)}(.)$ and a ninety-nine percent pointwise interval for $\eta_{(i)}(x)$, given the five original observations and the addtional twelve sampled points. The 99% bounds only deviate enough from $m^{**}_{(i)}(.)$ to be visible when $|x| > 4.334$, which is outside the range of inputs where $\eta(.)$ has been sampled or observed. As $x \sim N(0, 4)$, the variance of $\eta_{(i)}(x)$ is negligible for all x of interest.

A complication can arise if two points $\mathbf{x}$ and $\mathbf{x}'$ in the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}'_1, \ldots, \mathbf{x}_{n'}\}$, the set of original design points and simulation design points, are too close together.

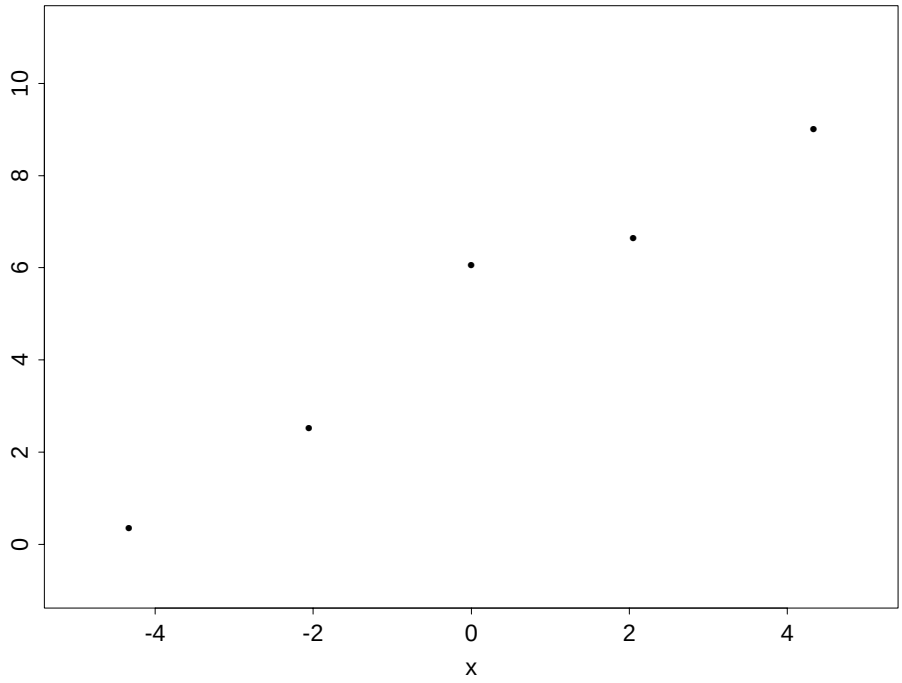

Figure 3.1: The five initial observations

As $c(\mathbf{x}, \mathbf{x}')$ increases as $|\mathbf{x} - \mathbf{x}'|$ decreases, $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ will be highly correlated. This can lead to columns in the correlation matrix that are very similar, and consequently, inverting this matrix accurately can be difficult. Thus in generating random functions, the following practical considerations arise:

1. How should we choose the simulation design points?

2. Is the conditioning of the correlation matrix causing large numerical errors?

3. What accuracy is lost by approximating $\eta_{(i)}(.)$ by $m^{**}_{(i)}(.)$?

We address each of these points in turn.

3.2.1 Choice of the simulation design points

Each randomly drawn function $\eta_{(i)}$ is approximated by its posterior mean $m^{**}_{(i)}(.)$. We require the error in the approximation to be minimal, and so design points should be chosen that will minimise the posterior variance of $\eta_{(i)}(.)$. Hence choosing

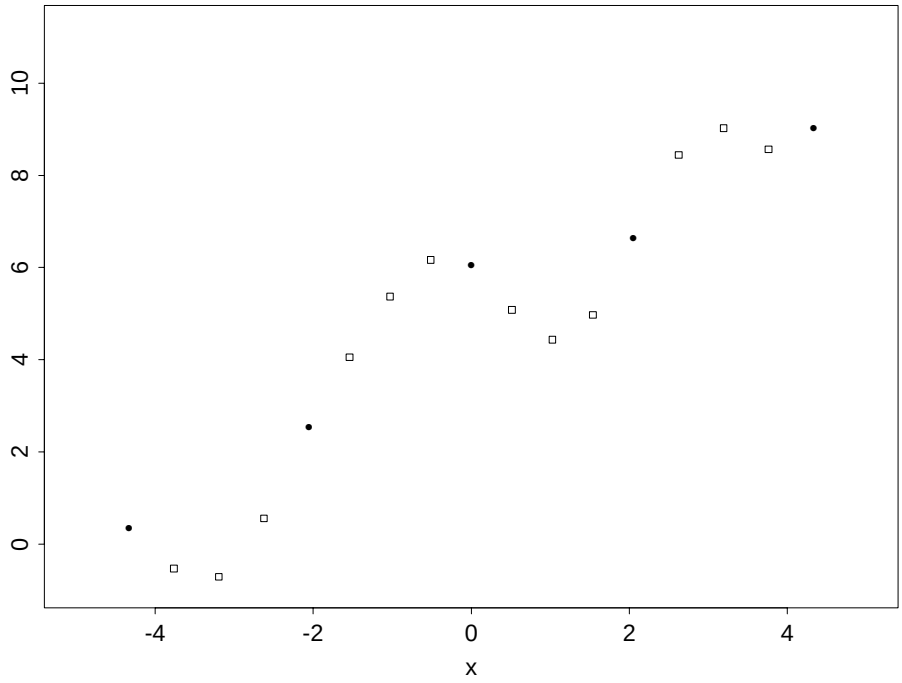

Figure 3.2: The five original points and twelve new observations

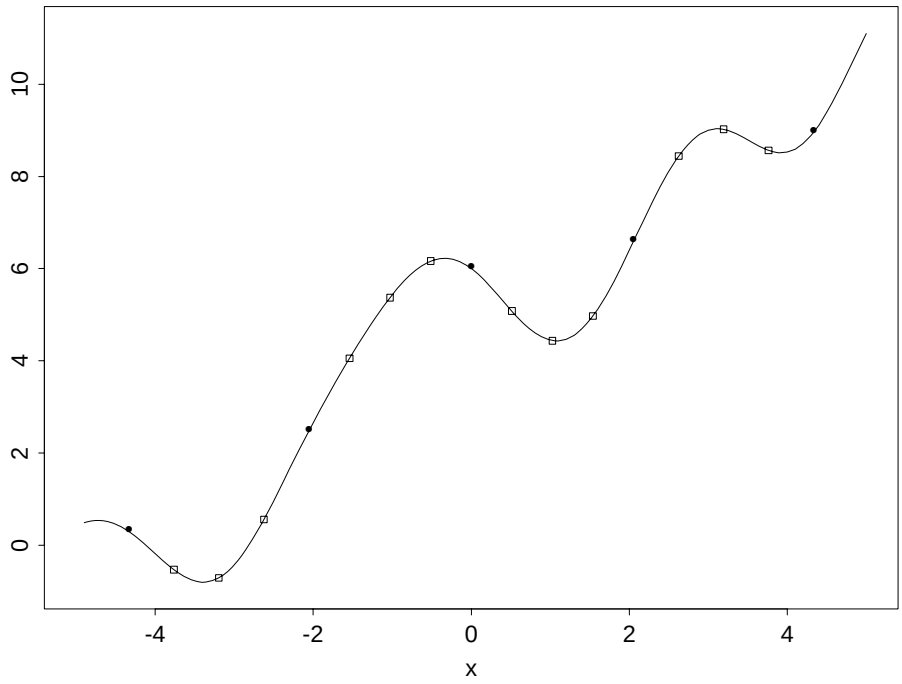

Figure 3.3: Interpolating the seventeen points by the posterior mean

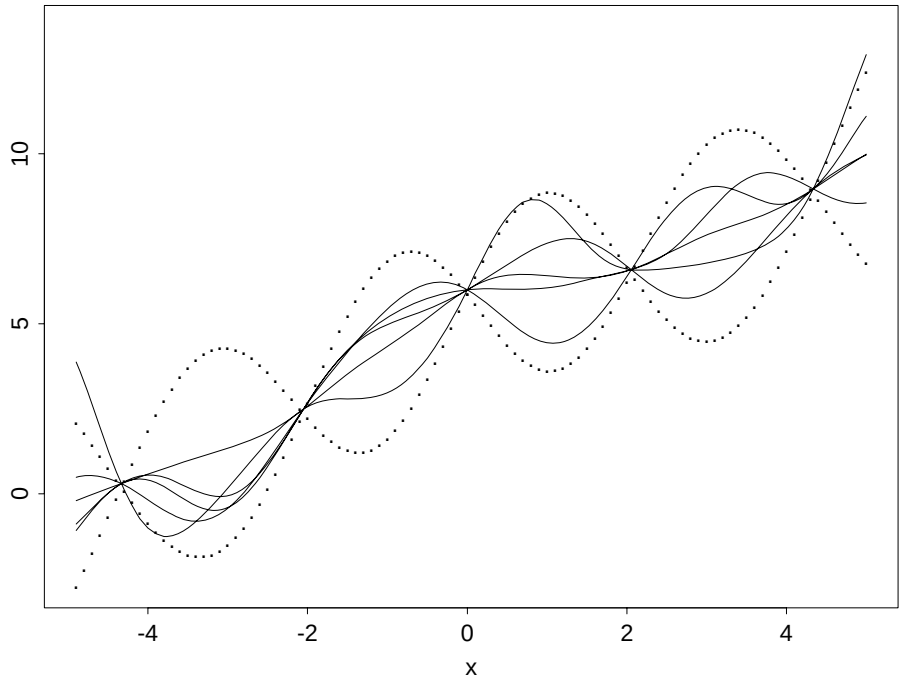

Figure 3.4: Five simulated functions

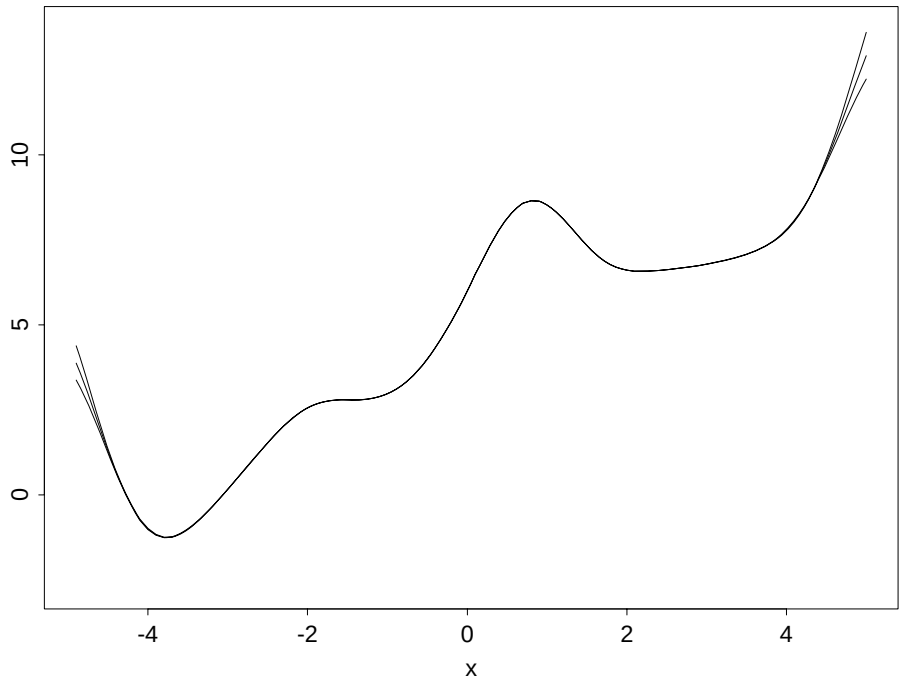

Figure 3.5: A 99% interval for one function after sampling

the simulation design points can be approached in the same way as choosing the initial design points. The only difference in this case will be in the influence of the distribution $G(\mathbf{x})$. For example, suppose $\mathbf{X}$ has a normal distribution. Then the initial design points will be chosen with the result that the variance of $\eta(\mathbf{x})$ is smaller around the mean of $\mathbf{X}$ than it is in the tails. However, when generating a random function, we may want $m^{**}_{(i)}(.)$ to be a good approximation of $\eta_{(i)}(.)$ everywhere over a central 99% (say) interval. In chapter six, we consider an example when we are only interested in the 95th percentile of the output, and so the simulation design points are chosen with this in mind. In this section we proceed on the assumption that we require the variance of $\eta_{(i)}(\mathbf{x})$ to be small everywhere over some interval for $\mathbf{x}$.

In the case of a one dimensional input x, it is usually sufficient to choose the simulation design points at regularly spaced intervals between the intitial design points. This is because we can sample $\eta(.)$ at enough inputs so that finding an optimal design is not crucial. Note that in the one-dimensional example, twelve design points were sufficient to ensure a negligible variance of $\eta_{(i)}(x)$ for all x of interest.

As the number of dimensions increases, the number of design points needed will increase rapidly. Consequently, more care will be needed in selecting the simulation design points. One possibility is to use a combination of Latin hypercube sampling, and the entropy criterion mentioned in section 2.2.3. Writing $\mathbf{x} = (x_1, \ldots, x_d)$, we assign a uniform distibution over the central 99% interval for each component x_i of $\mathbf{x}$. A Latin hypercube sample of inputs, $(\mathbf{x}'_{k_1}, \ldots, \mathbf{x}'_{k'_n})$, is generated. Defining the matrix V_k to have $i,j-$th element $c^{**}(\mathbf{x}'_{k_i}, \mathbf{x}'_{k_j})$, we now evaluate $\det(V_k)$. This procedure is carried out for $k = 1, \ldots, N$, and the sample k is chosen with the largest value of $\det(V_k)$. This should result in design points that are well spaced out, and a small posterior variance. We could also consider using an optimisation routine that will maximise $\det(V)$ as a function of the simulation design points. However, this is likely to be a computationally expensive procedure, as n' will be large. A practical problem with using the entropy criterion in this case is that there can be problems in computing $\det(V)$ due to ill conditioning of the matrix V, which we discuss in

the next subsection.

On a final note, in certain cases it might be possible to tailor the choice of design points to the inference of interest. For example, suppose we wish to estimate $P(Y \leq y)$ for some specific value of y. If for a proposed simulation design point $\mathbf{x}'$ we are almost certain that $\eta(\mathbf{x}') \leq y$, or $\eta(\mathbf{x}') > y$, then it might not be necessary to simulate values of $\eta(\mathbf{x}')$. An example of this is given in chapter six, where we are attempting to estimate a specific percentile of Y.

3.2.2 Conditioning of the correlation matrix

There are two equivalent methods for simulating the values $y'_1 = \eta(\mathbf{x}'_1), \ldots, y'_{n'} = \eta(\mathbf{x}'_{n'})$ that result in two different correlation matrices that will need inverting. One option is to generate the observations sequentially. We simulate in order $(y'_1|\mathbf{y})$, $(y'_2|\mathbf{y}, y'_1)$, $\ldots$, $(y'_{n'}|\mathbf{y}, y'_1, \ldots, y'_{n'-1})$. To simulate y'_j, we need to invert the $(n+j'-1)\times(n+j'-1)$ matrix A defined in (2.16), which comprises of all the prior correlations $c(\mathbf{x}, \mathbf{x}')$ between all the inputs $(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}'_1, \ldots, \mathbf{x}'_{j-1})$. It is the conditioning of the matrix A when we simulate $y'_{n'}$ that needs to be considered.

A more efficient method is to use Cholesky decomposition. Let V be the variance-covariance matrix of $\mathbf{y}_{(i)}$, where the $j,k-$th element of V is given by $c^{**}(\mathbf{x}'_j, \mathbf{x}'_k)$. Let U be the Cholesky square root of V. We now make the transformation $\mathbf{r} = U^{-1}\mathbf{y}_{(i)}$. Then for all i, we have

$$Var(U^{-1}\mathbf{y}_{(i)}) = U^{-1}VU^{-1^T} = U^{-1}UU^TU^{-1^T} = I, \tag{3.2}$$

where I is the identity matrix. So instead, we can generate a vector of independent random variables $\mathbf{r}$, and then make the required transformation back to $\mathbf{y}_{(i)}$. When using this approach, we consider the conditioning of the matrix V.

A further simplification can be made by first simulating $\boldsymbol{\beta}$ and σ^2 from $f(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$. Now we have

$$\eta(\mathbf{x})|\boldsymbol{\beta}, \sigma^2, \mathbf{y} \sim N\{m^*(\mathbf{x}), \sigma^2 c^*(\mathbf{x})\}, \tag{3.3}$$

with $m^*(\mathbf{x})$ and $c^*(\mathbf{x})$ defined in (2.18) and (2.19). The covariance between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ is given by $\sigma^2\left\{c(\mathbf{x}, \mathbf{x}') - \mathbf{t}(\mathbf{x})^T A^{-1}\mathbf{t}(\mathbf{x}')\right\}$.

As long as all the points in the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}'_1, \ldots, \mathbf{x}_{n'}\}$ are distinct, the inverses of A and V will always exist. However, in certain cases we have noted that it can be hard to compute the inverses accurately. When computing A^{-1} for example, we obtain $A^{-1} + E_1$, where E_1 is some unknown error term E_1. We would like to know what effect the matrix E_1 has on the simulation procedure. We suggest the following procedure.

In the process of generating one random function, we generate the random outputs $y'_1 = \eta(\mathbf{x}'_1), \ldots, y'_{n'} = \eta(\mathbf{x}'_{n'})$ directly from their joint t distribution. Suppose we have generated N functions, and consider the sample $[\eta_{(1)}(\mathbf{x}'_i), \ldots, \eta_{(N)}(\mathbf{x}'_i)]$. From the distribution of $\eta(\mathbf{x}'_i)$, we know that the expected mean of this sample is $m^{**}(\mathbf{x}'_i)$, and the variance of the sample mean is $\frac{1}{N}\sigma^2 c^{**}(\mathbf{x}'_i)$. Hence we can calculate

$$Z = \frac{\frac{1}{N}\sum_{j=1}^{N} \eta_{(j)}(\mathbf{x}'_i) - m^{**}(\mathbf{x}'_i)}{\sqrt{\frac{1}{N}\sigma^2 c^{**}(\mathbf{x}'_i)}}, \tag{3.4}$$

and since $Z \sim N(0,1)$, large values of $|Z|$ will suggest that poor conditioning has lead to unacceptable numerical errors. We illustrate this with the function described in (3.1) and the same five initial observations. We simulate 1000 functions and the values of Z for a twelve point simulation design are given in table 3.1, and the values of Z for a twenty point design are given in table 3.2. For each function, the outputs are generated sequentially, with the value of the input increasing at each stage. If ill conditioning is resulting in errors, then we expect the effects to become more apparent as more outputs are generated. In the twenty point case, there is some evidence that poor conditioning has resulted in unreliable samples.

If ill conditioning is resulting in unacceptable numerical errors, there are various methods of resolving the problem. We should first check that all the simulation design points are necessary to achieve a suitably small variance. If no points can be omitted from the design, then an idea used in Neal (1998) to deal with ill conditioning is the introduction of an artificial measurement error, or "jitter". If we are attempting to invert an ill conditioned matrix M, we first add in a small error term along the leading diagonal:

$$M' = M + \epsilon I, \tag{3.5}$$

for some small ϵ. We then approximate M^{-1} by M'^{-1}. The inverse of M' can be

x_i'	Z	x_i'	Z
-3.764	0.55	0.5135	0.68
-3.194	0.79	1.027	0.23
-2.624	0.80	1.5465	-0.05
-1.5405	0.17	2.624	0.69
-1.027	-0.10	3.194	0.37
-0.5135	-0.73	3.764	-1.29

Table 3.1: Values of Z for a twelve point simulation design

x_i'	Z	x_i'	Z
-3.954	0.50	0.342	0.34
-3.574	1.15	0.685	0.95
-3.194	0.28	1.027	1.42
-2.814	0.83	1.369	2.24
-2.434	0.68	1.712	2.23
-1.712	1.40	2.434	-2.90
-1.369	2.55	2.814	-3.22
-1.027	1.14	3.194	-3.13
-0.685	1.51	3.574	-2.49
-0.342	1.72	3.954	-2.45

Table 3.2: Values of Z for a twenty point simulation design

computed more reliably. A small value of ϵ should have little effect on the final inference.

We investigate the effects of introducing jitter in the previous example with the twenty simulation design points. We replace A by $A + \epsilon I$, generate 1000 functions as before and calculate the Z values, for different values of ϵ. In each case, the same set of random numbers are used. The values of Z are shown in table 3.3.

We should also consider the effect of including jitter in the model itself. We plot the mean and 95% interval for $\eta(x)$ after the five initial observations, when $\epsilon = 0$, as dotted lines. We also plot the mean and 95% interval for $\eta(x)$ as solid lines when $\epsilon = 0.1$ in figure 3.6 and when $\epsilon = 0.01$ in figure 3.7.

There are other techniques for dealing with ill conditioned matrices, though they have not offered much improvement in our examples. One idea is to note that if $\eta(\mathbf{x}_i)$ and $\eta(\mathbf{x}_j)$ are highly correlated, then the correlation between $\frac{1}{2}\{\eta(\mathbf{x}_i) + \eta(\mathbf{x}_j)\}$ and $\frac{1}{2}\{\eta(x_i) - \eta(x_j)\}$ is small. Hence we could then simulate $\frac{1}{2}\{\eta(\mathbf{x}_i) + \eta(\mathbf{x}_j)\}$ and $\frac{1}{2}\{\eta(x_i) - \eta(x_j)\}$, as deriving the joint distribution of the sum and the difference is straightforward.

3.2.3 Is there any error in the final inference resulting from approximating each realisation by $m^{**}_{(i)}(.)$?

In low dimensional problems, we can sample $\eta(.)$ at enough inputs such that the variance of any one realisation $\eta_{(i)}(.)$ is minimal. This may not be possible in higher dimensions, and so we need to investigate the effect of approximating $\eta_{(i)}(.)$ by $m^{**}_{(i)}(.)$. This will depend on the inference $S\{\eta(.), G(\mathbf{X})\}$ of interest, since we are approximating $S\{\eta_{(i)}(.), G(\mathbf{X})\}$ by $S\{m^{**}_{(i)}(.), G(\mathbf{X})\}$.

First consider the case when $\eta_{(i)}(\mathbf{x}) < \eta_{(j)}(\mathbf{x})\ \forall \mathbf{x}$ implies either $S\{\eta_{(i)}(.), G(\mathbf{X})\} < S\{\eta_{(j)}(.), G(\mathbf{X})\}$ or $S\{\eta_{(i)}(.), G(\mathbf{X})\} > S\{\eta_{(j)}(.), G(\mathbf{X})\}$, for some particular inference $S\{\eta(.), G(\mathbf{X})\}$. If we can find upper and lower bounds for the function $\eta(.)$, we can find upper and lower bounds for $S\{\eta(.), G(\mathbf{X})\}$. Thus if the distance between the upper and lower bounds for $S\{\eta_{(i)}(.), G(\mathbf{X})\}$ is small, then $S\{m^{**}_{(i)}(.), G(\mathbf{X})\}$ will be an adequate approximation of $S\{\eta_{(i)}(.), G(\mathbf{X})\}$.

x'	$\epsilon = 0$	$\epsilon = 0.01$	$\epsilon = 0.1$
-3.954	0.50	0.50	0.50
-3.574	0.98	1.15	1.44
-3.194	0.98	0.28	-0.27
-2.814	0.75	0.83	1.04
-2.434	0.57	0.68	0.27
-1.712	-0.53	1.40	1.76
-1.369	-0.26	2.55	2.76
-1.027	0.35	1.14	0.03
-0.685	1.14	1.51	0.87
-0.342	1.83	1.72	1.24
0.342	-2.26	0.34	1.04
0.685	-2.08	0.95	1.70
1.027	-1.76	1.42	1.40
1.369	-1.31	2.24	1.85
1.712	-0.67	2.23	1.25
2.434	1.41	-2.90	-1.2
2.814	-2.67	-3.22	-1.52
3.194	-3.75	-3.13	-1.52
3.574	-4.47	-2.49	-0.93
3.954	-4.47	-2.45	-1.22

Table 3.3: Values of Z when jitter is used

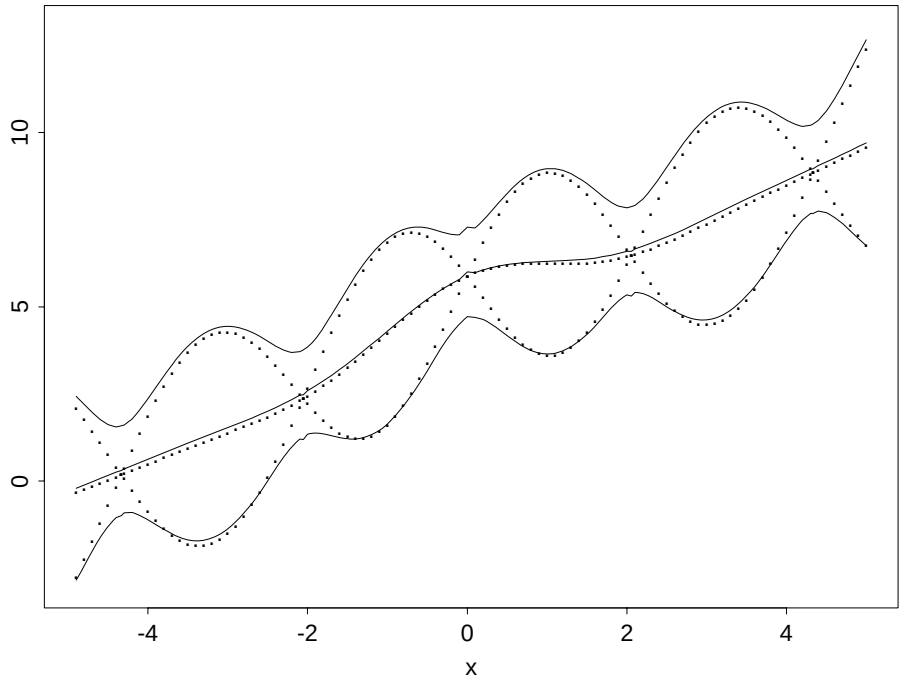

Figure 3.6: The distribution of $\eta(x)$ when $\epsilon = 0.1$

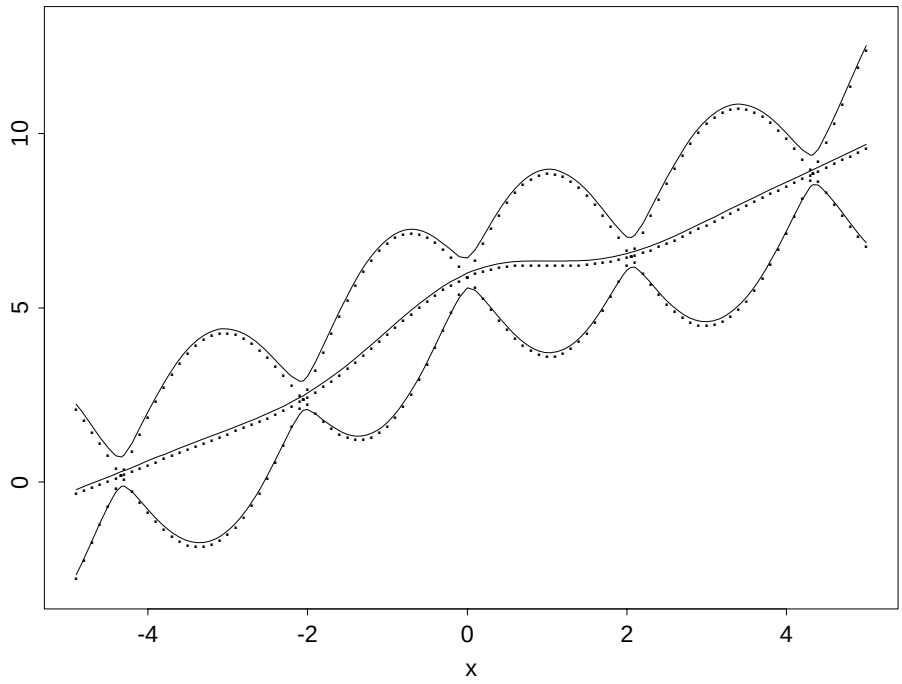

Figure 3.7: The distribution of $\eta(x)$ when $\epsilon = 0.01$

An example of this is when we wish to estimate the distribution function,

$$P\{\eta_{(i)}(\mathbf{X}) \leq y\}, \tag{3.6}$$

which we denote by $F_{(i)}(y)$, for a randomly generated function $\eta_{(i)}(.)$. We estimate $F_{(i)}(y)$ by

$$\hat{F}_{(i)}(y) = \frac{1}{N}\sum_{j=1}^{N} I\left\{m^{**}_{(i)}(\mathbf{x}^*_j) \leq y\right\}, \tag{3.7}$$

where $(\mathbf{x}^*_1, \ldots, \mathbf{x}^*_N)$ are randomly drawn from $G(\mathbf{x})$. We can also obtain lower and upper bounds, $F^L_{(i)}(y)$ and $F^U_{(i)}(y)$ for $F_{(i)}(y)$, using

$$F^L_{(i)}(y) = \frac{1}{N}\sum_{j=1}^{N} I\left\{m^{**}_{(i)}(\mathbf{x}^*_j) + \alpha\hat{\sigma}_{(i)}\sqrt[2]{c^{**}_{(i)}(\mathbf{x}^*_j)} \leq y\right\} \tag{3.8}$$

$$F^U_{(i)}(y) = \frac{1}{N}\sum_{j=1}^{N} I\left\{m^{**}_{(i)}(\mathbf{x}^*_j) - \alpha\hat{\sigma}_{(i)}\sqrt[2]{c^{**}_{(i)}(\mathbf{x}^*_j)} \leq y\right\}, \tag{3.9}$$

for some appropriate value of α. If $F^L_{(i)}(y)$ and $F^U_{(i)}(y)$ are sufficiently close (or preferably indistinguishable), then the error in the approximation is acceptable.

If the condition described previously does not hold, then we can not find bounds for $S\{\eta(.), G(\mathbf{X})\}$ in this manner. In higher dimensions it is more likely that $m^{**}_{(i)}(.)$ will be an overly smooth approximation of $\eta_{(i)}$. The main concern is that the sample $S\{\eta_{(1)}(.), G(\mathbf{X})\}, \ldots, S\{\eta_{(N)}(.), G(\mathbf{X})\}\}$ is sufficiently varied so that we do not underestimate the uncertainty in $S\{\eta(.), G(\mathbf{X})\}$. A method for checking this is as follows. The simulation design points are chosen and a sample $S\{\eta_{(1)}(.), G(\mathbf{X})\}, \ldots,$ $S\{\eta_{(N)}(.), G(\mathbf{X})\}\}$ is obtained. Then conditional on both $y_1 = \eta(\mathbf{x}_1), \ldots, y_n = \eta(\mathbf{x}_n)$ and $y'_1 = \eta(\mathbf{x}'_1), \ldots, y'_{n'} = \eta(\mathbf{x}'_{n'})$, we find a small number of additional inputs where the variance of the outputs are largest, and include these into the simulation design points. We then obtain a second sample $S\{\eta_{(1)}(.), G(\mathbf{X})\}', \ldots, S\{\eta_{(N)}(.), G(\mathbf{X})\}\}'$ and compare various summaries of the two samples. If there is little difference between the two, then we can be confident that we have captured the variability of $S\{\eta(.), G(\mathbf{X})\}$ adequately.

3.3 Example: uncertainty analysis with alternative covariance functions

We now give a simple example of the use of the simulation approach in uncertainty analysis. We again consider the function

$$\eta(x) = 5 + x + \cos x, \tag{3.10}$$

with the true input $X \sim N(0, 4)$. Haylock (1997) gave expressions for the mean and variance of $\eta(X)$, which could be computed analytically when $c(x, x') = \exp\{-b(x - x')^2\}$. We now consider the covariance function

$$c(x, x') = \exp\left(-\frac{|d|}{2}\right). \tag{3.11}$$

With this function, computing the formulae in Haylock (1997) analytically is not a straightforward task. We use the simulation approach to estimate $M = E\{\eta(X)\}$. We first evalute $\eta(x)$ at five inputs to obtain data $\mathbf{y}$. A realisation of $\eta(.)$, denoted by $\eta_{(i)}(.)$ is generated, and $E\{\eta_{(i)}(X)\}$ is estimated by drawing 1000 inputs x^* from $N(0, 4)$ and calculating the sample mean of $\eta_{(i)}(x^*)$. This process is repeated to obtain 1000 generated functions $\eta_{(i)}(.)$ and 1000 simulated values of $E\{\eta_{(i)}(X)\}$. From the sample we estimate $E(M|\mathbf{y}) = 5.06$ and $Var(M|\mathbf{y}) = 0.2$. We can check these answers by evaluating $\eta(x)$ at a very large number of inputs and calculating the sample mean. The true value of the expected output is 5.13.

3.4 Conclusions

We have presented a method that can in principle be used to make any inference about Y, based on a small number of runs of the computer code. The procedure relies on being able to simulate enough observations so that the posterior variance of any particular realisation is small. As the number of dimensions in the input $\mathbf{x}$ increases, the number of simulation design points required may increase rapidly, and the simulation method may not be as effective for computational reasons.

Chapter 4

The distribution and percentile functions

4.1 Introduction

We now consider means of quantifying the uncertainty in the true output Y. Suppose that the computer code is modelling some situation where a certain output c is considered 'critical'. The user of the model may have to make a decision on the basis of whether of not the true output will exceed this critical value. A natural question to ask is, what is the probability that Y is less than c? Clearly, a summary of interest is the distribution function

$$F_Y(y) = \int_{\mathcal{X}} I\{\eta(\mathbf{x}) \leq y\} dG(\mathbf{x}), \tag{4.1}$$

where $I\{.\}$ denotes the indicator function. If $F_Y(c)$ is close to 0.5, then we have high uncertainty about whether or not Y will exceed c, as a consequence of not knowing $\mathbf{X}$. Thus even if the model describes reality accurately, the unknown inputs have made the model less useful in the decision making process. Alternatively, values of $F_Y(c)$ close to 0 or 1 imply that the uncertainty in the inputs has not induced much uncertainty in the output at the critical value.

As discussed in chapter three, $F_Y(y)$ is a random variable, because we do not know the value of $\eta(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$. Thus in this chapter, we estimate (4.1), and find means of quantifying the uncertainty in our estimate. The simulation technique

can be used for estimating (4.1), and this will be considered in section 4.3.1. We will begin with an analytic approach.

Note that estimating distribution functions is a subject that has long been of interest in the statistical community. An early example is Kaplan and Meier (1958), who were interested in estimating distribution functions based on incomplete data. Nonparametric approaches are frequently adopted, and a common method involves the use of a Dirichlet process prior for $F_Y(y)$ (see for example O'Hagan, 1995). Further discussion is in Walker et al. (1999), who consider generalisations of the Dirichlet prior. A novel approach will be necessary here. Firstly, the data may not be a sample from the distribution of Y, if the design points are not chosen at random from $G(\mathbf{x})$. In addition, we also have the opportunity here to utilise the information that the data gives us about $\eta(.)$.

4.2 Posterior moments of $F_Y(y)$

We derive the first two posterior moments of $F_Y(y)$. Using the result in (2.33), we obtain

$$\begin{aligned} E\left\{F_Y(y)\right\} &= \int_{\mathcal{X}} E\left[I\left\{\eta(\mathbf{x}) \leq y\right\}\right] dG(\mathbf{x}) \\ &= \int_{\mathcal{X}} P\left\{\eta(\mathbf{x}) \leq y\right\} dG(\mathbf{x}) \\ &= \int_{\mathcal{X}} P\left\{\frac{\eta(\mathbf{x}) - m^{**}(\mathbf{x})}{\hat{\sigma}\sqrt{c^{**}(\mathbf{x})}} \leq \frac{y - m^{**}(\mathbf{x})}{\hat{\sigma}\sqrt{c^{**}(\mathbf{x})}}\right\} dG(\mathbf{x}) \\ &= \int_{\mathcal{X}} \int_{-\infty}^{\frac{y-m^{**}(\mathbf{X})}{\hat{\sigma}\sqrt{c^{**}(\mathbf{X})}}} f_{T_{n-q}}(t)\, dt\, dG(\mathbf{x}), \end{aligned} \tag{4.2}$$

where $f_{T_{n-q}}(t)$ is the density function of a t random variable with $n-q$ degrees of freedom. For the posterior covariance we require

$$E\{F_Y(s_1)F_Y(s_2)\} = E\left[\int_{\mathcal{X}}\int_{\mathcal{X}} I\left\{\eta(\mathbf{x}) \leq s_1\right\} I\left\{\eta(\mathbf{z}) \leq s_2\right\} dG(\mathbf{x})\, dG(\mathbf{z})\right]. \tag{4.3}$$

Since

$$I\left\{\eta(\mathbf{x}) \leq s_1\right\} I\left\{\eta(\mathbf{z}) \leq s_2\right\} = I\left[\left\{\eta(\mathbf{x}) \leq s_1\right\} \bigcap \left\{\eta(\mathbf{z}) \leq s_2\right\}\right] \tag{4.4}$$

We will require

$$P\left[\left\{\eta(\mathbf{x}) \leq s_1\right\} \bigcap \left\{\eta(\mathbf{z}) \leq s_2\right\}\right] = P\left\{\eta(\mathbf{z}) \leq s_2\right\} P\left\{\eta(\mathbf{x}) \leq s_1 | \eta(\mathbf{z}) \leq s_2\right\}$$

$$= \int_{-\infty}^{s_2} P\{\eta(\mathbf{x}) \leq s_1 \mid \eta(\mathbf{z}) = k\} f_{\eta(\mathbf{z})}(k) dk \quad (4.5)$$

where $f_{\eta(\mathbf{z})}(k)$ is the density function of $\eta(\mathbf{z})$. Now $\eta(\mathbf{x})|\eta(\mathbf{z}) = k$ has a t distribution similar to that of $\eta(\mathbf{x})$, but with the addition of one extra data point, $\eta(\mathbf{z}) = k$. Hence

$$E\{F_Y(s_1)F_Y(s_2)\} = \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{-\infty}^{s_2}\int_{-\infty}^{\frac{s_1 - m_k^{**}(\mathbf{X})}{\hat{\sigma_k}\sqrt{c_k^{**}(\mathbf{X})}}} f_{T_{n-q+1}}(t)\, f_{\eta(\mathbf{z})}(k)\, dt\, dk\, dG(\mathbf{x})\, dG(\mathbf{z}) \quad (4.6)$$

where $m_k^{**}(\mathbf{x})$ and $\hat{\sigma}_k^2 c_k^{**}(\mathbf{x})$ and are the posterior mean and variance of $\eta(\mathbf{x})$ conditional on both the data $\mathbf{y}$ and the additional 'observation' $\eta(\mathbf{z}) = k$. We evaluate these integrals numerically.

Two difficulties are encountered with this approach to inference about $F_Y(y)$. Firstly, we note that (4.2) may not be a good location summary in the tails. Clearly, $F_Y(y)$ is constrained to take values between zero and one. At a high or low value of y, it is possible that the distribution of $F(y)$ will be skewed. Hence if we estimate $F_Y(y)$ by its mean we may overestimate $F_Y(y)$ at low values of y, and underestimate $F_Y(y)$ at high values of y.

Secondly, care will be needed when evaluating (4.2) at outputs y_j when $\eta(\mathbf{x}_j) = y_j$ has been observed. In this case we will have $P\{\eta(\mathbf{x}_j) \leq y_j\} = 1$, but since we know little about the derivative of $\eta(.)$ at x_j, we will have considerably less certainty as to whether $\eta(\mathbf{x}_j + \boldsymbol{\delta})$ exceeds y_j or not for any infinitesimal $\boldsymbol{\delta}$. Thus there is a discontinuity in the integrand at $\mathbf{x} = \mathbf{x}_j$. Writing

$$P\{\eta(\mathbf{x}) \leq y\} = \int_{-\infty}^{\frac{y - m^{**}(\mathbf{X})}{\hat{\sigma}\sqrt{c^{**}(\mathbf{X})}}} f_{T_{n-q}}(t)dt \quad (4.7)$$

we now show that

$$\lim_{\delta \searrow 0} \frac{\eta(\mathbf{x}_j) - m^{**}(\mathbf{x}_j + \delta\mathbf{x}_0)}{\hat{\sigma}\sqrt{c^{**}(\mathbf{x} + \delta\mathbf{x}_0)}} = -\lim_{\delta \nearrow 0} \frac{\eta(\mathbf{x}_j) - m^{**}(\mathbf{x}_j + \delta\mathbf{x}_0)}{\hat{\sigma}\sqrt{c^{**}(\mathbf{x} + \delta\mathbf{x}_0)}}, \quad (4.8)$$

with the result that

$$\lim_{\delta \searrow 0} P\{\eta(\mathbf{x}_j + \delta\mathbf{x}_0) \leq y_j\} = 1 - \lim_{\delta \nearrow 0} P\{\eta(\mathbf{x}_j + \delta\mathbf{x}_0) \leq y_j\}, \quad (4.9)$$

for any constant $\mathbf{x}_0$.

We suppose that the input $\mathbf{x}$ has d dimensions and we write $\mathbf{x} = (x_1, \ldots, x_d)^T$. We also write $\mathbf{x}_0 = (x_{0_1}, \ldots, x_{0_d})^T$. We define $\mathbf{b}^T = (b_1, b_2, \ldots, b_d)$ to be the vector of d smoothing parameters. First consider the mean of $\eta(\mathbf{x}_j)$ in the neighbourhood of $\mathbf{x}_j$, when $y_j = \eta(\mathbf{x}_j)$ is known. We have

$$m^{**}(\mathbf{x}_j + \delta\mathbf{x}_0) = \mathbf{h}(\mathbf{x}_j + \delta\mathbf{x}_0)^T\boldsymbol{\beta} + \mathbf{t}(\mathbf{x}_j + \delta\mathbf{x}_0)^T A^{-1}(\mathbf{y} - H\boldsymbol{\beta}). \tag{4.10}$$

We suppose that $\eta(.)$ has been evaluated at n design points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. We write $c(\mathbf{x}_i, \mathbf{x}) = t_i(\mathbf{x})$, where $\mathbf{x}_i$ is a design point, and we assume that $\mathbf{h}(x)$ is differentiable, and that $c(\mathbf{x}, \mathbf{x}')$ is of the exponential product form given in (2.12). We then have $\mathbf{t}(\mathbf{x})^T = \{t_1(\mathbf{x}), \ldots, t_n(\mathbf{x})\}$, and we write

$$A^{-1} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & \ddots & & \\ \vdots & & & \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}. \tag{4.11}$$

Then

$$m^{**}(\mathbf{x}_j + \delta\mathbf{x}_0) = \mathbf{h}(\mathbf{x}_j + \delta\mathbf{x}_0)^T\boldsymbol{\beta} + \sum_{k=1}^{n}\left\{\sum_{i=1}^{n} t_i(\mathbf{x}_j + \delta\mathbf{x}_0)a_{ik}\right\}\left\{\eta(\mathbf{x}_k) - \mathbf{h}(\mathbf{x}_k)^T\boldsymbol{\beta}\right\}. \tag{4.12}$$

Note that $\mathbf{t}(x_j)^T$ is the j-th row of A, and so

$$\sum_{i=1}^{n} t_i(\mathbf{x}_j)a_{ik} = \begin{cases} 1 & \text{if} \quad k = j \\ 0 & \text{otherwise.} \end{cases} \tag{4.13}$$

We use the notation

$$\nabla f(\mathbf{x}) = \{\frac{\partial}{\partial x_1} f(\mathbf{x}), \ldots, \frac{\partial}{\partial x_d} f(\mathbf{x})\}, \tag{4.14}$$

and

$$D\{f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})\}^T = \{\nabla f_1(\mathbf{x}), \ldots, \nabla f_n(\mathbf{x})\}^T. \tag{4.15}$$

Then if $\mathbf{h}(.)$ and $c(.,.)$ are differentiable we can write

$$t_i(\mathbf{x}_j + \delta\mathbf{x}_0) = t_i(\mathbf{x}_j) + \delta\left\{\nabla t_i(\mathbf{x}_j)\right\}\mathbf{x}_0 + O(\delta^2) \tag{4.16}$$

and

$$\mathbf{h}(\mathbf{x}_j + \delta\mathbf{x}_0)^T\boldsymbol{\beta} = \mathbf{h}(\mathbf{x}_j)^T\boldsymbol{\beta} + \delta\mathbf{x}_0^T\{D\mathbf{h}(\mathbf{x}_j)^T\}\boldsymbol{\beta} + O(\delta^2). \tag{4.17}$$

Then using the result in (4.13) we can write

$$\begin{aligned} m^{**}(\mathbf{x}_j + \delta\mathbf{x}_0) - \eta(\mathbf{x}_j) &= \mathbf{h}(\mathbf{x}_j)^T\boldsymbol{\beta} + \left\{\eta(\mathbf{x}_j) - \mathbf{h}(\mathbf{x}_j)^T\boldsymbol{\beta}\right\} \\ &\quad +\delta\mathbf{x}_0^T\left[\{D\mathbf{h}(\mathbf{x}_j)^T\}\boldsymbol{\beta} + \{D\mathbf{t}(\mathbf{x}_j)^T\}A^{-1}(\mathbf{y} - H\boldsymbol{\beta})\right] \\ &\quad -\eta(\mathbf{x}_j) + O(\delta^2) \qquad (4.18) \\ &= m^{**}(\mathbf{x}_j) + \delta\left\{\nabla m^{**}(\mathbf{x}_j)\right\}\mathbf{x}_0 - \eta(\mathbf{x}_j) + O(\delta^2) \qquad (4.19) \\ &= \delta c_1 + O(\delta^2), \qquad (4.20) \end{aligned}$$

for some constant c_1. Hence $m^{**}(\mathbf{x}_j) + \delta\mathbf{x}_0) - \eta(\mathbf{x}_j)$ is of order δ.

We now show that $c^{**}(\mathbf{x}_j + \delta\mathbf{x}_0)$ is of order δ^2. First consider the function $\mathbf{t}(\mathbf{x}_j + \delta\mathbf{x}_0)^T A^{-1}\mathbf{t}(\mathbf{x}_j + \delta\mathbf{x}_0)$. If we write $\mathbf{x}_i = (x_{i_1}, \ldots, x_{i_d})$ then

$$\begin{aligned} \frac{\partial}{\partial x_l}t_i(\mathbf{x}) &= \frac{\partial}{\partial x_l}\prod_{k=1}^{d}\exp\{-b_k(x_k - x_{i_k})^2\} \qquad (4.21) \\ &= -2b_l(x_l - x_{i_l})\prod_{k=1}^{d}\exp\{-b_k(x_k - x_{i_k})^2\}, \qquad (4.22) \end{aligned}$$

which is equal to zero if $\mathbf{x} = \mathbf{x}_i$. Also,

$$\begin{aligned} \frac{\partial^2}{\partial x_l^2}t_i(\mathbf{x}) &= \left[4b_l^2\exp\{-b_l(x_l - x_{i_l})^2\} - 2b_l\exp\{-b_l(x_l - x_{i_l})^2\}\right] \\ &\quad \times\prod_{k=1}^{d}\exp\{-b_k(x_k - x_{i_k})^2\}, \qquad (4.23) \end{aligned}$$

which is equal to $-2b_l$ if $\mathbf{x} = \mathbf{x}_i$. It then follows that

$$\begin{aligned} \mathbf{t}(\mathbf{x}_j + \delta\mathbf{x}_0)^T A^{-1}\mathbf{t}(\mathbf{x}_j + \delta\mathbf{x}_0) &= \mathbf{t}(\mathbf{x}_j)^T A^{-1}\mathbf{t}(\mathbf{x}_j) + \delta\{D\mathbf{t}(\mathbf{x}_j)^T\}\mathbf{x}_0 A^{-1}\mathbf{t}(\mathbf{x}_j) \\ &\quad +\delta\mathbf{t}(\mathbf{x}_j)^T A^{-1}\left\{D\mathbf{t}(\mathbf{x}_j)\right\}\mathbf{x}_0 \\ &\quad +\delta\{D\mathbf{t}(\mathbf{x}_j)^T\}\mathbf{x}_0 A^{-1}\left\{D\mathbf{t}(\mathbf{x}_j)\right\}\mathbf{x}_0\delta \\ &\quad +\frac{\delta^2}{2}\{D^2\mathbf{t}(\mathbf{x}_j)^T\}\mathbf{x}_0 A^{-1}\mathbf{t}(\mathbf{x}_j) \\ &\quad +\frac{\delta^2}{2}\mathbf{t}(\mathbf{x}_j)^T A^{-1}\{D^2\mathbf{t}(\mathbf{x}_j)^T\}\mathbf{x}_0 \\ &\quad +O(\delta^3) \qquad (4.24) \\ &= 1 - 2\mathbf{b}^T\mathbf{x}_0 + \delta^2\{D\mathbf{t}(\mathbf{x}_j)^T\}\mathbf{x}_0 A^{-1}\left\{D\mathbf{t}(\mathbf{x}_j)\right\}\mathbf{x}_0 \\ &\quad +O(\delta^3), \qquad (4.25) \end{aligned}$$

since $\mathbf{t}(\mathbf{x}_j)A^{-1}$ and $A^{-1}\mathbf{t}(\mathbf{x}_j)^T$ are the $j-$th row and column of the identity matrix respectively. Now consider $\mathbf{h}(\mathbf{x}_j+\delta\mathbf{x}_0)^T-\mathbf{t}(\mathbf{x}_j+\delta\mathbf{x}_0)^TA^{-1}H$. We write

$$\begin{aligned}\mathbf{h}(\mathbf{x}_j+\delta\mathbf{x}_0)^T-\mathbf{t}(\mathbf{x}_j+\delta\mathbf{x}_0)^TA^{-1}H &= \mathbf{h}(\mathbf{x}_j)^T-\mathbf{t}(x_j)^TA^{-1}H\\ &\quad+\delta\left[\{D\mathbf{h}(\mathbf{x}_j)^T\}\mathbf{x}_0+\{D\mathbf{t}(\mathbf{x}_j)^T\}\mathbf{x}_0A^{-1}H\right]\\ &\quad+O(\delta^2) \qquad (4.26)\end{aligned}$$

Since $\mathbf{h}(\mathbf{x}_j)$ is the $j-$th row of H, we have $\mathbf{h}(\mathbf{x}_j)-\mathbf{t}(\mathbf{x}_j)^TA^{-1}H=0$. Hence we have the result that

$$\begin{aligned}c^{**}(\mathbf{x}_j+\delta\mathbf{x}_0) &= \delta\left[\{D\mathbf{h}(\mathbf{x}_j)^T\}\mathbf{x}_0+\{D\mathbf{t}(\mathbf{x}_j)^T\}\mathbf{x}_0A^{-1}H\right](H^TA^{-1}H)^{-1}\\ &\quad\times\left[\{D\mathbf{h}(\mathbf{x}_j)^T\}\mathbf{x}_0+\{D\mathbf{t}(\mathbf{x}_j)^T\}\mathbf{x}_0A^{-1}H\right]^T\delta+\mathbf{b}^T\mathbf{x}_0\\ &\quad-\delta^2\{D\mathbf{t}(\mathbf{x}_j)^T\}\mathbf{x}_0A^{-1}\left\{D\mathbf{t}(\mathbf{x}_j)\right\}\mathbf{x}_0-O(\delta^3). \qquad (4.27)\end{aligned}$$

and the result in (4.8) is proved, since δ^2 is strictly positive.

For a sufficiently small value of δ, we could also derive $P\{\eta(\mathbf{x}_j+\delta\mathbf{x}_0)\le y_j\}$ by considering the derivative of $\eta(\mathbf{x})$ at $\mathbf{x}=\mathbf{x}_j$ in the direction of $\mathbf{x}_0$. If δ is small and positive, then $\eta(\mathbf{x}_j+\delta\mathbf{x}_0)$ is less than y_j if $\{\nabla\eta(\mathbf{x}_j)\}\mathbf{x}_0$ is negative. From the previous analysis we have the result that

$$P\{\eta(\mathbf{x}_j+\delta\mathbf{x}_0)\le y_j\}=\int_{-\infty}^{\frac{-\delta\left\{\nabla m^{**}(\mathbf{x}_j)\right\}\mathbf{x}_0+O(\delta^2)}{\hat{\sigma}\sqrt{\delta^2v(\mathbf{x}_j)-O(\delta^3)}}}f_{T_{n-q}}(t)dt, \qquad (4.28)$$

where

$$\begin{aligned}v(\mathbf{x}_j) &= \left[\{D\mathbf{h}(\mathbf{x}_j)^T\}\mathbf{x}_0+\{D\mathbf{t}(\mathbf{x}_j)^T\}\mathbf{x}_0A^{-1}H\right](H^TA^{-1}H)^{-1}\\ &\quad\times\left[\{D\mathbf{h}(\mathbf{x}_j)^T\}\mathbf{x}_0+\{D\mathbf{t}(\mathbf{x}_j)^T\}\mathbf{x}_0A^{-1}H\right]^T+\mathbf{b}^T\mathbf{x}_0\\ &\quad-\delta\{D\mathbf{t}(\mathbf{x}_j)^T\}\mathbf{x}_0A^{-1}\left\{D\mathbf{t}(\mathbf{x}_j)\right\}\mathbf{x}_0 \qquad (4.29)\end{aligned}$$

Alternatively, $\{\nabla\eta(\mathbf{x}_j)\}\mathbf{x}_0|\mathbf{y}$ also has a t distribution with $n-q$ degrees of freedom. The mean of $\{\nabla\eta(\mathbf{x}_j)\}\mathbf{x}_0|\mathbf{y}$ is $\{\nabla m^{**}(\mathbf{x}_j)\}\mathbf{x}_0$, and the variance of $\{\nabla\eta(\mathbf{x}_j)\}\mathbf{x}_0|\mathbf{y}$ is $\sigma^2v(\mathbf{x}_j)$. Hence

$$P\left[\{\nabla\eta(\mathbf{x}_j)\}\mathbf{x}_0\le 0\right]=\int_{-\infty}^{\frac{\left\{\nabla m^{**}(\mathbf{x}_j)\right\}\mathbf{x}_0}{\hat{\sigma}\sqrt{v(\mathbf{x}_j)}}}f_{T_{n-q}}(t)dt, \qquad (4.30)$$

which is equivalent to the previous expression for sufficiently small δ. The result in (4.9) can easily be understood in terms of derivatives. For a small positive δ we have $P\{\eta(\mathbf{x}_j + \delta\mathbf{x}_0) \leq y_j\} \simeq P\left[\{\nabla\eta(\mathbf{x}_j)\}\,\mathbf{x}_0 \leq 0\right]$ and $P\{\eta(\mathbf{x}_j - \delta\mathbf{x}_0) < y_j\} \simeq 1 - P\left[\{\nabla\eta(\mathbf{x}_j)\}\,\mathbf{x}_0 \leq 0\right]$.

Despite this discontinuity, (4.2) can still be evaluated numerically to a reasonable accuracy. For example, suppose $\mathbf{x} = (x_1, x_2)$, with $x_1 \in (-\infty, \infty)$ and $x_2 \in (-\infty, \infty)$, and that we have observed $\eta(x_{1_j}, x_{2_j}) = c_j$. Then for some small value of ε we can use the approximation

$$\begin{aligned}\int_{\mathcal{X}} P\{\eta(\mathbf{x}) \leq c_j\}dG(\mathbf{x}) \;\simeq\; & \int_{-\infty}^{\infty}\int_{-\infty}^{x_{1_j}-\varepsilon} P\{\eta(\mathbf{x}) \leq c_j\}dG(x_1)dG(x_2) \\ & + \int_{-\infty}^{\infty}\int_{x_{1_j}+\varepsilon}^{\infty} P\{\eta(\mathbf{x}) \leq c_j\}dG(x_1)dG(x_2). \quad (4.31)\end{aligned}$$

In addition, we will only have to use this approximation for the n observed outputs. However, calculating (4.6) numerically is harder. This is because for each $k \in \mathcal{X}$ with $k \leq y_2$ we condition on $\eta(\mathbf{z}) = k$ and determine $P(\eta(\mathbf{x}) \leq y_1 | \eta(\mathbf{z}) = k)$. This function will then have a discontinuity at $\mathbf{x} = \mathbf{z}$ if $k = y_1$. The problem is most acute when we are calculating the variance of $F_Y(y)$. In this case $y_1 = y_2$, and so there will be a discontinuity for all values of y_1, even if they have not been observed.

We illustrate the discontinuity in (4.2) with the simple two dimensional function

$$\eta(x_1, x_2) = 5 + x_1 + x_2 + \cos(x_1) + 2\sin(x_2) \quad (4.32)$$

We set

$$h(x_1, x_2)^T = \begin{pmatrix} 1 & x_1 & x_2 \end{pmatrix} \quad (4.33)$$

$$c\{(x_1, x_2), (x_1', x_2')\} = \exp\left\{-\begin{pmatrix} x_1 & x_2 \end{pmatrix}\begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}\begin{pmatrix} x_1' & x_2' \end{pmatrix}^T\right\}. \quad (4.34)$$

We evaluate $\eta(.)$ at nine points, including the input $(0, 0)$ where we observe $\eta(0, 0) = 6$. In figure 4.1 we plot $P\{\eta(x_1, x_2) \leq 6\}$ for different values of x_1 and x_2.

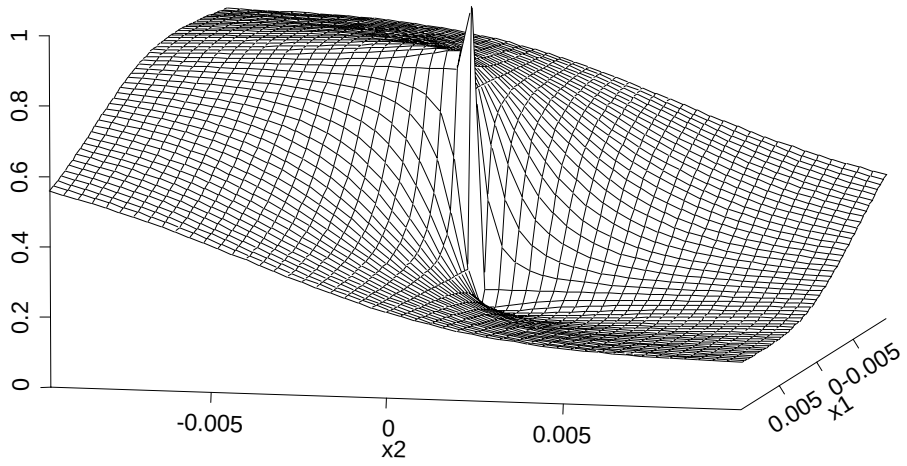

Figure 4.1: The integrand of (4.2) at an observed output

4.3 Alternative methods for estimating $F_Y(y)$

4.3.1 Using simulation

The simulation approach described in chapter three can be used to make inferences about $F_Y(y)$. We first draw a large sample of inputs randomly from $G(\mathbf{x})$, which we denote by $\mathbf{x}_1^*, \ldots, \mathbf{x}_N^*$. Then for one randomly generated function $\eta_{(i)}(.)$, we estimate its distribution function $P\{\eta_{(i)}(\mathbf{X}) \leq y\}$ by

$$\hat{F}_{Y_{(i)}}(y) = \frac{1}{N} \sum_{j=1}^{N} I\{m_{(i)}^{**}(\mathbf{x}_j^*) \leq y\}, \tag{4.35}$$

where $I\{.\}$ denotes the indicator function. We repeat this procedure a large number of times and obtain a sample of distribution functions $F_{Y_{(1)}}(y), \ldots, F_{Y_{(N')}}(y)$. We can then obtain summaries of $F_Y(y)$ from this sample. In addition to estimating the mean and variance of $F_Y(y)$, we can also check to see whether or not the distribution of $F_Y(y)$ is skewed at high and low values of y. We might then consider using the median as an estimate.

4.3.2 Using $m^{**}(.)$ as a surrogate for $\eta(.)$

One might consider estimating $F_Y(y)$ by

$$\hat{F}_Y(y) = \frac{1}{N}\sum_{j=1}^{N} I\{m^{**}(\mathbf{x}_j^*) \leq y\}. \tag{4.36}$$

This should result in a good estimate if $m^{**}(.)$ is a good approximation of $\eta(.)$, and is a simple way of obtaining an estimate. The difficulty lies in quantifying the uncertainty in the estimate. One could consider estimates of the distribution function based on extreme estimates of $\eta(.)$:

$$\hat{F}_Y^L(y) = \frac{1}{N}\sum_{j=1}^{N} I\{m^{**}(\mathbf{x}_j^*) + t_{\alpha,n-q}\hat{\sigma}\sqrt{c^{**}(\mathbf{x})} \leq y\} \tag{4.37}$$

$$\hat{F}_Y^U(y) = \frac{1}{N}\sum_{j=1}^{N} I\{m^{**}(\mathbf{x}_j^*) - t_{\alpha,n-q}\hat{\sigma}\sqrt{c^{**}(\mathbf{x})} \leq y\}, \tag{4.38}$$

where $t_{\alpha,n-q}$ is some appropriate percentage point from the t distribution with $n-q$ degrees of freedom. However, it is not clear how to interpret these bounds in terms of the quantities that we are interested in. The expressions (4.37) and (4.38) give the distribution function for $\alpha\%$ pointwise bounds of $\eta(\mathbf{X})$, when the real interest is in $\alpha\%$ pointwise bounds of the distribution function of $\eta(\mathbf{X})$.

If we are only interested in esimtating $P(Y \leq y_1)$ for a single output y_1 , we might consider the following estimate:

$$\hat{F}_Y(y_1) = \frac{1}{N}\sum_{j=1}^{N} I\{\phi(\mathbf{x}_j^*) \leq y_1\}, \tag{4.39}$$

where

$$\phi(\mathbf{x}) = \begin{cases} \eta(\mathbf{x}) & \text{if } |m^{**}(\mathbf{x}) - y_1| \leq t_{\alpha,n-q}\hat{\sigma}\sqrt{c^{**}(\mathbf{x})} \\ m^{**}(\mathbf{x}) & \text{otherwise} \end{cases}, \tag{4.40}$$

for some suitable percentile α of the t distribution with $n-q$ degrees of freedom. In this case, the only interest is in whether or not $\eta(\mathbf{x})$ will exceed y_1 for any particular $\mathbf{x}$. If for example, we are certain that $\eta(\mathbf{x})$ is significantly less than y_1, then it may not be necessary to determine the exact value of $\eta(\mathbf{x})$. Thus for each random value of $\mathbf{x}$, if we believe that $\eta(\mathbf{x})$ is far enough away from y_1 such that $P\{\eta(\mathbf{x}) \leq y_1\}$ is close to zero or one, then we do not need to know the exact value of $\eta(\mathbf{x})$ and we can approximate it by $m^{**}(\mathbf{x})$. Otherwise we evaluate $\eta(\mathbf{x})$ so that there is no error. In

addition, each time we evaluate $\eta(\mathbf{x})$, our uncertainty about $\eta(.)$ is reduced, $c^{**}(\mathbf{x})$ will decrease for all untested values of $\mathbf{x}$, and so the frequency with which we will have to evaluate $\eta(\mathbf{x})$ instead of approximating it by $m^{**}(\mathbf{x})$ will decrease.

4.4 Estimating the percentile function

In addition to inference about $F(.)$ we can consider the corresponding quantile function. Define p_α to be the 100α percentile, such that $F(p_\alpha) = \alpha$. The distribution of p_α is given by

$$P\{p_\alpha \leq y\} = P\{F(y) \geq \alpha\}. \tag{4.41}$$

We can estimate $P\{F(y) \geq \alpha\}$ using the simulation approach. We obtain a random sample $F_{(1)}(y), \ldots, F_{(N)}(y)$, and then estimate $P\{F(y) \geq \alpha\}$ by

$$\hat{P}\{F(y) \geq \alpha\} = \frac{1}{N}\sum_{i=1}^{N} I\{F_{(i)}(y) \geq \alpha\}, \tag{4.42}$$

where $I\{.\}$ denotes the indicator function. We can then estimate p_α by its median. Alternatively, we can find $p_{(i)\,\alpha}$,the α percentile for realisation i, for $i = 1 \ldots N$, and then estimate p_α by its sample mean.

4.5 Example: the ^{131}I algorithm

We now introduce a computer model that we will use as an example. The model is concerned with the behaviour of radioactive iodine in the human body. Iodine occurs naturally in the human body and tends to concentrate in the thyroid gland. If radioactive iodine, ^{131}I is taken into the body, it will also concentrate in the thyroid. It will undergo decay, which causes an increase in the risk of developing thyroid cancer. Exposure to ^{131}I could result from being in the vicinity of a nuclear accident. To assess the risk of an individual developing thyroid cancer, it is necessary to determine the dose of radiation received by the individual. A standard measure used is the committed effective dose equivalent (CEDE), which quantifies the detriment over a fifty year period following the initial exposure. However, it is difficult to measure the CEDE directly, and so a mathematical model must be used.

Various models have been developed to describe the movement of ^{131}I in the body. They all contain parameters whose exact values are unknown. The model we will use is an algorithm proposed by Adams and Fell (1988). A graphical representation is given in figure 4.2. Each box (or compartment) in the diagram represents

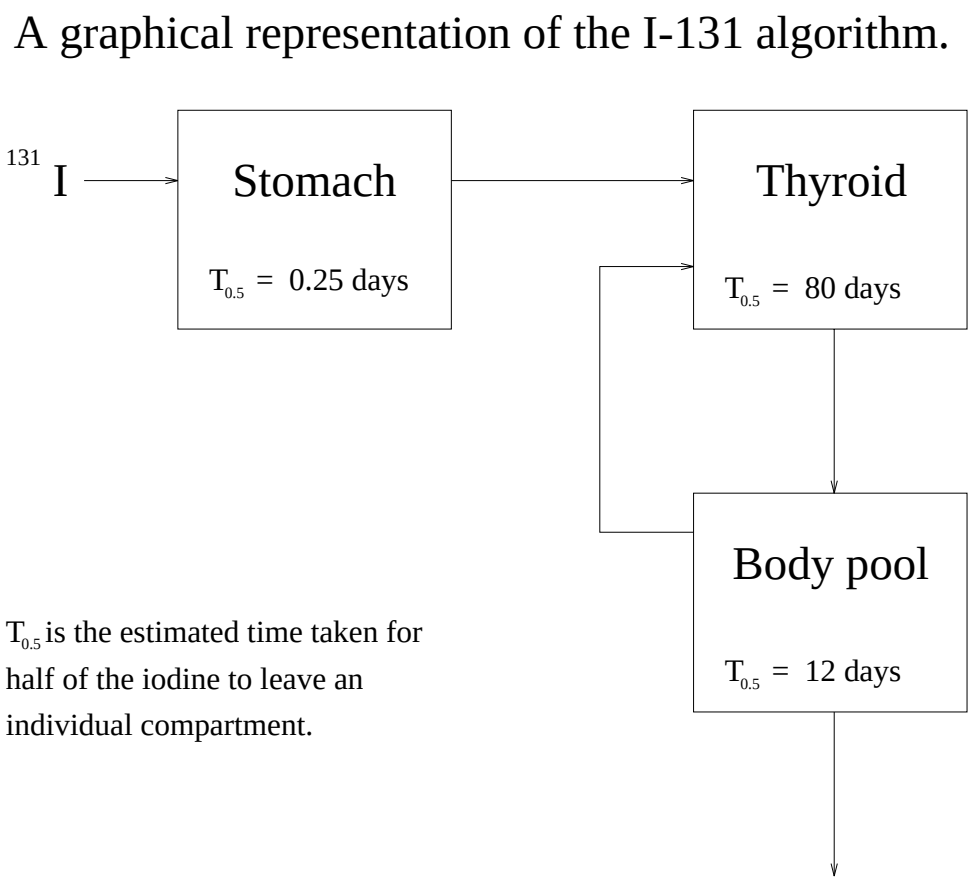

Figure 4.2: A graphical representation of the ^{131}I model

a part of the body, and each arrow has a transfer rate describing the flow of ^{131}I from one region to another. The quantity $T_{0.5}$ is the time taken for half the amount of ^{131}I to be removed from a particular compartment. In addition, the ^{131}I will be undergoing radioactive decay to form stable iodine. The radioactive half life for ^{131}I is approximately eight days.

It is assumed that a unit quantity of iodine is ingested. The model will then give the amount of ^{131}I in each of the three compartments at any time following ingestion. The CEDE can then be calculated from these values. We will consider the output of this model to be a single value, the CEDE. Two parameters are required

by the model which are not practical to measure. These are:

1. w, the mass of the thyroid gland
2. f, the fraction of ^{131}I contained in a unit quantity of blood that is taken up by the thyroid.

Hence for any individual, the output of the model run at the correct values of w and f will be unknown. After the development of the algorithm the NRPB carried out an analysis to determine the uncertainty in the output induced by the uncertainty in the model parameters. The ^{131}I algorithm is not computationally expensive, so we will conduct an uncertainty analysis using both the Bayesian approach and conventional Monte Carlo methods for comparison.

The first stage is to specify prior distributions for w and f. A study was carried out by Dunning et al. (1988), and log-normal distributions were found to be appropriate for both parameters. We have

$$\log w \sim N(2.889, \quad 0.463) \tag{4.43}$$

$$\log f \sim N(-1.315, \quad 0.355). \tag{4.44}$$

w and f are believed to be independent, and this completes the specification for $G(.)$.

4.5.1 The distribution function of the ^{131}I algorithm

We write the two inputs of the model to be $\log w$ and $\log f$, so that $\mathbf{x} = (\log w \quad \log f)^T$. As before, we denote the algorithm by $\eta(\mathbf{x})$. Haylock (1997) set

$$h(\mathbf{x}) = (1 \quad \log w \quad \log f)^T, \tag{4.45}$$

and used the correlation function

$$c(\mathbf{x}, \mathbf{x}') = \exp\{-(\mathbf{x} - \mathbf{x}')^T B(\mathbf{x} - \mathbf{x}')\}. \tag{4.46}$$

Note that the transformation to the log scale will alter the correlation structure. The correlation between the outputs at (w, f) and (w', f') now depends on $|\log w - \log w'|$

and $|\log f - \log f'|$. Haylock (1997) chose this form because it was believed that function might be less smooth for values of w and f near the origin. In addition, this transformation is appropriate if the output tends to some bound for large values of the input. In this case we would expect the outputs at two distinct large values of the input to be highly correlated, even if these two inputs themselves are far apart.

A weak prior distribution is used for $\boldsymbol{\beta}$ and σ^2. We evaluate $\eta(.)$ at nine inputs. A product design for the inputs is chosen, using a suggested design in O'Hagan (1991), for convenience. We use estimates of the smoothing parameters given in Haylock (1997), so

$$B = \begin{pmatrix} 0.7407 & 0 \\ 0 & 0.8696 \end{pmatrix}. \tag{4.47}$$

The posterior distribution of $\eta(.)$ can now be derived.

We first calculate the 'true' distribution function using Monte Carlo methods based on a sample of 100000 runs of the algotrithm. We then calculate the expected distribution function using the Bayesian approach based on nine runs. These are both plotted in figure 4.3, with the 'true' function shown as the dotted line. We have written 'dose' as the CEDE multiplied by 10^8. It can be seen that with only nine observations we have obtained an accurate estimate of the distribution function. To assess the uncertainty induced by the two unknown inputs, the user of the model must first decide what they believe a critical value of the dose is. A particular value could be considered 'critical' if the course of action to be taken following exposure will depend on whether or not the dose exceeds this particular value. Suppose for example that a dose around 2.25 is considered to be critical. Then examining figure 4.3, we can see that the expected probability of the model output when run at the correct inputs exceeding 2.25 is approximately 0.5. Hence the two unknown inputs have induced high uncertainty in the model output. Even if the model predicts the CEDE very accurately, the decision making process has been made considerably more difficult by the uncertainty in the inputs. Alternatively, suppose the critical value is around 8. We can now see that the probability of the model output exceeding 8 is small, and so little uncertainty has been induced by the unknown inputs.

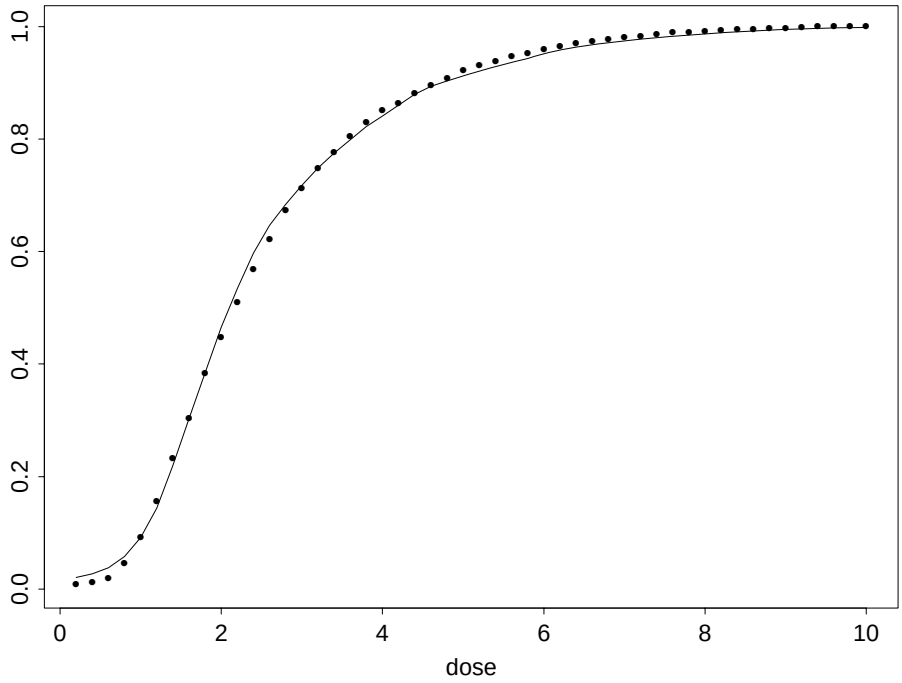

Figure 4.3: The expected distribution function

4.5.2 Using simulation

In addition to estimating the distribution function, we need to determine the uncertainty in the estimate. Due to the reasons discussed in section 4.2, this is difficult to do analytically. Using the simulation approach, it is relatively straightforward.

We must first choose the simulation design points. The initial nine design points are in a product design form. We also use a product design for the simulation design points. In each dimension, the initial (standardised) design points are of the form $(-x, 0, x)$. We choose two new points $x_1', -x_1'$ positioned plus and minus two standard deviations from the mean of the unknown input. We then find the input $x_2' \in (0, x)$ with the largest value of $c^{**}(x)$. This will give us inputs $-x_1', -x_2', x_2', x_1'$ in each of the two dimensions. The product design will give us sixteen new inputs.

We take a random sample of 1000 inputs $\mathbf{x}_1^*, \ldots, \mathbf{x}_{1000}^*$ from $G(\mathbf{x})$. For one randomly generated function $\eta_{(i)}(.)$, we estimate $F_Y(y)$ using (4.35). Recall that the output of the generated function $\eta_{(i)}(.)$ is only known at twenty five inputs (the initial nine design points and the sixteen simulation design points), and so we have

to approximate $\eta_{(i)}(.)$ by $m^{**}_{(i)}(.)$. We first need to check the accuracy that is lost in this approximation. In figure 4.4, we plot $\hat{F}_{Y_{(i)}}(y), \hat{F}^L_{Y_{(i)}}(y)$ and $\hat{F}^U_{Y_{(i)}}(y)$ as defined in subsection 3.2.3. It can be seen that all three distribution functions are very similar, and so there will be little error in the approximation.

We also need to check that there are no major numerical errors arising from ill conditioning. We simulate 1000 functions and compare the sample means of the outputs at the sixteen inputs with the known true means, as described in chapter three. For each simulation design point $\mathbf{x}'_i$ we calculate

$$Z = \frac{\sum_{j=1}^{1000} \left\{ \eta_{(j)}(\mathbf{x}'_i) - m^{**}(\mathbf{x}'_i) \right\}}{\sqrt{1000 \sigma^2 c^{**}(\mathbf{x}'_i)}}. \tag{4.48}$$

In table 4.1 we give the Z values for each $\mathbf{x}'_i$. We now obtain a sample of 1000

$\mathbf{x}'_i$	Z	$\mathbf{x}'_i$	Z
$1.833, -2.0464$	-1.32	$3.2207, -2.0464$	-1.77
$1.833, -1.5444$	-0.86	$3.2207, -1.5444$	0.30
$1.833, -1.0856$	-0.72	$3.2207, -1.0856$	-0.91
$1.833, -0.6016$	-0.13	$3.2207, -0.6016$	1.44
$2.5573, -2.0464$	-1.25	$3.8950, -2.0464$	-0.03
$2.5573, -1.5444$	1.37	$3.8950, -1.5444$	0.79
$2.5573, -1.0856$	-0.89	$3.8950, -1.0856$	0.46
$2.5573, -0.6106$	1.91	$3.8950, -0.6106$	1.04

Table 4.1: Checking for numerical errors arising from possible ill conditioning

values of $F_Y(y)$, and plot the median, 2.5 and 97.5 percentiles in figure 4.5. The true distribution is shown as the dotted line. Note that we may have underestimated the uncertainty about $F_Y(.)$, since we have not accounted for the uncertainty about B.

Another possible source of error here is simulation error; each generated distribution function is estimated using 1000 outputs, and the percentiles of the true distribution are estimated using a sample of 1000 generated distribution functions.

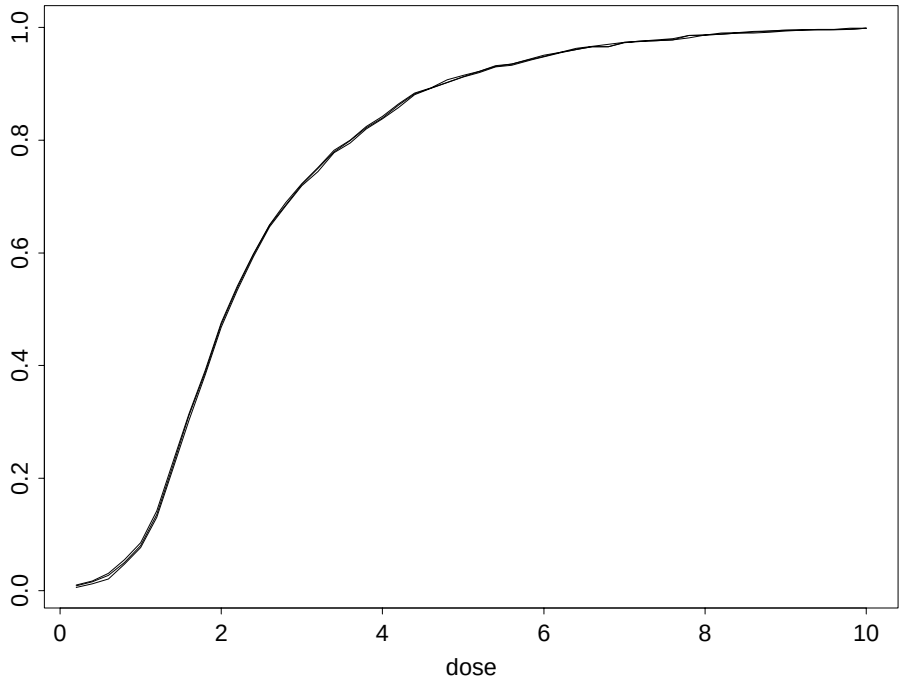

Figure 4.4: Examining the possible error in approximating $\eta_{(i)}(.)$ by $m^{**}_{(i)}(.)$

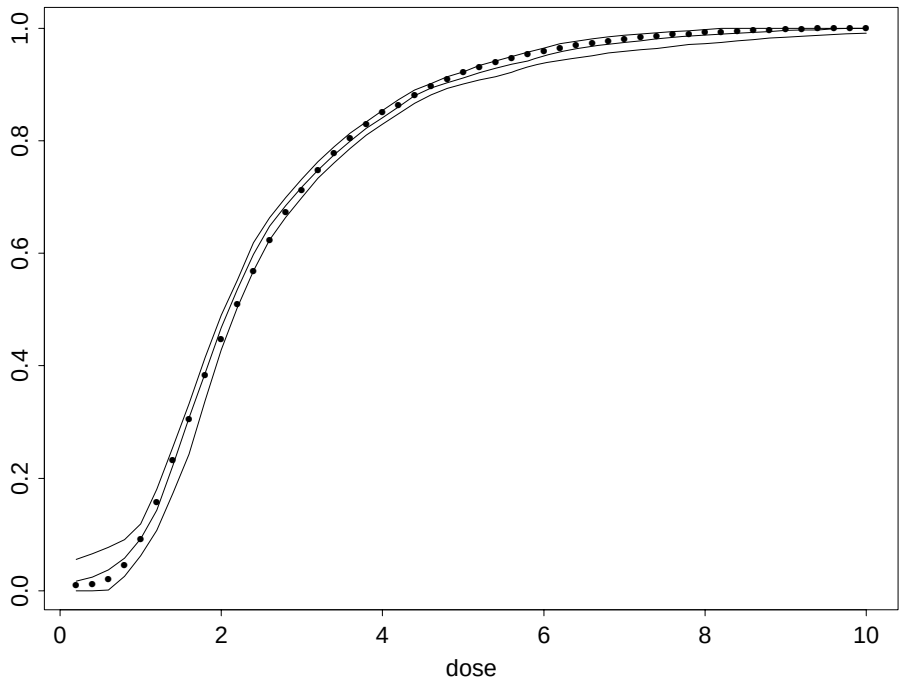

Figure 4.5: Percentiles of the distribution function

In practise, increasing the sample size in each case results has a minimal effect on the final estimates.

The simulation approach can be used to investigate the skewness of $F_Y(y)$ for extreme values of y. For $y = 0.4$ we obtain 1000 random values of $F_Y(y)$, and in figure 4.6 we plot the kernel density estimate of $F_Y(y)$. In figure 4.7 we plot the true distribution function (as the solid line), the mean distribution function (as the squares), and the median distribution function (as the circles), for low values of the dose. It can be seen that the median gives a fractionally better estimate, but the improvement is not large.

4.5.3 Other approaches to estimating $F_Y(y)$

We estimate the distribution function of the ^{131}I algorithm using the methods discussed in section 4.3. We first consider using $m^{**}(\mathbf{x})$ as a surrogate for $\eta(\mathbf{x})$. In figure 4.8 we plot the estimate given in (4.36), and the bounds given in (4.37) and (4.38). In (4.37) and (4.38) we use the 97.5 percentage point of the t distribution with 6 degrees of freedom. Note that the bounds are wider than the 95% interval for $F_Y(y)$. This highlights the difficulty in quantifying the uncertainty in this estimate.

We now investigate using the estimator suggested in (4.39), assuming we are only interested in $F_Y(y)$ for a single value of y. We consider a sample of 10000 inputs, and consider updating the distribution of $\eta(.)$ n times for $n = 0, 1, 2, 3$, when $\eta(\mathbf{x})$ has been evaluated. For three values of y and setting α to be 0.99, we show in table 4.2 how many times $\eta(\mathbf{x})$ would need to be evaluated.

n	$y = 1$	$y = 2$	$y = 5$
0	2306	3679	931
1	1867	2350	582
2	1225	989	306
3	1089	522	172

Table 4.2: Number of runs of $\eta(.)$ when estimating $F_Y(y)$ using (4.39) with updating

Observe that the number of runs required falls faster for $y = 2$ than it does for

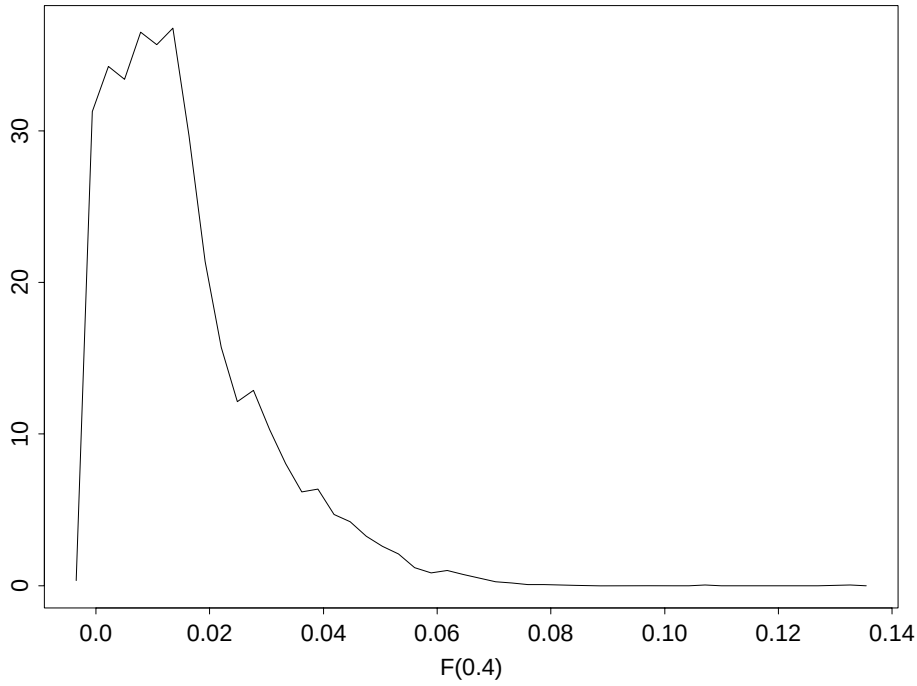

Figure 4.6: Density estimate of $F_Y(0.4)$

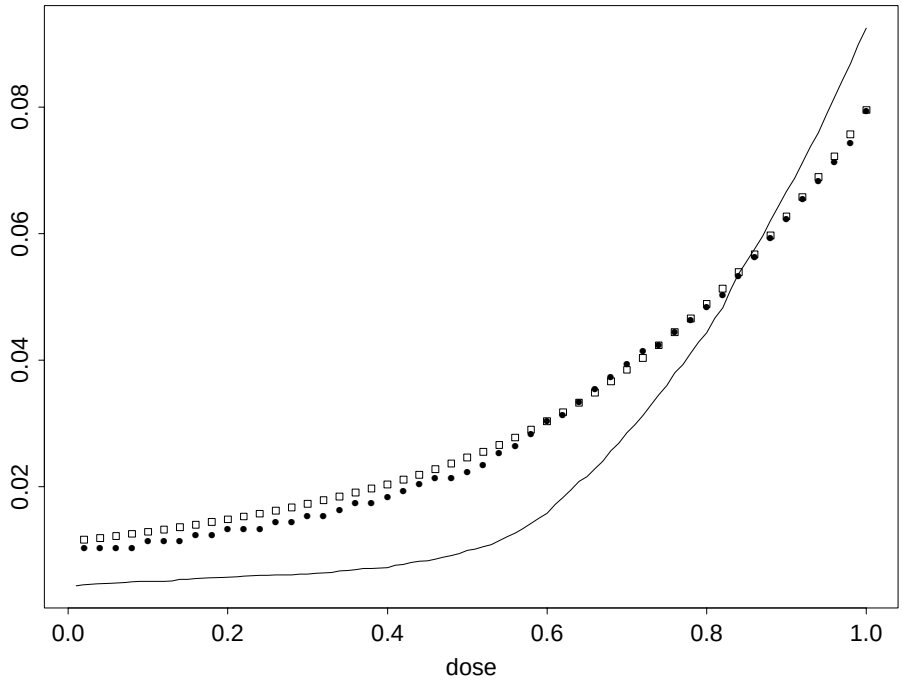

Figure 4.7: The mean, median and true distribution functions for low values of the dose

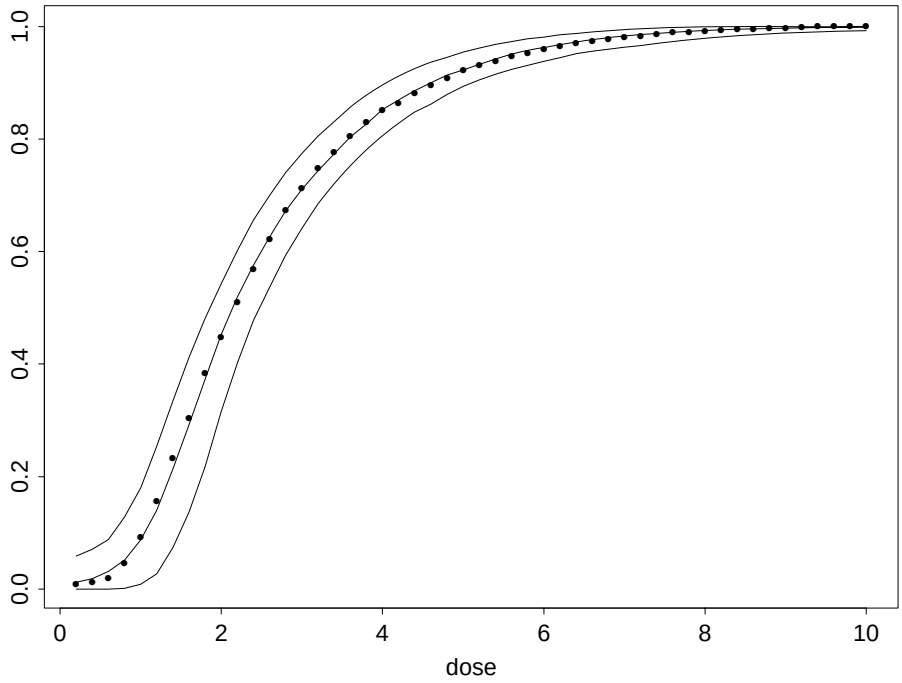

Figure 4.8: Using $m^{**}(x)$ to estimate $F_Y(y)$

$y = 1$. Suppose we obtain the estimate with no updating. There will be a subset of the 10000 inputs $\{\mathbf{x}_1^*, \ldots, \mathbf{x}_N^*\}$ where we have had to determine the true value of the output. If $\eta(\mathbf{x}_1^*)$ is highly correlated with the outputs in the set $\{\eta(\mathbf{x}_2^*), \ldots, \eta(\mathbf{x}_N^*)\}$, then we would expect it to be beneficial to update the distribution of $\eta(.)$ to include the observation $\eta(\mathbf{x}_1^*)$. For the 2306 inputs in the case $y = 1$, we find the average of the absolute values of the 2305 correlations is 0.64, and 4% of these absolute values are greater than 0.95. For the 3679 inputs in the case $y = 2$, we find the average of the absolute values of the 3678 correlations is 0.64, and 17% of these absolute values are greater than 0.95.

4.6 Comparison between the Bayesian and Monte Carlo approaches

The objective of the Bayesian approach is to obtain an accurate estimate of $F_Y(y)$ based on a small number of runs of the algorithm. We now consider how many

runs would be needed to obtain comparable accuracy using Monte Carlo methods; obtaining a random sample $\eta(\mathbf{x}_1), \ldots, \eta(\mathbf{x}_N)$ with $\mathbf{x}_1, \ldots, \mathbf{x}_N$ randomly drawn from $G(\mathbf{x})$, and estimating $F_Y(y)$ using the empirical distribution function. Given the true distribution function, we can calculate theoretical 95% pointwise intervals for Monte Carlo estimates of $F_Y(y)$ based on a sample of N observations. We use the normal approximation and calculate the bounds

$$F_Y(y) - 1.96\sqrt{\frac{1}{N}F_Y(y)\{1 - F_y(y)\}}, \tag{4.49}$$

and

$$F_Y(y) + 1.96\sqrt{\frac{1}{N}F_Y(y)\{1 - F_y(y)\}}. \tag{4.50}$$

Intervals for various sample sizes are shown in figure 4.9. For comparison, the 95% interval and median estimate for $F_Y(y)$ using the Bayesian approach are shown. It can be seen that the Bayesian estimate using nine runs is comparable in accuracy to the Monte Carlo estimate using 500 runs. For computationally expensive algorithms, the computing time is reduced considerably.

4.7 Conclusions

We have described a means of quantifying the uncertainty in a computer model output induced by unknown inputs. The plot of the estimated distribution function provides a graphical summary that can easily be interpreted by the user of the computer model. In the ^{131}I algorithm example, the estimate was obtained using a small number of runs of the code. We believe the Bayesian approach will offer substantial reductions in computing time when each run of the code is computationally expensive. In addition, we have provided means of quantifying our uncertainty about the distribution function, through its variance and its percentiles. The simulation method has been useful here, as limitations have been found with the analytic approach.

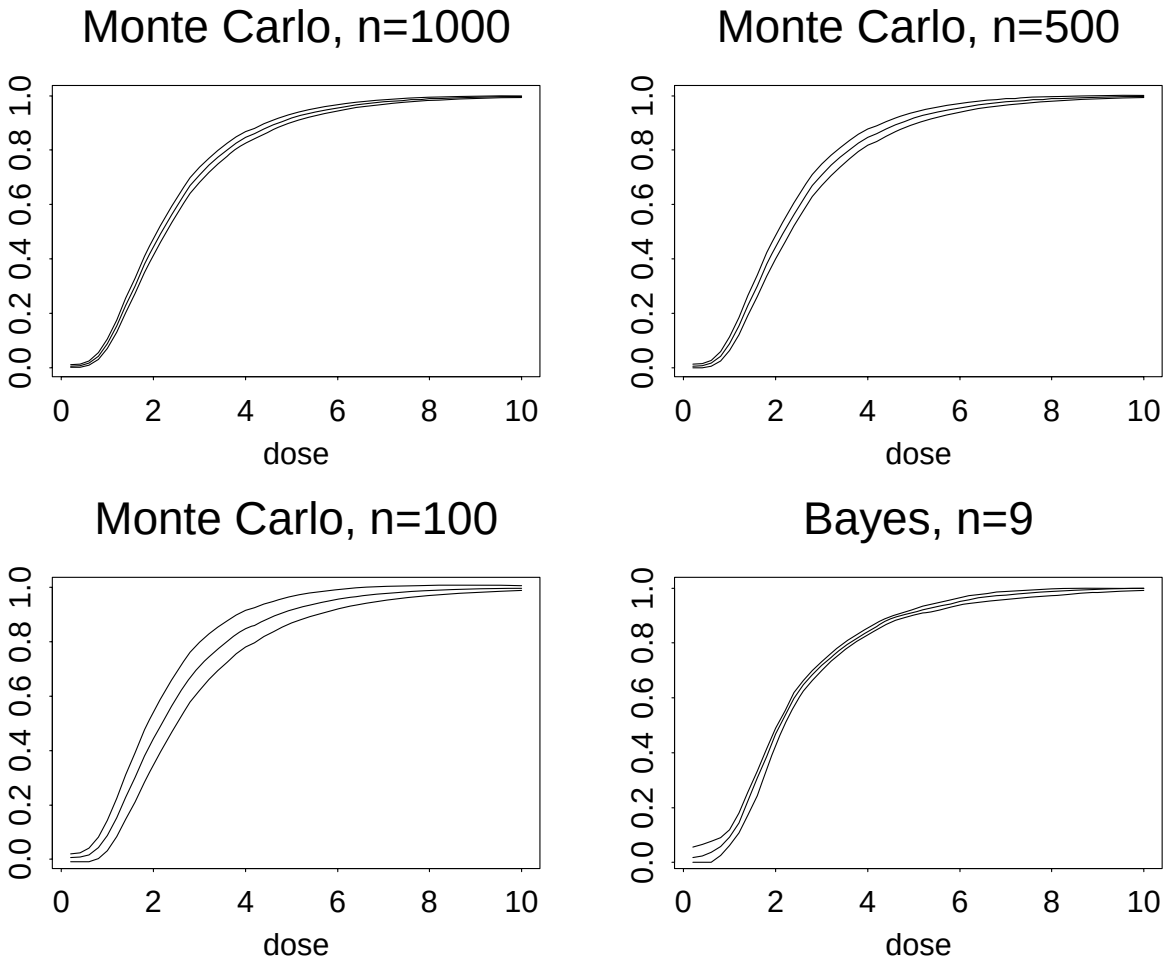

Figure 4.9: Monte Carlo and Bayesian estimates of $F_Y(y)$

Chapter 5

Estimating the density function

5.1 Introduction

In the previous chapter we estimated the distribution function $F_Y(y)$ as a means of quantifying the uncertainty in Y. We now turn our attention to the density function of Y, denoted by $f_Y(y)$ and defined by

$$P\{\eta(X) \leq y\} = \int_{\mathbf{X}} I\{\eta(x) \leq y\} dG(x) = \int_{-\infty}^{y} f_Y(t)\, dt. \tag{5.1}$$

Plotting an estimate of the density function of Y will provide a useful graphical summary of the uncertainty about Y, and may highlight features such as multimodality that would not necessarily be apparent from a plot of the distribution function.

As in the previous chapter, density estimation is itself a large area of interest. Various methods are described in Silverman (1986), and examples of the Bayesian approach are Escobar and West (1995), Hjort (1996) and Roeder and Wasserman (1997). For the same reasons as given in the previous chapter, a novel approach particular to this scenario is required here.

Differentiating (5.1) to derive the density function, we obtain

$$\begin{aligned} f_Y(y) &= \lim_{h \to 0} \frac{1}{h} \{F(y) - F(y-h)\} \\ &= \int_{\mathcal{X}} \lim_{h \to 0} \frac{1}{h} I\{y - h < \eta(\mathbf{x}) \leq y\} dG(\mathbf{x}) \end{aligned} \tag{5.2}$$

We can see that $f_Y(y)$ is a function of the random variable $\lim_{h\to 0} \frac{1}{h} I\{y-h < \eta(\mathbf{x}) \leq y\}$ and consequently, $f_Y(y)$ is also unknown.

The problem of inference about $f_Y(y)$ can be thought about in terms of a transformation of a random variable. We have a random variable $\mathbf{X}$, with a known distribution $G(\mathbf{x})$, and an unknown transformation of $\mathbf{X}$ given by $Y = \eta(\mathbf{X})$. The interest here is in the density of the transformed random variable. However, the density function of Y may not be defined everywhere depending on the derivative of $\eta(\mathbf{x})$. We illustrate this for a one-dimensional function $\eta(x)$.

Let X have the distribution function $F_X(x)$. We require the density of $Y = \eta(X)$. We consider a general function $\eta(x)$ and assume that the value of $\eta(x)$ is known for all x, and that $\eta(x_j) = c_j$, $\eta'(x_j) = 0$, and $\eta''(x_j) = k_1 < 0$. Consider the density of Y at c_j. For given $h > 0$, let $\varepsilon(h)$ satisfy

$$\eta(x_j + \varepsilon(h)) = c_j - h. \tag{5.3}$$

Now

$$\eta(x_j + \varepsilon(h)) = \eta(x_j) + \eta'(x_j)\varepsilon(h) + \frac{1}{2}\eta''(x_j)\varepsilon(h)^2 + O\{\varepsilon(h)^3\}, \tag{5.4}$$

$\Rightarrow$

$$c_j - h = c_j + 0 + \frac{k_1}{2}\varepsilon(h)^2 + O\{\varepsilon(h)^3\}, \tag{5.5}$$

$\Rightarrow$

$$\varepsilon(h) \simeq \pm\sqrt{\frac{-2h}{k_1}}, \tag{5.6}$$

for sufficiently small h. Then

$$\begin{aligned}
f_Y(c_j) &= \lim_{h\to 0} \frac{1}{h} P(c_j - h < Y \leq c_j) && (5.7)\\
&\geq \lim_{h\to 0} \frac{1}{h} \{F_X(x_j) - F_X(x_j - \varepsilon(h))\} && (5.8)\\
&\simeq \lim_{h\to 0} \frac{1}{h} f_X(x_j)\, \varepsilon(h) && (5.9)\\
&= \lim_{h\to 0} \frac{1}{h} f_X(x_j) \sqrt{\frac{-2h}{k_1}} && (5.10)\\
&= f_X(x_j) \sqrt{\frac{-2}{k_1}} \lim_{h\to 0} \frac{1}{\sqrt{h}}, && (5.11)
\end{aligned}$$

which will be infinite. We can also write

$$f_Y(c_j) = \lim_{h\to 0} \frac{1}{h} P(c_j \leq Y < c_j + h). \tag{5.12}$$

If in addition $\eta(x) \neq c_j \quad \forall x \neq x_j$, so that c_j is the maximum of $\eta(x)$, then we have $P(c_j \leq Y < c_j + h) = 0$ and $\lim_{h\to 0} \frac{1}{h} P(c_j \leq Y < c_j + h)$ is finite. Hence the density at c_j is not defined. Alternatively, c_j may just be a local maximum. Suppose for example we have $\eta(x_j^*) = c_j$ with $\eta'(x_j^*) > 0$, then

$$\lim_{h\to 0} \frac{1}{h} P(c_j \leq Y < c_j + h) = \lim_{h\to 0} \left\{ \frac{1}{h} f_X(x_j) \sqrt{\frac{-2}{k_1}} + \frac{f_X(x_j^*)}{|\eta'(x_j^*)|} \right\}, \tag{5.13}$$

and

$$\lim_{h\to 0} \frac{1}{h} P(c_j - h < Y \leq c_j) = \lim_{h\to 0} \left\{ 0 + \frac{f_X(x_j^*)}{|\eta'(x_j^*)|} \right\}, \tag{5.14}$$

and so the density is also undefined at c_j. The other case to consider is when $\eta(x_j) = c_j$ is a point of inflexion. Now we have $\eta''(x_j) = 0$ and we suppose that $\eta'''(x_j) = k_2$ where $k_2 > 0$. Then

$$\lim_{h\to 0} \frac{1}{h} P(c_j - h < Y \leq c_j) = \lim_{h\to 0} \left\{ \frac{1}{h} f_X(x_j) \sqrt[3]{\frac{6h}{k_2}} \right\}, \tag{5.15}$$

and

$$\lim_{h\to 0} \frac{1}{h} P(c_j \leq Y < c_j + h) = \lim_{h\to 0} \left\{ \frac{1}{h} f_X(x_j) \sqrt[3]{\frac{6h}{k_2}} \right\}, \tag{5.16}$$

and the density at c_j is infinite.

Returning to the uncertainty analysis context where the value of $\eta(x)$ is only known at a small sample of inputs, then depending on our observed data $\eta(x_1)$, $\eta(x_2)$, $\ldots$, $\eta(x_n)$, (assuming that $x_1 < x_2 < \ldots < x_n$) we may either have certainty that $\eta'(x) = 0$ at some value x, or at least non-zero probability that this is case, since if $\eta(x_j) < \eta(x_{j+1})$, then for all $x \in (x_j,\, x_{j+1})$ it is true that $P\{\eta(x) < \eta(x_j)\} > 0$, and similarly if $\eta(x_j) > \eta(x_{j+1})$.

5.2 Posterior moments

We first derive the expected density function. We have

$$\begin{aligned} E\{f_Y(y)\} &= E\left[\lim_{h\to 0} \frac{1}{h} \int_{\mathcal{X}} I\{y - h < \eta(\mathbf{x}) \leq y\} dG(\mathbf{x})\right] \\ &= \int_{\mathcal{X}} \lim_{h\to 0} \frac{1}{h} E\left[I\{y - h < \eta(\mathbf{x}) \leq y\}\right] dG(\mathbf{x}) \\ &= \lim_{h\to 0} \frac{1}{h} \int_{\mathcal{X}} P\{y - h < \eta(\mathbf{x}) \leq y\} dG(\mathbf{x}) \\ &= \int_{\mathcal{X}} f_{\eta(\mathbf{x})}(y) dG(\mathbf{x}), \end{aligned} \tag{5.17}$$

where $f_{\eta(\mathbf{x})}(y)$ denotes the density function of $\eta(\mathbf{x})$. For the covariance we need

$$E\left\{f_Y(y_1)f_Y(y_2)\right\} = E\left\{\lim_{h_1\to 0}\frac{F(y_1)-F(y_1-h_1)}{h_1}\lim_{h_2\to 0}\frac{F(y_2)-F(y_2-h_2)}{h_2}\right\}$$

$$= \lim_{h_1\to 0}\lim_{h_2\to 0}\frac{1}{h_1h_2}E\left\{\begin{array}{c} F(y_1)F(y_2)+F(y_1-h_1)F(y_2-h_2) \\ -F(y_1)F(y_2-h_2)-F(y_2)F(y_1-h_1)\end{array}\right\}$$

$$= \lim_{h_1\to 0}\lim_{h_2\to 0}\frac{1}{h_1h_2}\int_X\int_X \left[\begin{array}{c} \int_{-\infty}^{y_2-h_2} P\left\{\eta(\mathbf{x})\le y_1-h_1 \mid \eta(\mathbf{z})=k\right\} f_{\eta(\mathbf{z})}(k)dk \\ +\int_{-\infty}^{y_2} P\left\{\eta(\mathbf{x})\le y_1 \mid \eta(\mathbf{z})=k\right\} f_{\eta(\mathbf{z})}(k)dk \\ -\int_{-\infty}^{y_2-h_2} P\left\{\eta(\mathbf{x})\le y_1-h_1 \mid \eta(\mathbf{z})=k\right\} f_{\eta(\mathbf{z})}(k)dk \\ -\int_{-\infty}^{y_2} P\left\{\eta(\mathbf{x})\le y_1-h_1 \mid \eta(\mathbf{z})=k\right\} f_{\eta(\mathbf{z})}(k)dk\end{array}\right] dG(\mathbf{x})\,dG(\mathbf{z})$$

$$= \lim_{h_1\to 0}\lim_{h_2\to 0}\frac{1}{h_1h_2}\int_X\int_X \left[\begin{array}{c} \int_{-\infty}^{y_2-h_2} P\left\{y_1-h_1<\eta(\mathbf{x})\le y_1 \mid \eta(\mathbf{z})=k)\right\} f_{\eta(\mathbf{z})}(k)dk \\ -\int_{-\infty}^{y_2} P\left\{y_1-h_1<\eta(\mathbf{x})< y_1 \mid \eta(\mathbf{z})=k\right\} f_{\eta(\mathbf{z})}(k)dk\end{array}\right] dG(\mathbf{x})\,dG(\mathbf{z})$$

$$= \lim_{h_1\to 0}\lim_{h_2\to 0}\frac{1}{h_1h_2}\int_X\int_X \int_{y_2-h_2}^{y_2} P\left\{y_1-h_1<\eta(\mathbf{x})\le y_1 \mid \eta(\mathbf{z})=k\right\} f_{\eta(\mathbf{z})}dk\,dG(\mathbf{x})\,dG(\mathbf{z})$$

$$= \int_X\int_X f_{\eta(\mathbf{x}),\eta(\mathbf{z})}(y_1,y_2)\,dG(\mathbf{x})\,dG(\mathbf{z}). \tag{5.18}$$

Examining (5.17), it is can be seen that the integrand will be infinite at certain inputs $\mathbf{x}$ if we are considering an output y that has been observed, since we will have certainty that $\eta(\mathbf{x}) = y$ for some input $\mathbf{x}$ and so $f_{\eta(\mathbf{x})}(y)$ will be infinite. We now prove that if the input is one dimensional, this has the result that the expected density is infinite at any observed output.

Suppose we have observed $\eta(x_j) = y_j$ and consider the expected density of $\eta(X)$ at y_j, which is given by

$$E\{f_{\eta(X)}(y_j)\} = k_1\int_{\mathcal{X}}\frac{1}{\sqrt{c^{**}(x)}}\left[1+\frac{\{y_j-m^{**}(x)\}^2}{(n-q)\hat{\sigma}^2c^{**}(x)}\right]^{-\frac{1}{2}(n-q+1)} dG(x), \tag{5.19}$$

where k_1 is some constant. We consider the integrand in the neighbourhood of x_j. We have

$$E\{f_{\eta(X)}(y_j)\} \;\ge\; k_1\int_0^{\varepsilon}\frac{1}{\sqrt{c^{**}(x_j+\delta)}}\times$$

$$\left[1+\frac{\{y_j-m^{**}(x_j+\delta)\}^2}{(n-q)\hat{\sigma}^2c^{**}(x_j+\delta)}\right]^{-\frac{1}{2}(n-q+1)} f_X(x^*)\,d\delta, \qquad (5.20)$$

where $f_X(x^*)$ is the minimum value of the density of x for x in the range $(x_j, x_j+\varepsilon)$. In chapter four we showed that for sufficiently small δ, $m^{**}(x_j+\delta)-\eta(x_j)$ is of order δ, and $c^{**}(x_j+\delta)$ is of order δ^2. It then follows that

$$\begin{aligned} E\{f_{\eta(X)}(y_j)\} &\geq k_1\int_0^{\varepsilon}\frac{1}{\sqrt{\delta^2k_2}}\times \\ &\qquad \left[1+\frac{(\delta k_3)^2}{(n-q)\hat{\sigma}^2\delta^2k_2}\right]^{-\frac{1}{2}(n-q+1)} f_X(x^*)\,d\delta \qquad (5.21)\\ &= k_4\int_0^{\varepsilon}\frac{1}{\delta}\,d\delta, \qquad (5.22)\end{aligned}$$

for some constants k_2, k_3 and k_4. We can see that the integrand in (5.19) behaves like $\frac{1}{x-x_j}$ for x in the neighborhood of x_j, and so the integral diverges.

For higher dimensional inputs, it appears that the expected density is finite at observed outputs. This is an area for futher investigation, and we do not give a complete proof here. In chapter four we showed that if $\eta(\mathbf{x}_j)=c_j$ is known, then $m^{**}(\mathbf{x}_j+\delta\mathbf{x}_0)-c_j$ is of order δ for any vector $\mathbf{x}_0$. Hence $f_{\eta(\mathbf{x}_j+\delta\mathbf{x}_0)}(c_j)$ is of order $\frac{1}{\delta}$ for any vector $\mathbf{x}_0$. Suppose for example that $\mathbf{x}$ is two dimensional. Then for $\mathbf{x}$ close to $\mathbf{x}_j$ the integral is of the form

$$\int_{R_1}\int_{R_2}\frac{1}{\sqrt{x^2+y^2}}dxdy, \qquad (5.23)$$

where R_1 and R_2 both contain zero. Transforming to polar coordinates so that $x=r\cos\theta$ and $y=r\sin\theta$, this integral becomes

$$\int_{R_1'}\int_{R_2'} drd\theta, \qquad (5.24)$$

which is finite.

5.3 Alternative approaches to estimating the density function

In the case of a one dimensional input, we may know that the density function is infinite at certain outputs y, if the data $\mathbf{y}$ shows $\eta(.)$ to be non-monotone, but not

necessarily at the outputs where $E\{f_Y(y)\}$ is infinite. The expectation of the density function may give a good estimate of $f_Y(y)$ for non-observed values of y, but as there will be spikes in the plot at each observed output, the plot may not provide a good graphical summary of the uncertainty in Y. We now consider various methods for dealing with this problem.

5.3.1 Introducing artificial measurement error

We can smooth out the spikes in the density plot by including a small measurement error at each tested input in the data $\mathbf{y}$, so that the posterior variance at a tested input $\mathbf{x}_j$ of $\eta(\mathbf{x}_j)$ is non zero. We write

$$\mathbf{y} = \begin{pmatrix} \eta(\mathbf{x}_1) \\ \vdots \\ \eta(\mathbf{x}_n) \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \tag{5.25}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent and have a normal distribution with mean zero and variance ε^2, with ε^2 fixed at some small value. Then

$$\mathbf{y} \sim N(H\boldsymbol{\beta}, \sigma^2 A + \varepsilon^2 I). \tag{5.26}$$

However, it is not possible to proceed analytically from this point. If we use the weak prior for σ^2, given by $p(\sigma^2) \propto \sigma^{-2}$ then the posterior density of σ^2 is given by

$$p(\sigma^2|\mathbf{y}) \propto \frac{1}{\sigma^2} |V|^{-1} \exp\left\{-\frac{1}{2}(\mathbf{y} - H\hat{\boldsymbol{\beta}})^T V^{-1} (\mathbf{y} - H\hat{\boldsymbol{\beta}})\right\}, \tag{5.27}$$

where

$$V = \{\sigma^2 A + \varepsilon^2 I\}^{-1} \tag{5.28}$$

When we combine $\eta(.)|\sigma^2, \mathbf{y}$ with $\sigma^2|\mathbf{y}$, we cannot derive $\eta(.)|\mathbf{y}$ analytically.

Instead, we introduce an error proportional to σ^2, so that

$$\mathbf{y} \sim N\{H\boldsymbol{\beta}, \sigma^2(A + \varepsilon^2 I)\}. \tag{5.29}$$

We can now just replace A by $A + \varepsilon^2 I$ in the expressions for the posterior mean and variance of $\eta(\mathbf{x})$.

We illustrate this with the function

$$\eta(x) = 5 + 1.001x + \cos(x), \tag{5.30}$$

(We have adjusted the coefficient of x from previous examples so that the density of $\eta(x)$ is defined everywhere). We again have $X \sim N(0, 4)$, and as before we set $h(x)^T = (1 \quad x)$ and $c(x, x') = \exp\{-\frac{1}{2}(x - x')^2\}$. We evaluate $\eta(x)$ at five inputs and obtain the five outputs $\mathbf{y}^T = (0.2922, 2.4793, 6.0, 6.5914, 8.9689)$. In figure 5.1 we plot the expected density with measurement error $\varepsilon^2 = 0.001$, and in figure 5.2 $\varepsilon^2 = 0.01$. The true density function is shown as the dotted line.

As expected, when comparing figures 5.1 and 5.2 we can see that increasing the error term has increased the smoothness of the density plot. However, increasing the error term also leads to a flatter density plot, and the true spike at $y = 6$ is not captured as well.

5.3.2 Observing derivatives

In some cases, the computer model is able to return the derivative $\eta'(\mathbf{x})$ in addition to the output $\eta(\mathbf{x})$ (see Mitchell et al., 1993). The expected density of Y at y_j can be made finite if we also observe the derivative of $\eta(x)$ at $x = x_j$, and the derivative is non-zero. Suppose that $\eta'(x_j) = d_j$, so that we have data $\mathbf{y}^T = \{\eta(x_1), \ldots, \eta(x_n), \eta'(x_j)\}$. It then follows that for sufficiently small δ

$$m^{**}(x_j + \delta) - y_j = \delta d_j + O(\delta^2) \tag{5.31}$$

We again consider the variance of $\eta(x)$ in the neighbourhood of x_j. It is assumed that the correlation function is of the form $c(x, x') = \exp\{-b(x - x')^2\}$. We have

$$\begin{aligned} t(x_j + \delta)^T \quad A^{-1} \quad t(x_j + \delta) = \{t(x_j)^T + \delta t'(x_j)^T + \frac{\delta^2}{2} t''(x_j)^T + \frac{\delta^3}{6} t'''(x_j)^T + O(\delta^4)\} \\ \times A^{-1} \{t(x_j) + \delta t'(x_j) + \frac{\delta^2}{2} t''(x_j) + \frac{\delta^3}{6} t'''(x_j) + O(\delta^4)\}. \end{aligned} \tag{5.32}$$

Note that when $x = x'$

$$\frac{d^2}{dxdx'} c(x, x') = b, \tag{5.33}$$

$$\frac{d^2}{dx^2} c(x, x') = -2b, \tag{5.34}$$

$$\frac{d^3}{dx^3} c(x, x') = 0. \tag{5.35}$$

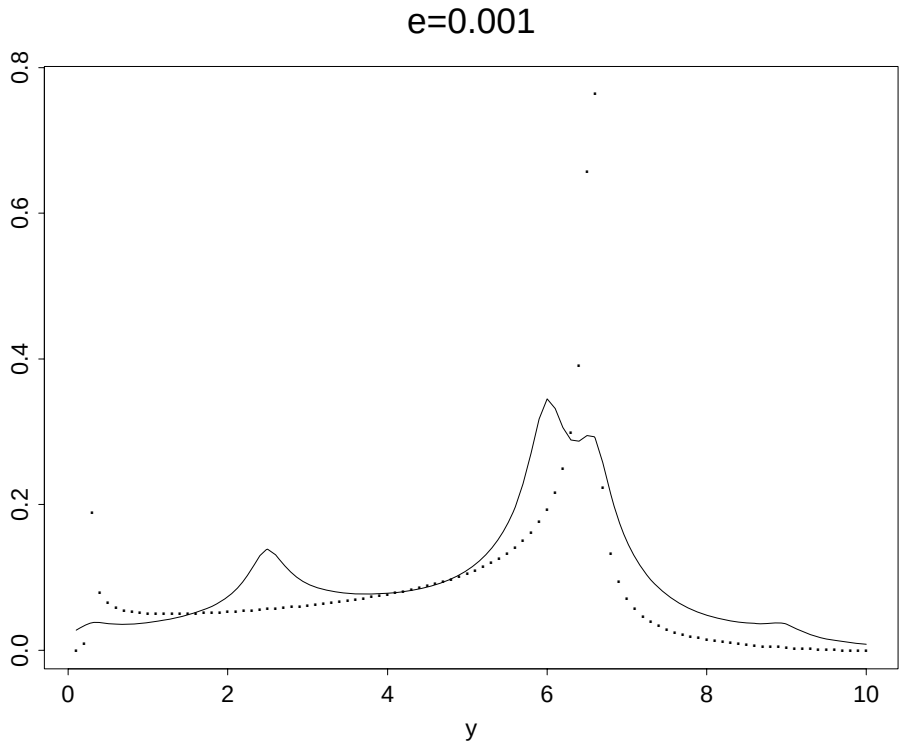

Figure 5.1: Density estimates with measurement error

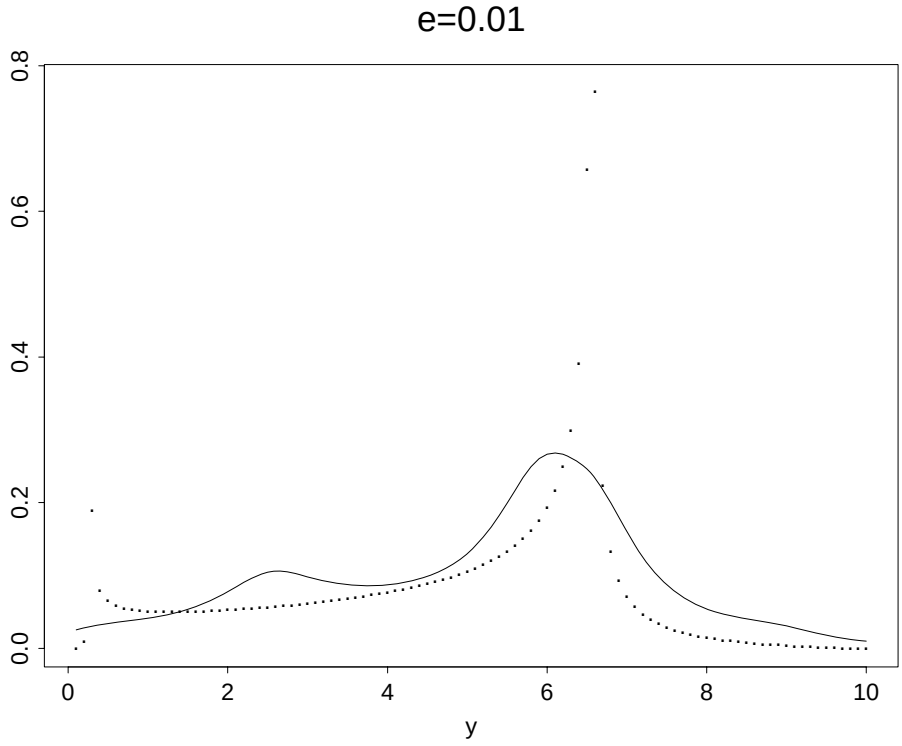

Figure 5.2: Density estimates with measurement error

Noting also that $t'(x_j)$ is the $n+1$th column of A, we have

$$\begin{aligned} t(x_j)^T A^{-1} t(x_j) &= 1, && (5.36)\\ \delta^2 t'(x_j)^T A^{-1} t'(x_j) &= b\delta^2, && (5.37)\\ \delta^2 t''(x_j)^T A^{-1} t(x_j) &= -2b\delta^2, && (5.38)\\ \delta^3 t'''(x_j)^T A^{-1} t(x_j) &= 0. && (5.39) \end{aligned}$$

We can then write

$$c(x_j+\delta, x_j+\delta) - t(x_j+\delta)^T A^{-1} t(x_j+\delta) = k_4\delta^4 + O(\delta^5), \tag{5.40}$$

for some constant k_4.

In addition we have

$$\begin{aligned} h(x_j+\delta)^T - t(x_j+\delta)^T A^{-1} H &= h(x_j)^T + \delta h'(x_j)^T - t(x_j)^T A^{-1} H \\ &\quad -\delta t'(x_j)^T A^{-1} H + O(\delta^2) && (5.41)\\ &= h(x_j)^T + \delta h'(x_j)^T - h(x_j)^T \\ &\quad -\delta h'(x_j)^T + O(\delta^2), && (5.42) \end{aligned}$$

and it can be seen that

$$c^{**}(x_j+\delta) = k_5\delta^4 + O(\delta^5), \tag{5.43}$$

for some constant k_5. Given that $n-q>2$, the integrand in (5.19) will tend to some finite limit as x tends to x_j, and so the expected density is finite at $y=y_j$

This gives us further insight into the distribution of the density function at observed outputs. Suppose we have observed an output $\eta(x_j)=y_j$, but not the corresponding derivative $\eta'(x_j)$. We know that conditional on the derivative, the expected density of Y at y_j is finite, assuming the derivative is non-zero. This implies that conditional on a non-zero derviative, the probability that the density of Y at y_j is infinite is zero. Then if we define E to be the event that the density of Y at y_j is infinite, we have

$$\begin{aligned} P(E) &= \int_{-\infty}^{\infty} P\{E|\eta'(x_j)=a\} f_{\eta'(x_j)}(a)da \\ &= \int_{0}^{0} P\{E|\eta'(x_j)=a\} f_{\eta'(x_j)}(a)da \\ &= 0. && (5.44) \end{aligned}$$

(Recall that $\eta'(x_j)$ has a Student-t distribution, and so $P\{\eta'(x_j) = 0\} = 0$).

Hence the distribution of the density at y_j must be sufficiently heavy tailed for the integral

$$\int_{-\infty}^{\infty} y f_Y(y) dy \tag{5.45}$$

to diverge.

In figure 5.3 we show the expected density function in the one dimensional example when the five derivatives have been observed in addition to the five outputs. The true density function is shown as the dotted line. The additional information from the derivatives has resulted in a reasonably good density estimate.

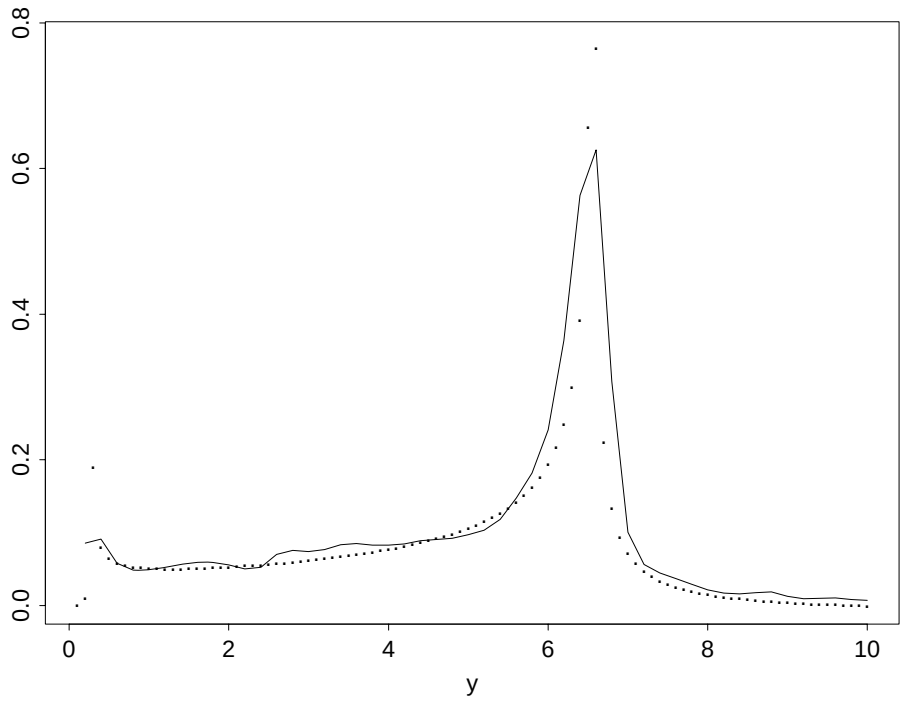

Figure 5.3: The expected density function when derivative information is available

5.4 Density estimation via the simulation approach

We proceed on the basis that estimating the density of $Y_{(i)} = \eta_{(i)}(\mathbf{X})$ for a cheap function $\eta_{(i)}(.)$ is relatively straightforward. We draw a large sample of inputs $\mathbf{x}_1^*, \ldots, \mathbf{x}_k^*$

from G, estimate $\eta_{(i)}(\mathbf{x}_j^*)$ by $m_{(i)}^{**}(\mathbf{x}_j^*)$, for $j = 1, \ldots, k$, and estimate the density function $f_{\eta_{(i)}(\mathbf{X})}(.)$ by kernel density estimation. Repeating this a large number of times will give us a sample of density functions $f_{Y_{(1)}}(.), \ldots, f_{Y_{(L)}}(.)$, for some large value of L. We can then consider estimating $f_Y(y)$ by the median of $f_{Y_{(1)}}(y), \ldots, f_{Y_{(L)}}(y)$. Other percentiles can be used to give bounds for $f_Y(y)$. This will be preferable to using the sample mean in the case when we know that the expected density is infinite at the observed outputs, since we know that the density is finite with probability one.

A difficulty is that the median density function will not necessarily integrate to one. As an example, suppose we had randomly drawn three density functions which are all uniform: $U[0,1]$, $U[\frac{1}{4}, \frac{3}{4}]$ and $U[\frac{1}{3}, \frac{2}{3}]$. The median density function of these three functions is given by

$$f_X(x) = \begin{cases} 0 & x < \frac{1}{4} \\ 1 & \frac{1}{4} \leq x < \frac{1}{3} \\ 2 & \frac{1}{3} \leq x < \frac{2}{3} \\ 1 & \frac{2}{3} \leq x < \frac{3}{4} \\ 0 & x \geq \frac{3}{4}, \end{cases} \tag{5.46}$$

which integrates to $\frac{5}{6}$. The median of the three density functions is shown as the bold line in figure 5.4.

5.4.1 Non-decreasing functions

In some cases we may have prior knowledge that the function $\eta(.)$ is monotonic. We can include this knowledge by rejecting any generated function $\eta_{(i)}(.)$ from the sample which is not monotonic. If $\eta_{(i)}(.)$ has been sampled at enough inputs, then the variance of $\eta'_{(i)}$ is small and we can approximate $\eta'_{(i)}(.)$ by $\frac{d}{dx}m^{**}(.)$ to establish monotonicity. We illustrate this with the one dimensional example in (5.30). In figure 5.5 graph a we show the pointwise 5th, 50th and 95th percentile of the density function, obtained using the simulation approach. The dotted line shows the true density functon. In graph b we use the simulation procedure as before, but reject any non-monotone functions from the sample. This can be seen to lead to an

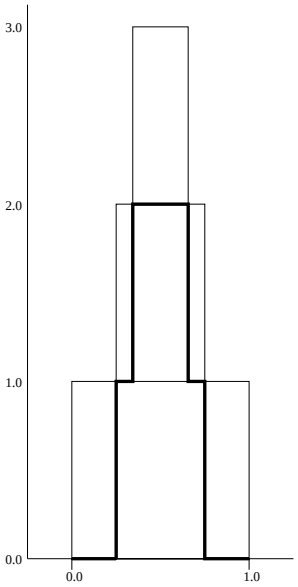

Figure 5.4: The median of three density functions

improved esimtate of the density function. However, this approach can require intensive computation, as the proportion of rejected functions may be large.

5.4.2 Investigating the density of $f_Y(y)$

We can use the simulation approach to explore the density function of $f_Y(y)$. The first step is to obtain a sample of values $f_{Y_{(1)}}(y), \ldots, f_{Y_{(N)}}(y)$ using simulation. We can then use kernel density estimation on this sample to estimate the density function of $f_Y(y)$. We have already noted that for outputs y_j that have been observed, $E\{f_Y(y_j)\}$ is infinite, even if $f_Y(y_j)$ is finite. In figure 5.6 we show density estimates of $f_Y(4)$ and $f_Y(6)$ in the one dimensional example, noting that we have observed $\eta(0) = 6$, and that the output $y = 4$ has not been observed. The density function of $f_Y(6)$ is the dotted line. As expected, the density function of $f_Y(6)$ is fairly flat compared to the density of $f_Y(4)$. In figure 5.7 we compare the density estimates of $f_Y(6)$ in the two cases when $\eta'(0)$ is known (the solid line) and when $\eta'(0)$ is unknown (the dotted line). We can now see that learning the derivative of $\eta(0)$ has resulted in a sharper density function, and so the expected value of $f_Y(6)$ will now be finite.

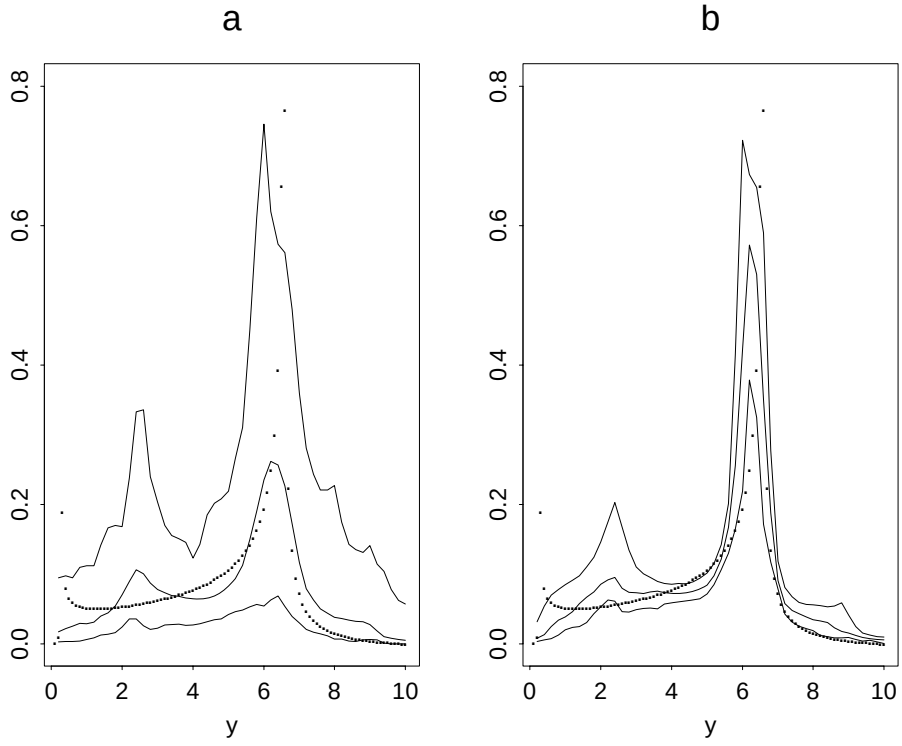

Figure 5.5: Density estimates using monotone functions

5.4.3 Example: the ^{131}I algorithm

We now return to the ^{131}I algorithm, and estimate the density function of the true output. We use the simulation procedure to obtain percentiles of the density function, and the kernel density estimates required during the procedure are performed using SPLUS. In figure 5.8 we plot the 5th, 50th and 95th percentiles of the distribution of the density function, in addition to the 'true' density function, shown by the dotted line. Again, by fixing B we have not accounted for the uncertainty in B, and so the uncertainty in the density function may have been underestimated. In figure 5.9 we plot the median density function as the solid line, and the mean density function as the dotted lines. The two functions are virtually indistinguishable, which suggests that the integral of the median density function will be very close to one.

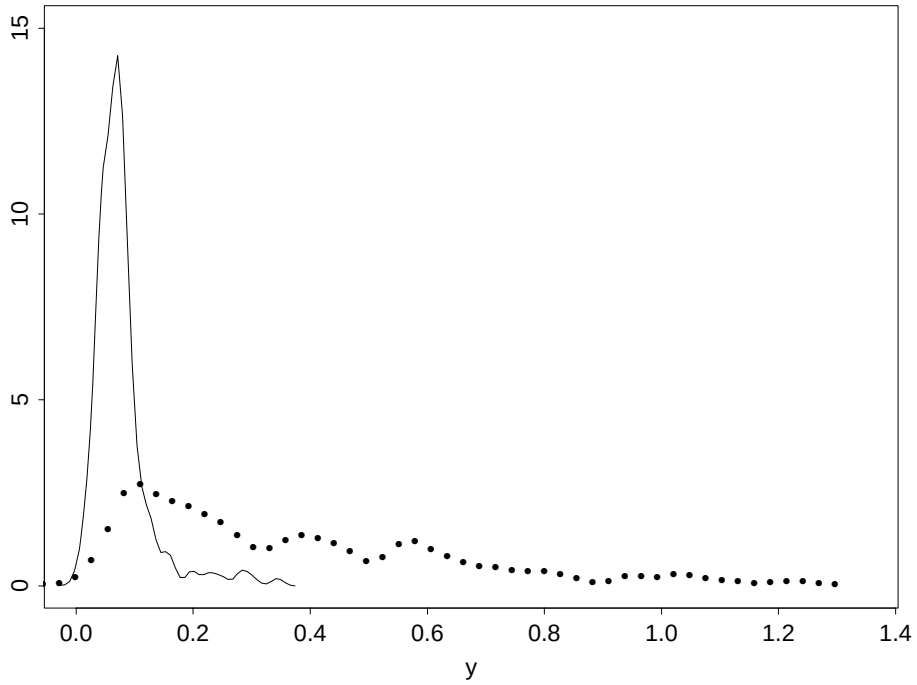

Figure 5.6: Density estimates of $f_Y(4)$ and $f_Y(6)$

Comparison with kernel density estimates based on Monte Carlo samples

We now consider classical kernel density estimates based on samples of various sizes. Four density estimates are plotted in figure 5.10 using sample sizes of 10,20,50 and 100. The true density function of Y is shown as the dotted line. In each case, the random inputs are generated using a maximin Latin hypercube scheme. We can see that median estimate of the density function is better than the classical estimate using 50 runs.

5.5 Conclusions

The aim in this chapter has been to provide a means of obtaining a graphical summary of the uncertainty in Y, using a small sample of data. This has been achieved, and in the iodine example, the estimate of the density function captured the shape of the true density function. One complication we have noted is that the true density function may not be defined everywhere, though we will not be able to establish this

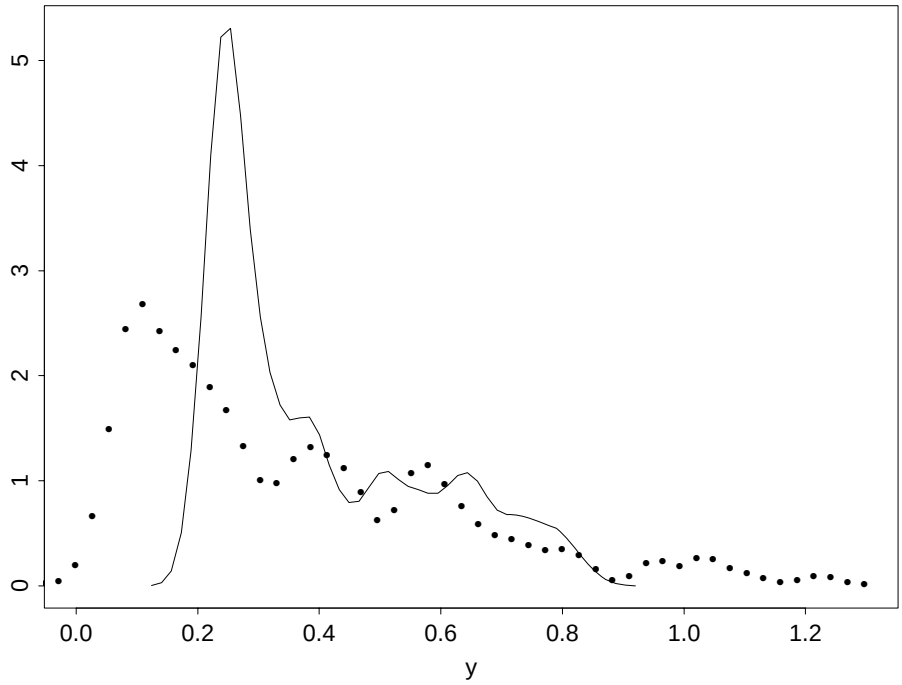

Figure 5.7: Density estimates of $f_Y(6)$ with and without knowledge of $\eta'(0)$

with our method.

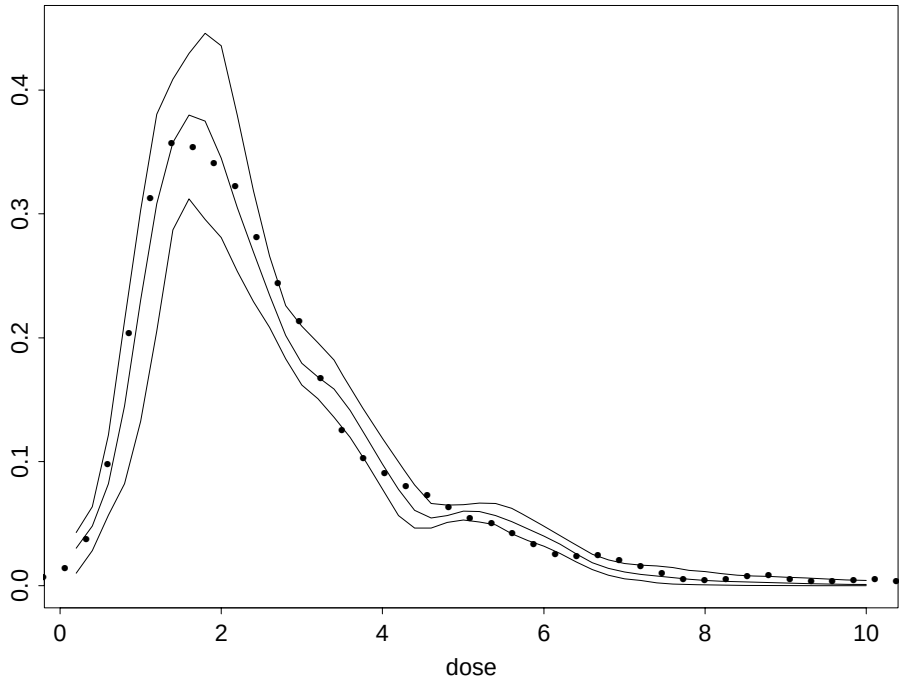

Figure 5.8: Estimate of the density function for the ^{131}I model

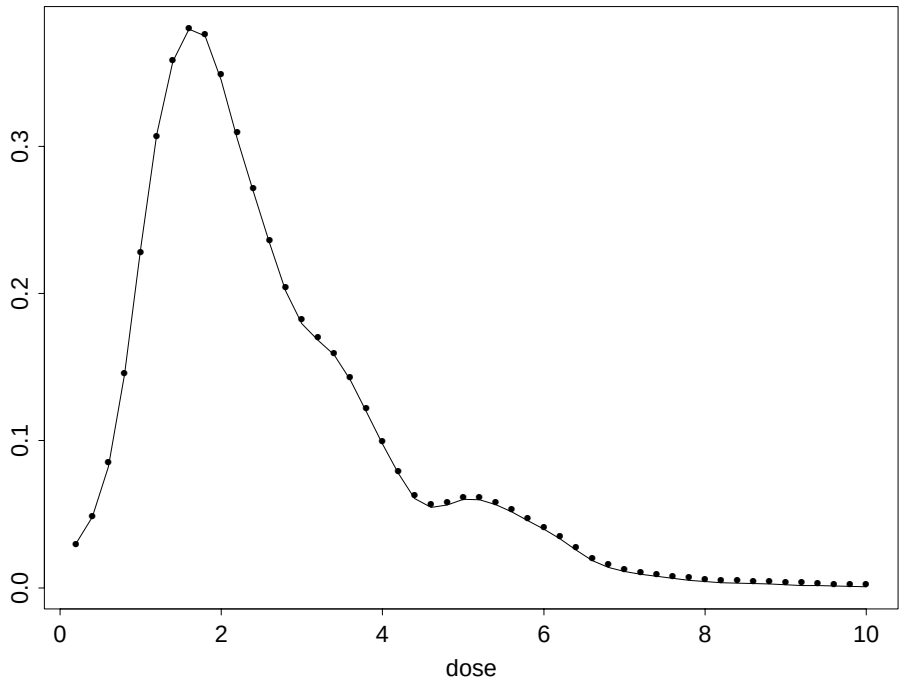

Figure 5.9: The mean and median density functions for the ^{131}I model

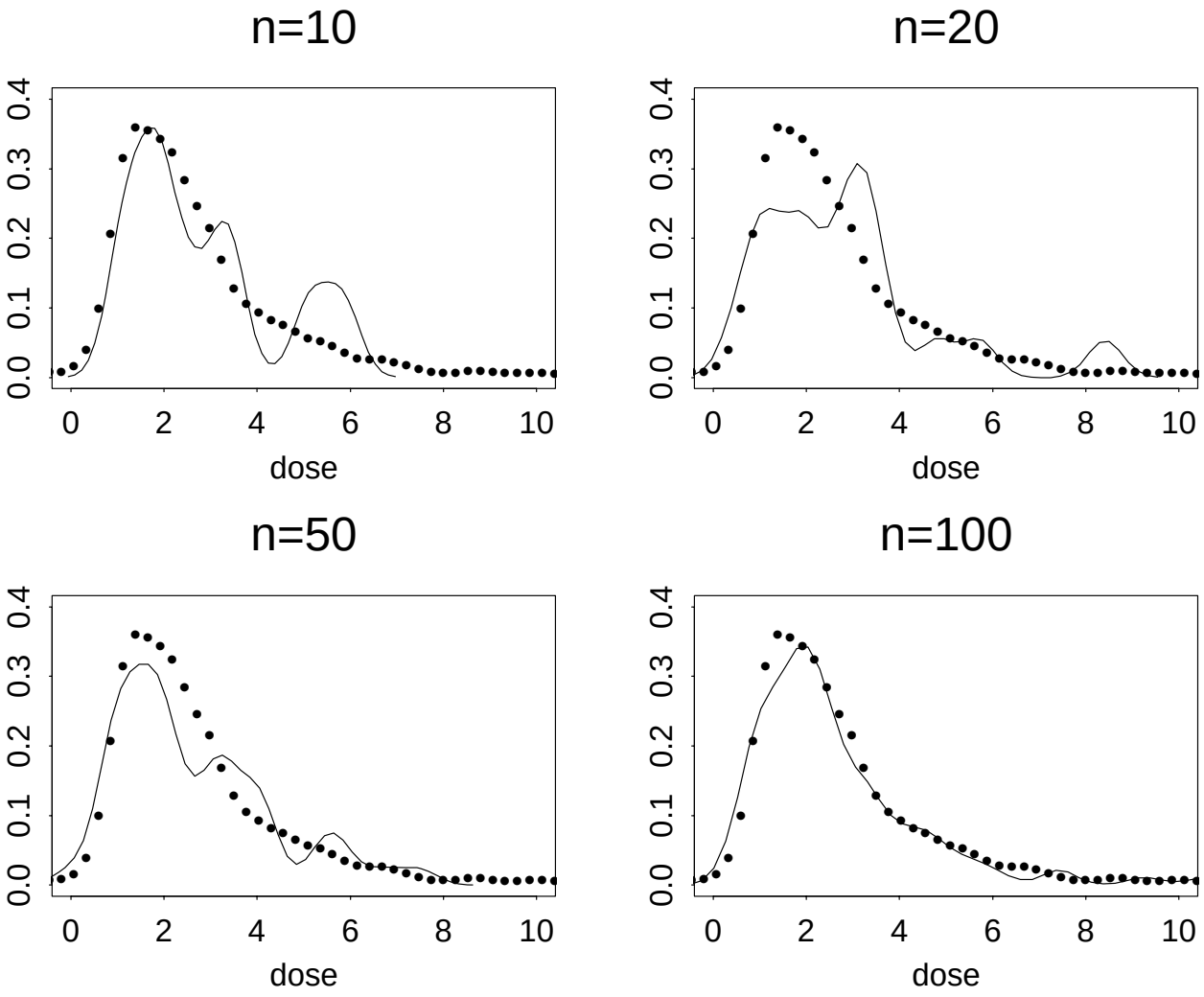

Figure 5.10: Classical kernel density estimates for the ^{131}I model

Chapter 6

Optimal designs for estimating percentiles

6.1 Introduction

This chapter is motivated by a specific problem given to us by the Water Research Centre (WRc). A computer model, known as the SIMPOL model, is to be used to make a prediction, but the true values of some of the inputs are unknown, and so the true output of the model is also unknown. WRc wish to know the 95th percentile of the true output. The interest is in whether it is possible to obtain an accurate estimate of the 95th percentile based on a small number of runs of the computer code. The full model has at least ten uncertain inputs. We develop methodology for estimating the 95th percentile, and test it on a simplified example where only four inputs are considered uncertain, and the rest are kept fixed. In this example we are able to obtain a 'true' estimate of the 95th percentile, and so we can check the accuracy of the estimate obtained by Bayesian methods.

In chapter four we presented a method for estimating any percentile of the output, using simulation. The simulation procedure will form the basis of our approach here. However, in this case, we will not necessarily need to know much information about the function $\eta(.)$ over the entire input space. If we know with certainty that the 95th percentile of the output will occur at some input $\mathbf{x}$, with $\mathbf{x} \in R$ for some

region R, then it will only be necessary to have accurate information about $\eta(.)$ over the region R. Hence estimating any specific percentile efficiently is a question of finding suitable design points to run the code at.

6.2 The SIMPOL model

The SIMPOL model is used as an aid in the design of Combined Sewer Overflows (CSOs) that are required to meet certain environmental standards. A combined sewer carries both sewage and surface water from urban drainage. These sewers are common in the UK. In the event of a storm, the excess of water may lead to an increase of pressure in the sewer pipes and a risk of flooding. To counter this, overflows (CSOs) are built into the sewer system at critical points. These overflows then discharge the excess water into rivers and streams. Clearly, there are environmental concerns when sewage is being discharged. Currently, CSOs are being upgraded to reduce both the number and volume of spills. One method is to provide storage for the storm flow at the CSO. Once the storm has subsided, the stored flow can then continue through the sewer system. The output of the SIMPOL model is the volume of storage required to meet specific environmental standards. These standards usually involve the condition of the river following a spill from the CSO.

6.2.1 The model inputs

Four inputs in the model are treated as unknown and the remainder are kept fixed. These are:

1. Maximum pass forward rate at the CSO, the capacity of the sewer downstream of the CSO. This determines the flow rate at which the storage starts to fill.

2. Average dry weather BOD concentration, the Biochemical Oxygen Demand of the sewer flow in dry weather. The breakdown of organic matter uses oxygen, and the BOD is a measure of the potential potency of the oxygen demand, expressed as a concentration.

3. BOD sediment load, the maximum load that can be built up in a sewer, if sufficient deposition occurs. In dry weather sedimentation can occur in sewers. During a storm, large sewer flows can erode the sediment, causing an increase in the BOD of the storm sewage.

4. BOD erosion concentration, the rate at which the sediment load is eroded.

WRc have assumed log normal distributions for the true values of all four inputs. We transform these inputs so that we can think of $\eta(\mathbf{x})$ as being a function of four inputs x_1, x_2, x_3 and x_4, where the true values all have standard normal distributions. Note that for certain values of the inputs, the required environmental standards are met or even exceeded without the need for any extra storage volume. The SIMPOL model returns a negative output, which is then corrected to zero. For our purposes, we retain the original negative outputs, as these will still give us information about how the output varies with the inputs.

6.3 Methodology for estimating the 95th percentile

As remarked earlier, we do not need to have precise information about $\eta(.)$ over the whole input space $\mathcal{X}$. The aim is to find an output $p_{0.95}$ such that

$$P\{\eta(X) \leq p_{0.95}\} = 0.95. \tag{6.1}$$

Alternatively, we can write this as

$$P(X \in S) = 0.95, \tag{6.2}$$

where $S = \{\mathbf{x} : \eta(\mathbf{x}) \leq p_{0.95}\}$. Now consider an input $\mathbf{x}^* \in \mathcal{X}$, and suppose that we believe with high probability that $\eta(\mathbf{x}^*) \in (a, b)$. If we can find a set $\mathcal{X}_L$ such that $P(X \in \mathcal{X}_L > 0.95)$ and we are certain that $\eta(\mathbf{x}) < a \quad \forall \mathbf{x} \in \mathcal{X}_L$, then we can be confident that the 95th percentile must be less than $\eta(\mathbf{x}^*)$, and it is not necessary to determine the exact value of $\eta(\mathbf{x}^*)$. Similarly, if we can find a set $\mathcal{X}_U$ such that $P(X \in \mathcal{X}_U > 0.05)$ and we are certain that $\eta(\mathbf{x}) > b \quad \forall \mathbf{x} \in \mathcal{X}_U$ then we can be confident that the 95th percentile must be greater than $\eta(\mathbf{x}^*)$, and again, we do not

need to know the exact value of $\eta(\mathbf{x}^*)$. If we can find a set W of all such inputs $\mathbf{x}^*$ that satisfy one of these two conditions, then it follows that we should concentrate the design points in the region $R = \mathcal{X}\backslash W$.

Determining whether or not a particular input $\mathbf{x}^*$ is in R can be done by generating random functions. Suppose we have generated N functions $\eta_{(1)}(.), \ldots, \eta_{(N)}(.)$, and consider the set S_1, the set of N simulated 95th percentiles, and S_2, the set of N simulated values of $\eta(\mathbf{x}^*)$. If every element of S_2 is less than every element of S_1, or every element of S_2 is greater than every element of S_1, then for sufficiently large N, we can be confident that $\mathbf{x}^* \notin R$.

With weak prior information we have no information about R until we evaluate $\eta(.)$ at various inputs. Consequently we propose a sequential procedure. The first stage is to choose a small set of inputs to minimise $\int_{\mathcal{X}} \eta(\mathbf{x}) dG(\mathbf{x})$, as we have done previously. For this purpose one of the methods described in chapter 3 can be used. We can then use the simulation approach to learn about R.

In practice, it is simpler to find inputs that we are certain are in R as opposed to finding inputs that are unlikely to be in R. This is done as follows. We generate a random function $\eta_{(i)}(.)$, and we define the 95th percentile of $\eta_{(i)}(\mathbf{X})$ to be $p_{(i)}^{0.95}$. We then estimate $p_{(i)}^{0.95}$ using the Monte Carlo approach; a set of inputs $\{\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_N^*\}$ is randomly drawn from $G(\mathbf{x})$, and $p_{(i)}^{0.95}$ is estimated by the 95th percentile of $\{m_{(i)}^{**}(\mathbf{x}_{(1)}^*), \ldots, m_{(i)}^{**}(\mathbf{x}_{(N)}^*)\}$. We use the notation $\mathbf{x}_{(i)}^*$ to denote the input in the set $\{\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_N^*\}$ whose expected output $m_{(i)}^{**}(\mathbf{x}_{(i)}^*)$ is the estimate of $p_{(i)}^{0.95}$. Now suppose we have performed the simulation procedure K times, and have obtained $\{p_{(1)}^{0.95} = m_{(1)}^{**}(\mathbf{x}_{(1)}^*), p_{(2)}^{0.95} = m_{(2)}^{**}(\mathbf{x}_{(2)}^*), \ldots, p_{(K)}^{0.95} = m_{(i)}^{**}(\mathbf{x}_{(K)}^*)\}$. We can now think of R as being the smallest region that contains $\{\mathbf{x}_{(1)}^*, \mathbf{x}_{(2)}^*, \ldots, \mathbf{x}_{(K)}^*\}$. The additional design points should then be chosen to minimise

$$\int_R c^{**}(\mathbf{x}) d\mathbf{x}. \tag{6.3}$$

Given the region R, the sample of inputs $\{\mathbf{x}_{(1)}^*, \mathbf{x}_{(2)}^*, \ldots, \mathbf{x}_{(K)}^*\}$ may not necessarily occur uniformly across R. Consequently, we introduce a weight function into (6.3), so that we instead choose design points to minimise

$$\int_R c^{**}(\mathbf{x}) w(\mathbf{x}) d\mathbf{x}, \tag{6.4}$$

for some function $w(\mathbf{x})$.

At this stage it is helpful to clarify the difference between choosing design points to minimise (6.4), and choosing design points to minimise

$$\int_{\mathcal{X}} c^{**}(\mathbf{x}) dG(\mathbf{x}). \tag{6.5}$$

When minimising (6.5) we are interested in learning about the output of the model when run at the true input $\mathbf{X}$. Our uncertainty about $\mathbf{X}$ is described by the distribution $G(\mathbf{x})$. When making inference about the 95th percentile of the output, we are predominantly interested in finding the inputs that give outputs at the high end of the output range, and specificaly the inputs that give outputs at the 95th percentile of the range. We can think of the function $w(\mathbf{x})$ as describing our uncertainty about the inputs of interest.

An alternative is to choose the new n' design points to be the n' inputs that occur the most number of times in the set $\{\mathbf{x}^*_{(1)}, \mathbf{x}^*_{(2)}, \ldots, \mathbf{x}^*_{(K)}\}$, assuming that the set $\{\mathbf{x}^*_1, \mathbf{x}^*_2, \ldots, \mathbf{x}^*_N\}$ is the same for each randomly generated function. This is a simpler approach, though it may result in design points that are too close together. To avoid this, we can make use of the entropy criterion discussed in chapter two. To find n' new design points, we first consider the $2n'$ inputs that occur the most number of times in $\{\mathbf{x}^*_{(1)}, \mathbf{x}^*_{(2)}, \ldots, \mathbf{x}^*_{(K)}\}$, and then find the subset of n' inputs that have the the largest value of V, where V is the posterior variance covariance matrix of outputs at the n' proposed inputs.

6.4 Choosing the simulation design points

The simulation procedure is used twice here; once after the first set of observations to determine where the second set of inputs should be chosen, and again after the second set of observations to obtain a final estimate of the 95th percentile. In both cases, for any generated function $\eta_{(i)}(.)$ we will only be interested in the 95th percentile of $\eta_{(i)}(X)$, and so the variance of $\eta_{(i)}(\mathbf{x})$ will only need to be small for certain inputs $\mathbf{x}$.

After running the model at the first set of inputs, we can find lower and upper

bounds for the 95th percentile without having to simulate any functions. We draw a large sample of N random inputs $\{\mathbf{x}_1^*, \ldots, \mathbf{x}_N^*\}$, and consider pointwise 99% intervals for $\eta(\mathbf{x})$. We then find the 95th percentiles of the sets $\{m^{**}(\mathbf{x}_1^*) - t_{n-q,0.005}\hat{\sigma}c^{**}(\mathbf{x}_1^*),$ $\ldots, m^{**}(\mathbf{x}_N^*) - t_{n-q,0.005}\hat{\sigma}c^{**}(\mathbf{x}_N^*)\}$ and $\{m^{**}(\mathbf{x}_1^*) + t_{n-q,0.005}\hat{\sigma}c^{**}(\mathbf{x}_1^*),$ $\ldots, m^{**}(\mathbf{x}_N^*) + t_{n-q,0.005}\hat{\sigma}c^{**}(\mathbf{x}_N^*)\}$. This gives us an interval which we denote by $(p_{0.95}^L, p_{0.95}^U)$ for the 95th percentile. When considering the simulation design points $\{\mathbf{x}_1', \ldots, \mathbf{x}_{n'}'\}$, for $i = 1, \ldots, n'$ we exclude $\mathbf{x}_i'$ from the design if

$$m^{**}(\mathbf{x}') - t_{n-q,0.005}\hat{\sigma}c^{**}(\mathbf{x}') > p_{0.95}^U, \tag{6.6}$$

or

$$m^{**}(\mathbf{x}') + t_{n-q,0.005}\hat{\sigma}c^{**}(\mathbf{x}') < p_{0.95}^L. \tag{6.7}$$

For simplicity, we first obtain a Latin hypercube sample as described in chapter two, and then remove unnecessary design points using the above procedure, to obtain suitable simulation design points.

6.5 Obtaining a true value of the 95th percentile

We have the output of the SIMPOL model for a particular sewer network evaluated at approximately 2300 different inputs, which are regularly spaced out to cover each unknown input's sample space. We denote this set of inputs by S_1. This in itself is not sufficient to obtain the true value of the 95th percentile, and so an interpolation procedure is used. For a given value of N, we first draw a sample of N random inputs from $G(\mathbf{x})$. This set of inputs is denoted by S_2. Then for each $\mathbf{x} \in S_2$, we find choose 81 inputs in S_1 using a product design based on the three closest inputs in each dimension. The outputs at these inputs are known, and so we fit the Gaussian process model to the function using this set of 81 outputs. Finally, we estimate $\eta(\mathbf{x})$, by its posterior mean. The variance of $\eta(\mathbf{x})$ should be small, since all 81 inputs are close to $\mathbf{x}$. Repeating this procedure gives us a sample of N outputs, and we use the 95th percentile of this sample as the 'true' value. This procedure is also used to obtain the 'true' output of the model at any input. Obtaining a large

sample of outputs in this way, we calculate the true value of the 95th percentile to be 1142.

6.6 Applying the Bayesian approach to the SIMPOL model

We set $\mathbf{h}(\mathbf{x}) = (1\ x_1\ x_2\ x_3\ x_4)^T$ and

$$c(\mathbf{x}, \mathbf{x}') = \exp\{-(\mathbf{x} - \mathbf{x}')^T B(\mathbf{x} - \mathbf{x}')\}, \tag{6.8}$$

for some diagonal matrix B. We have no proper prior information about the function $\eta(.)$, and so we choose two points in each dimension and use a sixteen point product design to make $\int_{\mathcal{X}} c^{**}(\mathbf{x})dG(\mathbf{x})$ small. We denote the intital sixteen observations by $\mathbf{y}_1$. After evaluating $\eta(\mathbf{x})$ at the sixteen points, we estimate B using the cross validation method.

6.6.1 Choosing simulation design points

An example to illustrate rejecting simulation design points

We now give an example to illustrate the idea of rejecting simulation design points as described in section 6.4. The initial bounds $p^L_{0.95} = 984$ and $p^U_{0.95} = 1364$ are obtained for the 95th percentile. We first consider an 81 point product design for the simulation design points, $\{\mathbf{x}'_1, \ldots, \mathbf{x}'_{81}\}$. For each design point $\mathbf{x}'_i$, we evalaute $m^{**}(\mathbf{x}') - t_{11,0.005}\hat{\sigma}c^{**}(\mathbf{x}')$ and $m^{**}(\mathbf{x}') + t_{11,0.005}\hat{\sigma}c^{**}(\mathbf{x}')$ and reject the design point $\mathbf{x}'_i$ if both these values lie outside the interval (984,1364). This gives us 32 accepted design points and 49 rejected design points. We now generate a function $\eta_{(j)}(.)$ using the 32 accepted points, and estimate the 95th percentile of $\eta_{(j)}(X)$, using Monte Carlo. The estimate is 1160, and by considering pointwise bounds for $\eta_{(j)}(\mathbf{x})$ as described in subsection 3.2.3 we obtain lower and upper bounds for the 95th percentile as 1077 and 1233.8. The exact value of $\eta_{(j)}(.)$ is known at the 16 intial design points and the 32 accepted design points. We now simulate the output of $\eta_{(j)}(.)$ at the 49 rejected inputs, so that the output is now known at 97 outputs in

total. The Monte Carlo estimate of the 95th percentile is 1153, and the new lower and upper bounds are 1081 and 1226.1. Thus generating the additional 49 outputs has made little difference to the estimate of the 95th percentile.

Choosing the simulation design points for the SIMPOL example

To obtain suitable simulation design points to learn about R, we begin with a Latin hypercube sample of size 200, and reduce this to 55 design points through rejecting unnecessary design points. We generate 1000 functions, and for each function we draw a random sample of 1000 inputs from $G(\mathbf{x})$. We obtain the sample of inputs $\{\mathbf{x}^*_{(1)}, \ldots, \mathbf{x}^*_{(1000)}\}$, where $m^{**}_{(i)}(\mathbf{x}^*_{(i)})$ is the estimate of $p^{0.95}_{(i)}$. Based on the intial 16 runs of the code, we estimate the 95th percentile of the output to be 1159.0, and a 95% interval for the 95th percentile is (1071,1255).

In equation (6.4), we considered using a weight function $w(\mathbf{x})$ when minimising the variance over the region R. To determine a suitable weight function, we might consider fitting a density function to $\mathbf{x}$ in the set $\{\mathbf{x}^*_{(1)}, \ldots, \mathbf{x}^*_{(1000)}\}$. We first consider x_1, x_2, x_3 and x_4 separately, and plot kernel density estimates in figures 6.1 to 6.4.

Within R, there are also correlations between some of the input parameters, and these are given in table 6.1 A multivariate normal distribution $N(\mathbf{a}, V)$ is fitted to

	x_1	x_2	x_3	x_4
x_1	1	0.010	0.224	−0.079
x_2	0.010	1	−0.578	0.206
x_3	0.224	−0.578	1	0.537
x_4	−0.079	0.206	0.537	1

Table 6.1: Correlations between the inputs in R

$\mathbf{x}$ within the region R. This distribution is supposed to represent the region of the input space where high outputs around the 95th percentile occur. Since we can obtain true outputs of the model easily, we check to see if this is the case. A large sample of inputs is drawn from $N(\mathbf{A}, V)$, and the output of the model at these inputs is evaluated. A kernel density estimate of these outputs is plotted in figure

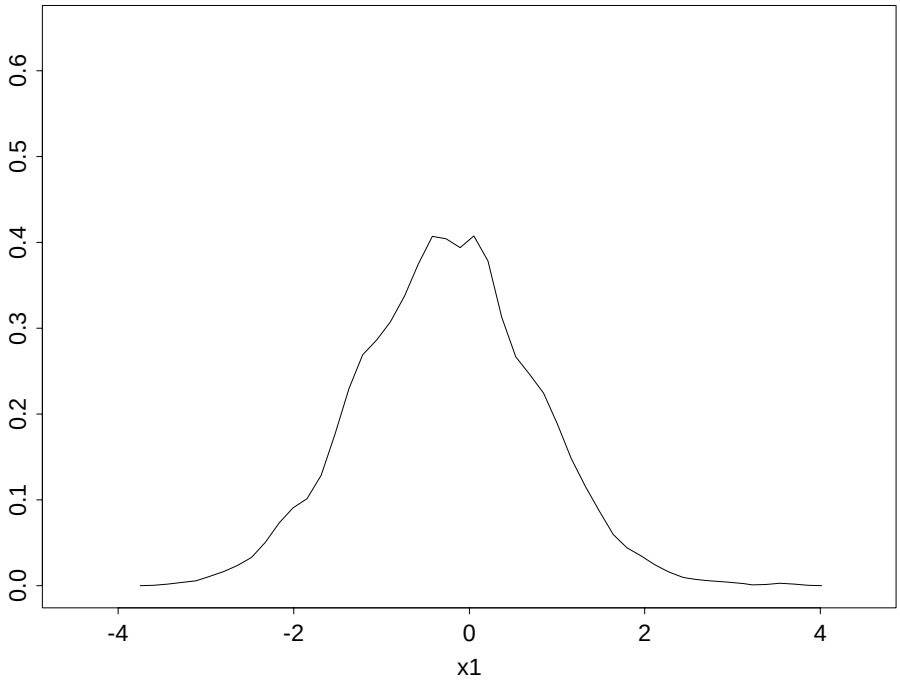

Figure 6.1: Density estimate of x_1 in R

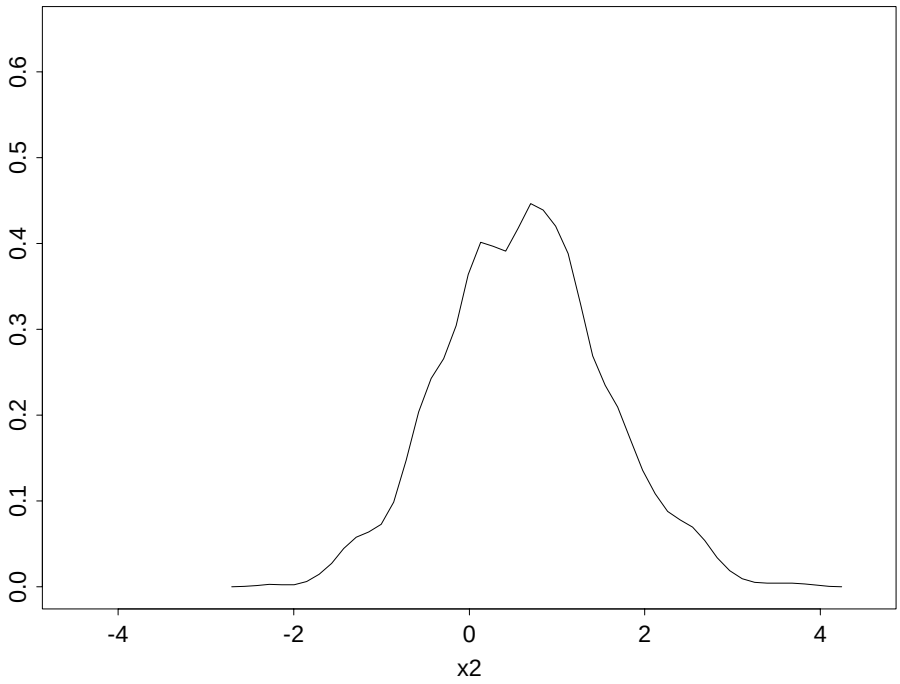

Figure 6.2: Density estimate of x_2 in R

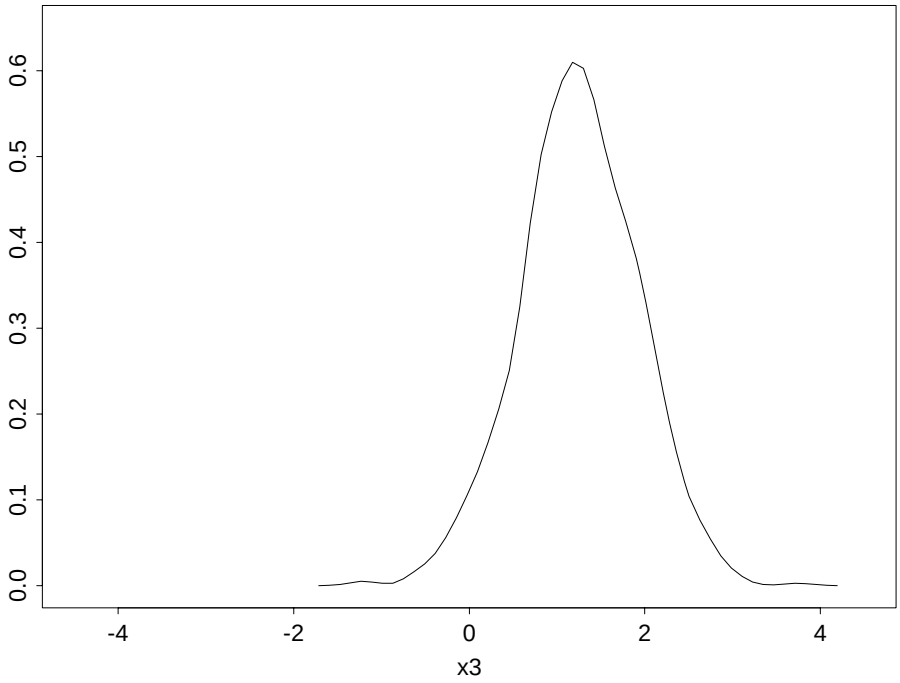

Figure 6.3: Density estimate of x_3 in R

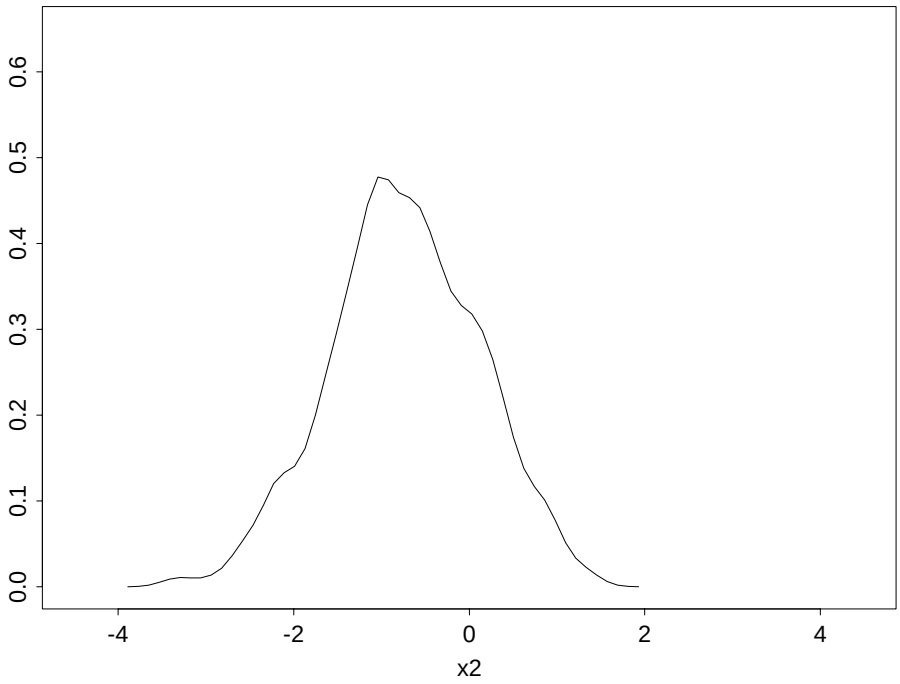

Figure 6.4: Density estimate of x_4 in R

6.5 as the solid line. Note that the modal output of this new density is very close to the true 95th percentile. The dotted line shows the density estimate of Y, based on a large sample of outputs whose inputs are drawn from $G(\mathbf{x})$.

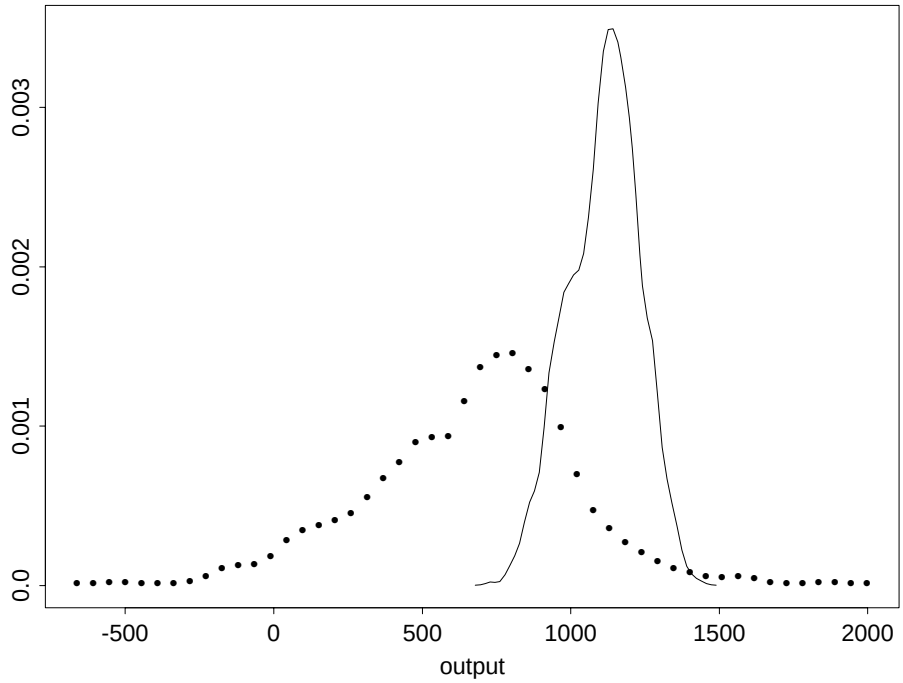

Figure 6.5: The density function of $\eta(\mathbf{X})$, and the fitted density of $\eta(\mathbf{x})$ for $\mathbf{x} \in R$

We now consider choosing eight new design points, to add to the intial sixteen. An eight point design is proposed as follows. We first generate a Latin Hypercube sample of eight inputs using standard normal distributions in each dimension. Denote these eight inputs to be $\mathbf{x}_1, \ldots, \mathbf{x}_8$. We then make the transformation

$$\mathbf{x}_i' = \mathbf{a} + U\mathbf{x}_i, \tag{6.9}$$

where U is the Cholesky square root of V, so that the transformed design points are drawn at random from $N(\mathbf{a}, V)$. We then calculate the determinant of the posterior variance covariance matrix of the output at these eight inputs. Repeating this procedure many times, we choose the new eight inputs to be those with the largest determinant of their variance covariance matrix. The eight new observations are denoted by $\mathbf{y}_2$.

We update the distribution of $\eta(.)$ after learning the eight new outputs, and use the simulation procedure again to obtain a final estimate of the 95th percentile. After simulating 1000 functions, we estimate the 95th percentile to be 1150.5, and a 95% interval for the 95th percentile is (1122.4,1172.1). In figure 6.6, we plot the median distribution function, and a pointwise 95% interval. As in previous examples, by fixing B we may have underestimated the uncertainty in the 95th percentile.

We now use the simpler approach of using the same set of random inputs when using Monte Carlo methods to determine the 95th percentile of each generated function, and finding the eight inputs that occur the most frequently in the set $\{\mathbf{x}^*_{(1)}, \ldots, \mathbf{x}^*_{(1000)}\}$. We denote the corresponding eight new outputs by $\mathbf{y}_3$. After updating the distribution of $\eta(.)$ and simulating 1000 functions, we estimate the 95th percentile to be 1123, and a 95% interval for the 95th percentile is (1094,1162). In figure 6.7, we plot the median distribution function, and a pointwise 95% interval. The interval for low values of the output is small, though this is artificial; we have not simulated $\eta(.)$ at inputs where the output is expected to be low.

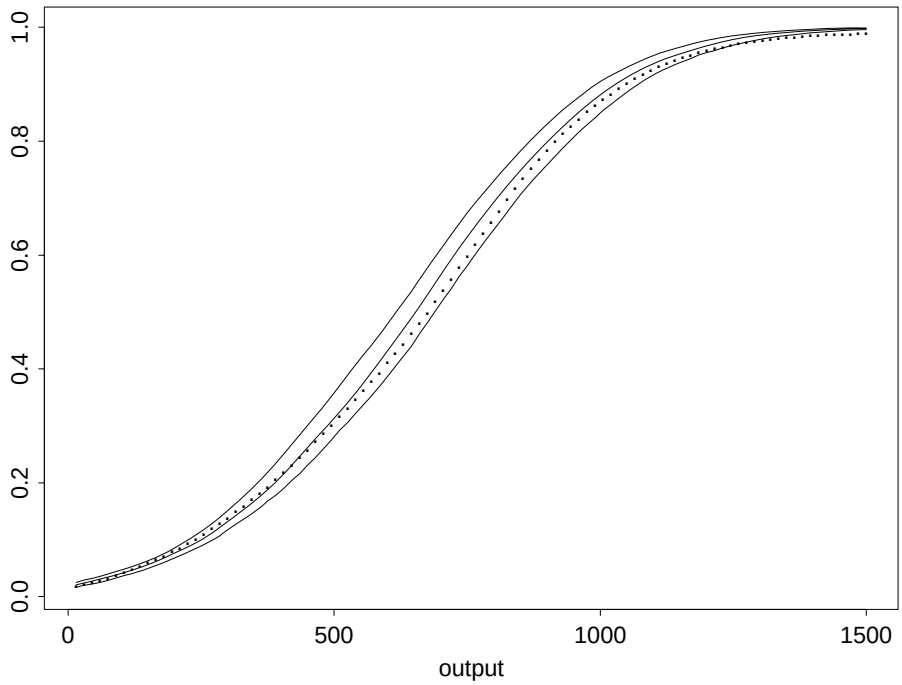

Figure 6.6: Estimate of the distribution function given data $\mathbf{y}_1$ and $\mathbf{y}_2$

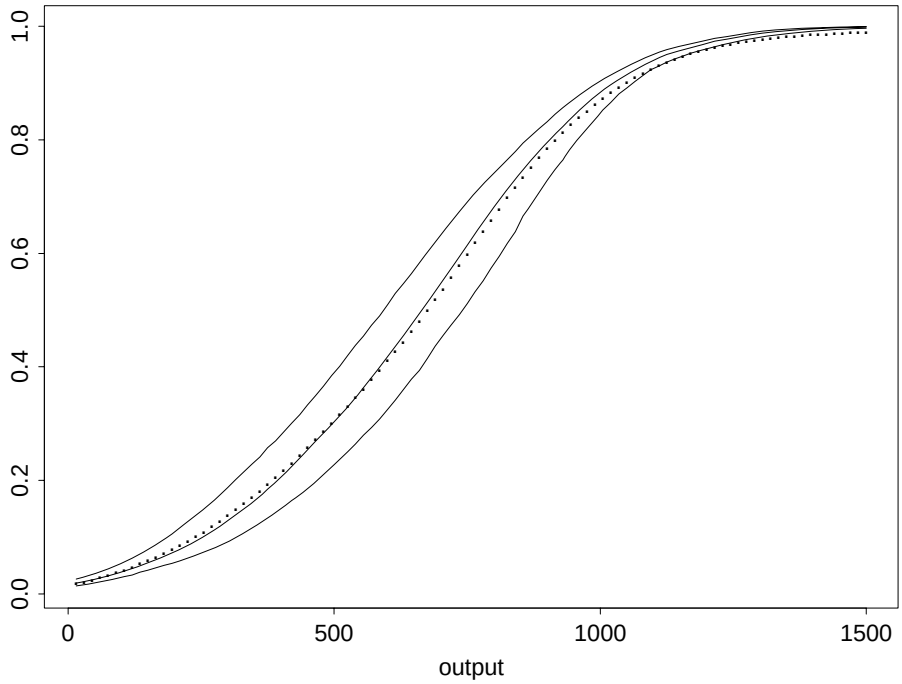

Figure 6.7: Estimate of the distribution function given data $\mathbf{y}_1$ and $\mathbf{y}_3$

We should also compare these two methods with a twenty four point one stage design. Twenty four points are chosen using the maximin Latin hypercube scheme (Mitchell and Morris, 1995), as described in chapter two. The estimate of the 95th percentile using these observations is 1162. A 95% interval for the 95th percentile is $(1125, 1199)$. We summarise these results in table 6.2. Recall that the true value of the 95th percentile is 1142. We can see that the simple two stage approach has not given better results than those achieved by choosing all twenty four design

Method	Median estimate	95% interval
Two stage design, fitting a normal density	1150.5	(1122.4,1172.1)
Simple two stage design	1123	(1094,1162)
One stage design	1162	(1125,1199)

Table 6.2: Estimates of the 95th percentile using different methods to choose the design points

Bayes, $n = 24$	$(1122.4, 1172.1)$
Monte Carlo, $n = 100$	$(1017.0, 1259.4)$
Monte Carlo, $n = 500$	$(1086.2, 1197.3)$
Monte Carlo, $n = 1000$	$(1102.5.0, 1182.5)$

Table 6.3: 95% intervals for the 95th percentile using Bayesian and Monte Carlo methods

points at once. However, fitting the normal density to R to choose the second set of eight points after the initial sixteen has given a smaller 95% interval for the 95th percentile.

We now compare the Bayesian estimate with Monte Carlo estimates based on various sample sizes. Campbell and Gardener (1988) give the following procedure for deriving confidence intervals for percentiles:

Given a sample of n observations, first arrange the observations into ascending order. Then the lower and upper bounds for a $100\alpha\%$ confidence interval for the 95th percentile are given by the r and sth observations respectively, where

$$r = n \times 0.95 - Z_{1-\frac{\alpha}{2}}\sqrt{n \times 0.95 \times 0.05} \tag{6.10}$$

$$s = 1 + n \times 0.95 + Z_{1-\frac{\alpha}{2}}\sqrt{n \times 0.95 \times 0.05}, \tag{6.11}$$

where $Z_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ percentage point from the standard normal distribution. If we determine the true distribution function using a sufficiently large sample of observations, then we can find confidence intervals for the 95th percentile based on samples of size n by finding y_l and y_u such that

$$F(y_l) + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{F(y_l) \times \{1 - F(y_l)\}}{n}} = 0.95, \tag{6.12}$$

and

$$F(y_u) - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{F(y_u) \times \{1 - F(y_u)\}}{n}} = 0.95. \tag{6.13}$$

In table 6.3 we give 95% intervals for the 95th percentile using both the Bayesian and Monte Carlo approaches. The sample size in each case is denoted by n. Estimates of the distribution functions and pointwise 95% intervals are plotted in figure 6.8.

We can see that the Bayes estimate is better than the Monte Carlo estimate using a 1000 runs of the code.

6.7 Conclusions

In this chapter the aim has been to obtain an accurate estimate of the 95th percentile based on a small number of runs of the code. This has been achieved, although the method used may not work as effectively when the input has a large number of dimensions, since the number of simulation design points needed can increase rapidly as the number of dimensions increases. In the SIMPOL example, the output is monotonic with respect to each input, and this has resulted in the region R being noticeably smaller than $\mathcal{X}$. In non-monotone cases, it might not be possible to concentrate the design points in one particular region of the input space, and more runs of the code may be necessary to obtain a good estimate.

An issue that has not been discussed in this particular example is the choice of the log scale in the correlation function. As in the iodine example, we transformed the inputs so that they had normal distributions, which then has implications for the correlation structure (see chapter four). We do not have prior knowledge about the roughness of the function, and whether or not it varies over the input space. In practice, an isotropic correlation function would not have been entirely appropriate here. Some of the four inputs in the model gave negative outputs when set at extreme values. Since any negative output is corrected to zero, we would wish to think of the outputs at any two of these extreme inputs as being highly correlated, irrespective of the distance between the inputs.

Finally, in this example, our choice of $\mathbf{h}(.)$ was not guided by any strong prior knowledge. Recall that in chapter two we noted that adding in regression terms in $\mathbf{h}(.)$ is not always beneficial, and a more complete analysis should investigate the effects of using alternative forms, for example, $\mathbf{h}(\mathbf{x}) = (1)$.

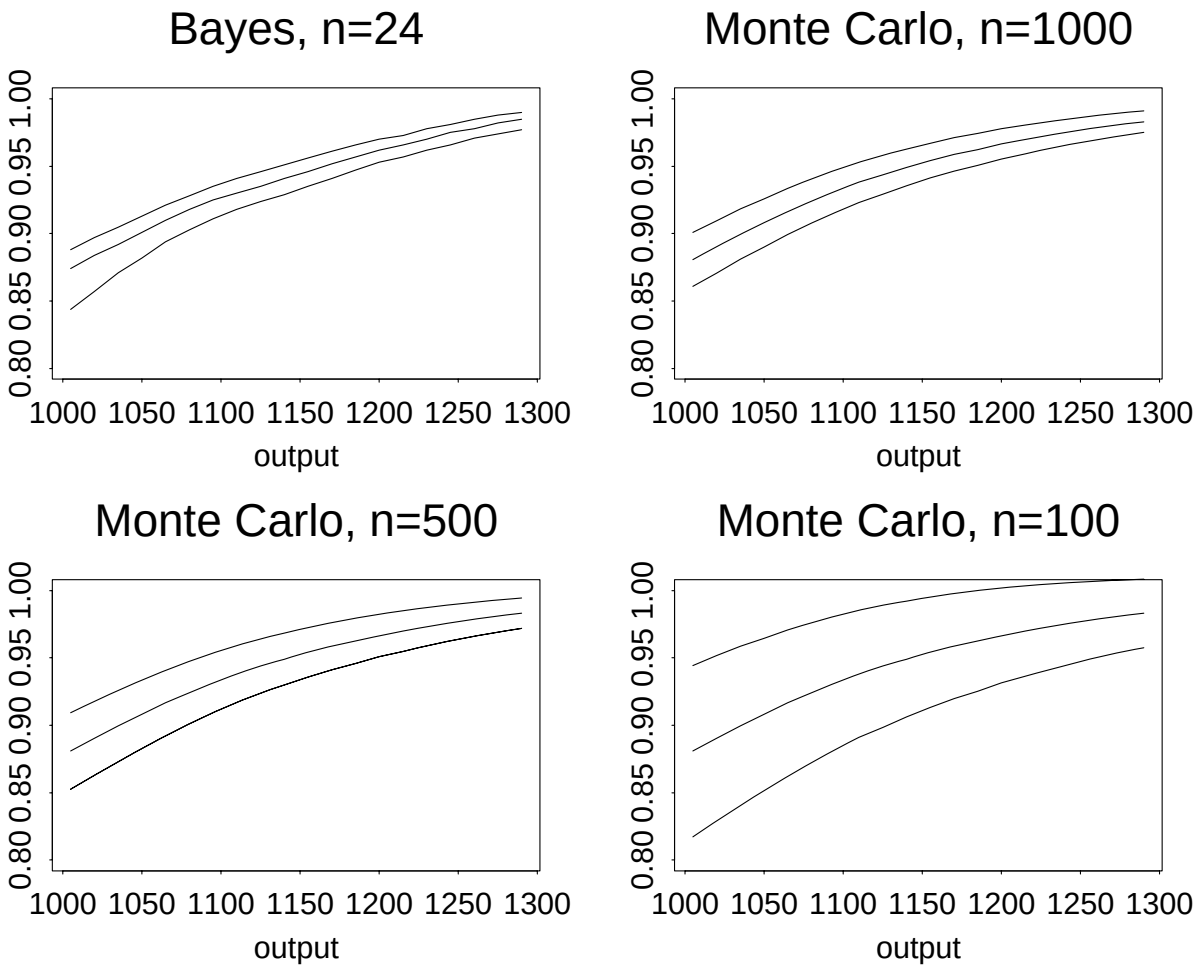

Figure 6.8: Estimates of the distribution function using Bayesian and Monte Carlo methods

Chapter 7

Eliciting Prior Beliefs

7.1 Introduction

In this chapter we consider exploiting a source of information that has not been fully utilised so far in this thesis: prior information about the function $\eta(.)$. Recall that our prior model for $\eta(.)$ is given by

$$\eta(.)|\boldsymbol{\beta}, \sigma^2 \sim N\{\mathbf{h}(.)^T\boldsymbol{\beta}, \sigma^2 C(.,.)\}. \tag{7.1}$$

To complete the prior specification, weak prior distributions for β and σ^2 are commonly used, and a conventional choice is

$$p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2} \tag{7.2}$$

This implies that for any $\mathbf{x}$, the variance of $\eta(\mathbf{x})$ is infinite. Clearly, an individual with knowledge of the computer model and the process being represented should have some idea of what range of outputs to expect, and how the output will vary with the input. Consequently, we expect that in most cases, proper prior information will be available. In addition, since our interest is in computationally expensive computer models, the quantity of data, the number of runs of the code, will be limited. We have also noted that prior information about the smoothing parameters can be useful. When choosing design points according to some criterion, it is usually nececssary to specify a value of the smoothing parameter before any data has been observed. Finding the posterior mode of B is complicated by the fact that the

likelihood for B is often very flat. A proper prior distribution would be of assistance here. We have two objectives in this chapter, firstly to develop a methodology for eliciting prior beliefs about a computer code, and secondly to assess the benefits of using proper prior distributions in uncertainty analysis.

7.2 Prior to posterior analysis

In principle, one should use whatever prior distribution for $\boldsymbol{\beta}$ and σ^2 best represents the expert's beliefs. For simplicity, we confine our attention to the conjugate prior distribution for $\boldsymbol{\beta}$ and σ^2, the normal-inverse-gamma distribution. For a $q \times 1$ vector $\boldsymbol{\beta}$, this distribution has the form

$$f(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-\frac{1}{2}(d+q+2)} \exp[-\{(\boldsymbol{\beta} - \mathbf{z})^T V^{-1} (\boldsymbol{\beta} - \mathbf{z}) + a\}/(2\sigma^2)], \tag{7.3}$$

with hyperparameters $a, d, \mathbf{z}$ and V. If we adopt this prior for $\boldsymbol{\beta}$ and σ^2, then after observing the $n \times 1$ data vector $\mathbf{y}$, O'Hagan (1993) shows that

$$f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\frac{1}{2}(d+q+n+2)} \exp[-\{(\boldsymbol{\beta} - \mathbf{z}^*)^T (V^*)^{-1} (\boldsymbol{\beta} - \mathbf{z}^* + a^*\}/(2\sigma^2)], \tag{7.4}$$

where

$$V^* = (V^{-1} + H^T A^{-1} H)^{-1}, \tag{7.5}$$

$$\mathbf{z}^* = V^*(V^{-1}\mathbf{z} + H^T A^{-1}\mathbf{y}), \tag{7.6}$$

$$a^* = a + \mathbf{z}^T V^{-1}\mathbf{z} + \mathbf{y}^T A^{-1}\mathbf{y} - (\mathbf{z}^*)^T (V^*)^{-1}\mathbf{z}^*. \tag{7.7}$$

In particular we have

$$\boldsymbol{\beta}|\mathbf{y}, \sigma^2 \sim N(\mathbf{z}^*, \sigma^2 V^*) \tag{7.8}$$

and

$$\sigma^2|\mathbf{y} \sim (n + d - q - 2) a^* \chi^2_{n+d-q}. \tag{7.9}$$

If we replace (2.27) and (2.28) by (7.8) and (7.9), then deriving the posterior distribution of $\eta(.)$ in the case of proper prior information is straightforward.

7.3 Example: a one dimensional function with a proper prior distribution

We give a simple example to show the value of including a proper prior distribution. Considering the one dimensional function described in chapter two, we first show in figure 7.1 the mean and 95% pointwise posterior interval for $\eta(x)$ when a weak prior distribution is used, and the function is evaluated at five inputs. We now

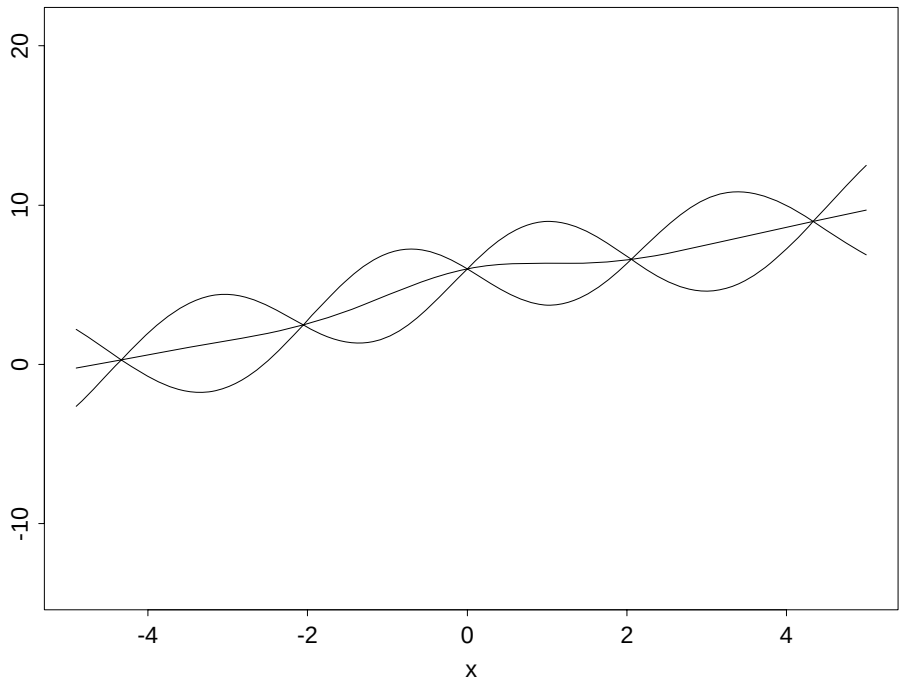

Figure 7.1: The posterior distribution for $\eta(.)$ using a weak prior

consider a normal-inverse-gamma-distribution prior for $\boldsymbol{\beta}$ and σ^2, with $a = 1$, $d = 3$, $\mathbf{z} = (1 \quad 3)^T$ and $V = \begin{pmatrix} 2 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}$. The prior mean and 95% pointwise intervals for $\eta(x)$ are shown in figure 7.2. The posterior mean and 95% pointwise intervals when using this prior are shown in figure 7.3. The posterior interval resulting from a weak prior is shown as the dotted lines. The smoothing parameter is fixed at 0.5 in both cases. We can see that the use of a proper prior distribution has resulted in a noticeable reduction in the posterior variance.

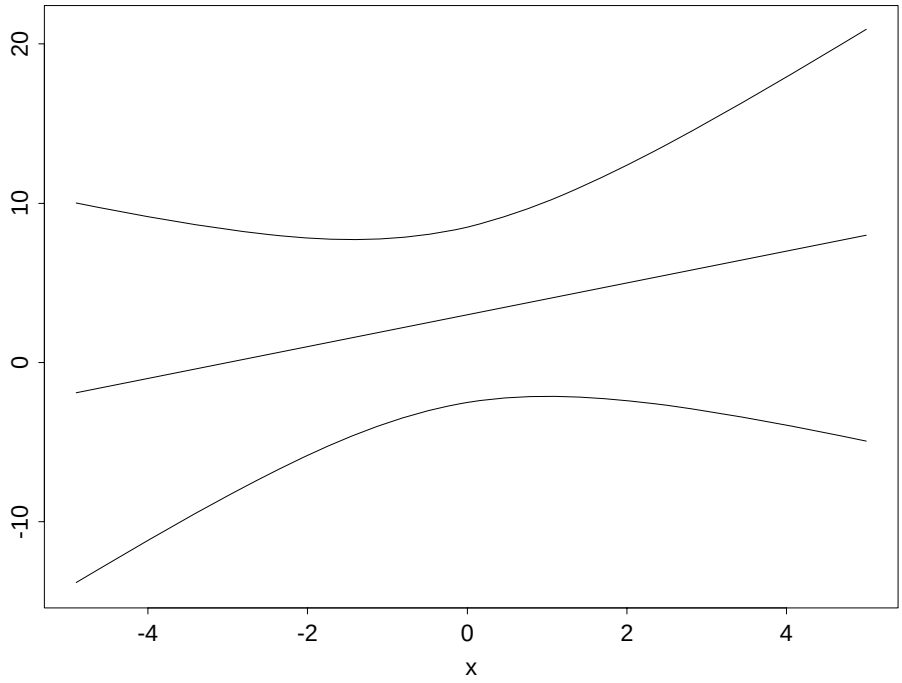

Figure 7.2: A prior distribution for $\eta(.)$

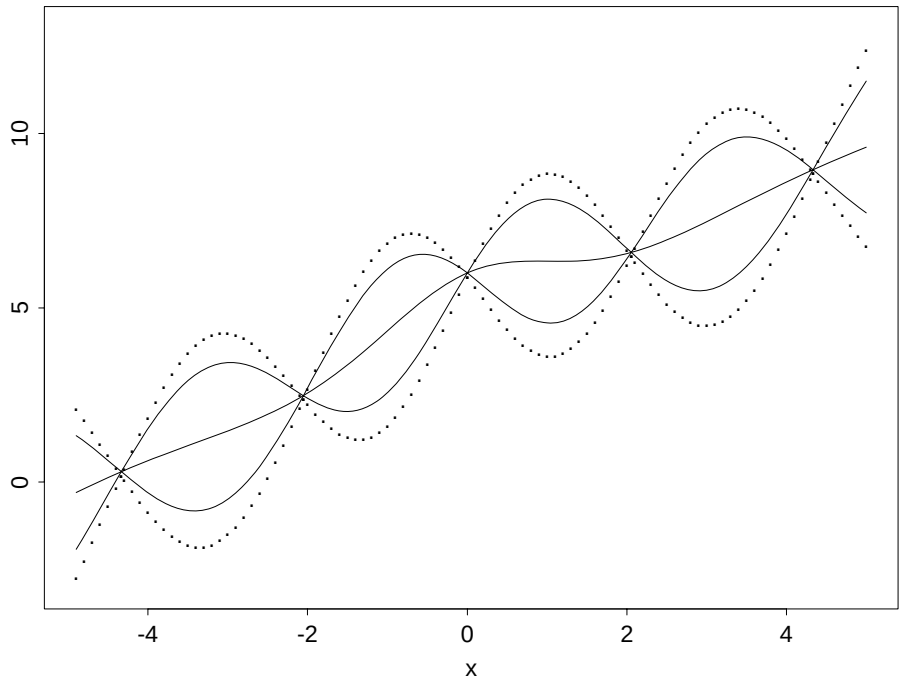

Figure 7.3: The posterior distribution for $\eta(.)$ using a proper prior

We now examine the benefits of using this prior distribution for uncertainty analysis. Suppose we are interested in M, the expected value of the true output, defined by

$$M = \int_{\mathcal{X}} \eta(x) dG(x). \tag{7.10}$$

In the one dimensional example, the true value of M is 5.154. Using a weak prior distribution and five evaluations of $\eta(.)$, we obtain $E(M) = 5.075$ and $Var(M) = 0.193$. If the proper prior distribution is used then we obtain $E(M) = 5.133$ and $Var(M) = 0.026$, and again, a clear improvement can be seen.

7.4 Methodology for eliciting prior beliefs

Prior information can be included through the choice of $\mathbf{h}$ and distributions for $\boldsymbol{\beta}, \sigma^2$ and B. It is generally considered preferable to elicit beliefs about observable quantities, rather than parameters in some statistical model (see for example Kadane and Wolfson, 1998 and O'Hagan, 1998). This is because the expert may have some difficulty in interpreting the meaning of some model parameters. We expect this to be the case here, as it is unlikely that the expert will view the output of a deterministic model as a random variable.

Making probability statements about $\boldsymbol{\beta}$ and σ^2 (given the function $\mathbf{h}(.)$) will automatically imply probability statements about $\eta(.)$ through the Gaussian process model. The underlying technique in the elicitation process used here is to ask the expert for probability statements about the observable quantity $\eta(\mathbf{x})$ at various values of $\mathbf{x}$, and then to determine what the equivalent statements about $\boldsymbol{\beta}$ and σ^2 are. This will only be valid if the expert believes the Gaussian process model to be an appropriate description of the uncertainty about $\eta(.)$. Thus we should first confirm that the Gaussian process model is suitable, before proceeding with any analysis (including or excluding prior elicitation). According to the definition of a Gaussian process, we could ask the expert if they believe that for any set of inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the corresponding outputs $y_1 = \eta(\mathbf{x}_1), \ldots, y_n = \eta(\mathbf{x}_n)$ have a multivariate normal distribution. However, the expert is quite likely to be unable to answer this question. Instead, we could ask the following questions regarding more general

properties of the Gaussian process model:

1. Is the function $\eta(.)$ continuous everywhere?

2. For any $\mathbf{x}$ and $\alpha \in [0,1]$, is your $100\alpha\%$ interval for $\eta(\mathbf{x})$ symmetrical about the median?

3. Given $\eta(\mathbf{x})$, does this give us information about $\eta(\mathbf{x} + \delta\mathbf{x}_0)$ for a small value of δ?

If the answer to 3 is negative, we could still use a Gaussian process model with the diagonal elements of B large. However, inference about the function is harder with a small sample of observations. If the correlation between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ is very small even when $\mathbf{x}$ is close to $\mathbf{x}'$, then the data will not give us much information about the value of $\eta(\mathbf{x})$ at untested values of $\mathbf{x}$.

The choice of correlation function also needs to be considered. Again, we would not expect the expert to propose a function directly. Our preferred choice of function is $c(\mathbf{x}, \mathbf{x}') = \exp\{-(\mathbf{x} - \mathbf{x}')^T B(\mathbf{x} - \mathbf{x}')\}$ since this is mathematically convenient for calculating expressions such as $E\{\eta(\mathbf{X})\}$ and $Var\{\eta(\mathbf{X})\}$ (see Haylock, 1997). In certain cases, this function may not be appropriate, for example, if the expert does not believe that $\eta(.)$ is differentiable everywhere. In this case, it may be necessary to consider one of the other correlation functions descibed in chapter two.

7.4.1 Choice of $\mathbf{h}(.)$

The first step is to choose an appropriate form for $\mathbf{h}(.)$, as this will determine the dimensionality of $\boldsymbol{\beta}$. The expert is asked to propose a crude approximation for $\eta(.)$, and $\mathbf{h}(.)$ is chosen on the basis of this approximation. The expert only needs to consider an approximation of $\eta(\mathbf{x})$ for the values of $\mathbf{x}$ of interest, as described by $G(\mathbf{x})$. We have noted earlier that the choice of $\mathbf{h}(.)$ is not entirely straightforward, and that the expert may be believe that several different approximations are plausible, corresponding to several different forms for $\mathbf{h}(.)$. We do not consider these issues further in this chapter.

7.4.2 Prior beliefs about $\boldsymbol{\beta}, \sigma^2$ and B

The expert's uncertainty about $\eta(\mathbf{x})$ is related to $\boldsymbol{\beta}$ and σ^2 through

$$E\{\eta(\mathbf{x})\} = \mathbf{h}(\mathbf{x})^T E(\boldsymbol{\beta}), \tag{7.11}$$

and

$$Var\{\eta(\mathbf{x})\} = E(\sigma^2) + \mathbf{h}(\mathbf{x})^T Var(\boldsymbol{\beta})\mathbf{h}(\mathbf{x}). \tag{7.12}$$

Assuming a normal inverse gamma distribution, we need to determine $E(\boldsymbol{\beta})$, $Var(\boldsymbol{\beta})$, $E(\sigma^2)$ and d, a degrees of freedom parameter. We use an approach similar to that of Kadane and Wolfson (1998). They were concerned with eliciting prior distributions for parameters in a normal linear model. Differences in their case are that different outputs can be observed at the same input, and that given $\boldsymbol{\beta}, \sigma^2, \mathbf{x}$ and $\mathbf{x}'$, two outputs $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ are independent.

We choose a set of inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and ask the expert to specify their median, 75th percentile and 95th percentile for each output. We denote these by $\eta_{0.5}(\mathbf{x}_i), \eta_{0.75}(\mathbf{x}_i)$ and $\eta_{0.95}(\mathbf{x}_i)$ respectively. The choice of inputs will depend on $G(\mathbf{x})$. We propose choosing well spaced out design points within some central 95% interval for $\mathbf{x}$ defined by $G(\mathbf{x})$.

Determining d

Kadane and Wolfson (1998) obtain a prior value for d by finding d which gives the best fit to

$$\frac{t_{d,0.95}}{t_{d,0.75}} = \frac{1}{n}\sum_{i=1}^{n} a(x_i) \tag{7.13}$$

where $t_{d,\alpha}$ is the 100α percentile of the t distribution with d degrees of freedom, and

$$a(x_i) = \max\left\{\frac{\eta_{0.95}(x_i) - \eta_{0.5}(x_i)}{\eta_{0.75}(x_i) - \eta_{0.5}(x_i)}, \frac{t_{\infty,0.95}}{t_{\infty,0.75}}\right\}. \tag{7.14}$$

The average of the $a(x_i)$s is taken, as there is likely to be some variation in the values of the proposed tail ratios. If one particular ratio is significantly different from all the others, the expert may wish to revise their assessment. In addition, if any single ratio is significantly smaller than $\frac{t_{\infty,0.95}}{t_{\infty,0.75}}$, the expert may also wish to reconsider their assessment.

An alternative method for choosing d is as follows. For a given d, we can estimate the expert's variance of $\eta(\mathbf{x}_i)$, which we denote by $v_{\mathbf{x}_i}$, using

$$\sqrt{v_{\mathbf{x}_i}} = \frac{\eta_{0.95}(\mathbf{x}_i) - \eta_{0.5}(\mathbf{x}_i)}{t_{d,0.95}}, \tag{7.15}$$

and then consider the fitted value of the 75th percentile, denoted by $\hat{\eta}_{0.75}(\mathbf{x}_i)$ and given by

$$\hat{\eta}_{0.75}(\mathbf{x}_i) = \eta_{0.5}(\mathbf{x}_i) + t_{d,0.75}\sqrt{v_{\mathbf{x}_i}}. \tag{7.16}$$

We can then choose d to minimise

$$\sum_{i=1}^{n} \{\hat{\eta}_{0.75}(\mathbf{x}_i) - \eta_{0.75}(\mathbf{x}_i)\}^2 \tag{7.17}$$

Before attempting to minimise (7.17), we should still consider the tail ratios given by $a(x_i)$ to see if any particular estimate needs re-assessing.

Determining $E(\sigma^2), Var(\boldsymbol{\beta})$ and B

Since we are fitting a normal-inverse-gamma distribution to $\boldsymbol{\beta}$ and σ^2, we have $Var(\boldsymbol{\beta}) = E(\sigma^2)V$ for some matrix V. In addition, from the Gaussian process model we have

$$Var\{\eta(\mathbf{x}_i)\} = E(\sigma^2) \times \{1 + \mathbf{h}(\mathbf{x}_i)^T V \mathbf{h}(\mathbf{x}_i)\}. \tag{7.18}$$

If $\mathbf{h}(.)$ is a $q \times 1$ vector, then for $n \geq \frac{1}{2}q(q+1)$, we can consider finding $E(\sigma^2)V$ to minimise

$$\sum_{i=1}^{n} \left[\eta_{0.5}(\mathbf{x}_i) + t_{d,0.95}\sqrt{Var\{\eta(\mathbf{x}_i)\}} - \eta_{0.95}(\mathbf{x}_i)\right]^2. \tag{7.19}$$

At this stage it will not be possible to determine $E(\sigma^2)$ and V separately, as long as the vector of functions $\mathbf{h}(\mathbf{x})$ contains a constant term. For example, suppose the input x is one dimensional, and $h(x)^T = (1\ x)$. Then if the i,j-th element of V is $v_{i,j}$, we have

$$Var\{\eta(x)\} = E(\sigma^2)\{1 + v_{1,1} + (v_{1,2} + v_{2,1})x + v_{2,2}x^2\}, \tag{7.20}$$

and due to the $E(\sigma^2)\{1 + v_{1,1}\}$ term, we can not solve to find $E(\sigma^2)$ for any value of n.

Once an estimate for $E(\sigma^2)V$ has been obtained, we can then obtain fitted values for the variances, which we denote by $\hat{v}_{\mathbf{x}_1}, \ldots, \hat{v}_{\mathbf{x}_n}$. We can then examine $v_{\mathbf{x}_i} - \hat{v}_{\mathbf{x}_i}$ for $i = 1, \ldots, n$, and the expert may wish to revise their estimate $v_{\mathbf{x}_i}$ if the residual $v_{\mathbf{x}_i} - \hat{v}_{\mathbf{x}_i}$ is particularly large.

We now need to elicit a prior value of σ^2, and a prior estimate of B. In principle this could be done as follows. A hypothetical observation $\eta(\mathbf{x}_0) = y_0$ is proposed. The expert is now asked to provide revised 95% intervals for $\eta(\mathbf{x}_0 + \delta\mathbf{x}^{(i)})$ and $\eta(\mathbf{x}_0 + (\delta + \epsilon)\mathbf{x}^{(i)}$, for positive δ and ϵ. We use the notation $\mathbf{x}^{(i)}$ to denote an input whose $i-$th element is 1 and all other elements are 0. This allows us to consider the smoothness of the function $\eta(.)$ in each dimension of the input in turn. We denote the expert's variances of $\eta(\mathbf{x}_0 + \delta\mathbf{x}^{(i)})$ and $\eta(\mathbf{x}_0 + (\delta + \epsilon)\mathbf{x}^{(i)})$ by v_δ and $v_{\delta+\varepsilon}$ respectively. We now find $s^2 = E(\sigma^2)$ and b_i that minimise

$$\left[s^2\{c^{**}_{b_i}(\mathbf{x}_0 + \delta\mathbf{x}^{(i)})\} - v_\delta\right]^2 + \left[s^2\{c^{**}_{b_i}(\mathbf{x}_0 + (\delta + \epsilon)\mathbf{x}^{(i)}\} - v_{\delta+\varepsilon}\right]^2, \tag{7.21}$$

where $c^{**}_{b_i}(\mathbf{x})$ is the posterior covariance function with the $i, i-$th element of B equal to b_i. This process is then repeated for each dimension of the input.

In practice, this procedure is unlikely to work very well, because even though there will only be one value of B and one value of $E(\sigma^2)$ that give the exact variance specified by the expert, for a range of values of B we will be able to find an appropriate value of $E(\sigma^2)$ that gives very similar results. We illustrate this with the prior described in section 7.3. In this prior we have $b = 0.5$ and $E(\sigma^2) = 1$. Now we propose the hypothetical observation $\eta(0) = 2$. Then conditional on this observation, we can calculate the mean and 95th percentile of $\eta(x)$ for any x. Now suppose we were to use a different value of b in the prior model. We now find a value of $E(\sigma^2)$ such that the posterior distribution of $\eta(x)|\eta(0) = 2$ is similar to the posterior distribution of $\eta(x)|\eta(0) = 2$ when $b = 0.5$ and $E(\sigma^2) = 1$. If the new value of b is larger than 0.5, then we achieve this by making $E(\sigma^2)$ smaller than 1, and vice versa if the new value of b is smaller than 0.5. In figure 7.4, we have four plots where in each case, the dotted lines show the mean and 95th percentile of $\eta(x)$ conditional on $\eta(0) = 2$ when $b = 0.5$ and $E(\sigma^2) = 1$. The solid lines show the mean and 95th percentile for different values of b and an appropriate value of $E(\sigma^2)$. If

the expert's original prior judgements correspond to $b = 0.5$ and $E(\sigma^2) = 1$, then even in the case $b = 5$, the expert may believe that the alternative distribution is acceptable.

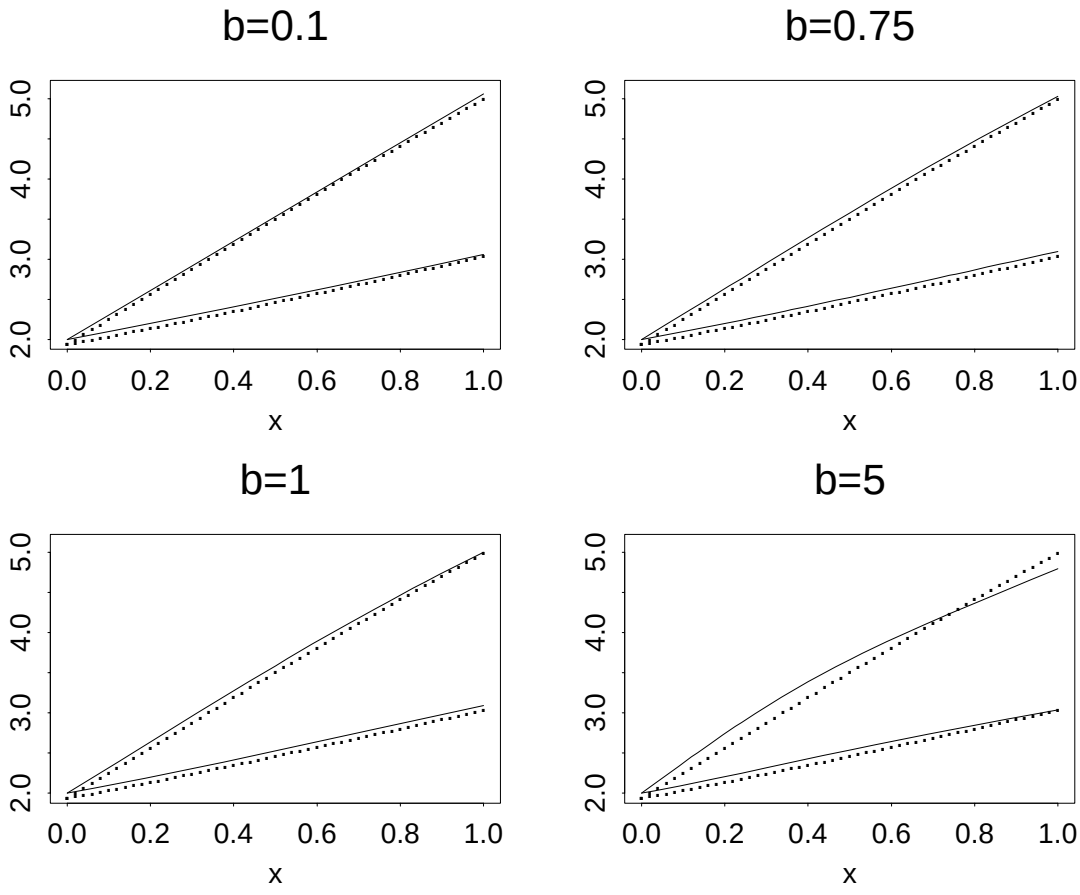

Figure 7.4: Posterior distributions with different values of b and $E(\sigma^2)$

Eliciting B

Clearly, the process of eliciting $E(\sigma^2)$ will be far simpler if we can keep B fixed. We now consider eliciting B separately using generated functions for different values of B. Each dimension is considered individually. We explain the technique using the one dimensional example in section 7.3, and then describe the adaptations needed for higher dimensions.

The idea is to generate functions from the prior distribution with a particular value of b, and then ask the expert to view these functions and make a judgement

about the plausibilty of the realisations in terms of their roughness. Suppose that we have elicited $E(\sigma^2)V$ and posterior intervals for various outputs conditional on a hypothetical observation. For a particular value of b, we find $E(\sigma^2)$ to fit the expert's posterior intervals. We then simulate functions from the prior distribution. However, for any one fixed value of b, the roughness of the generated functions will vary in terms of their appearance. We therefore need to show the expert a range of simulated functions that will illustrate the various degrees of roughness that can result from any single value of b. We propose the following measure to quantify the roughness of each function. If the functions are generated between the inputs x_l and x_u, then for each function $\eta_{(i)}(.)$ we quantify its roughness r_i to be

$$r_i = \int_{x_l}^{x_u} |\eta''_{(i)}|(x)dx, \tag{7.22}$$

so that r_i takes its minimum value when the function $\eta_{(i)}(x)$ is a straight line. We approximate $\eta''_{(i)}(x)$ by $\frac{d^2}{dx^2}m^{**}_{(i)}(x)$, which will be a good approximation if a sufficient number of outputs have been generated. We now generate a large sample of N random functions for a given b, evaluate r_i for $i = 1, \ldots, N$, arrange the functions into ascending order according to their value of r_i, and select functions $j \times \frac{N}{10}$ for $j = 1, \ldots, 10$ to show to the expert. This will give us a suitable set of functions to show the expert. We give an illustration of this in figure 7.5. Following this, the simplest option is for the expert to choose which value of b seems most appropriate, based on these plots. We may then choose to keep b fixed at this value, since we have noted that in some cases the data give little information about b. A more complex approach is to ask the expert to assess their probability that the true function $\eta(x)$ is at least as rough as the functions shown, for each value of b. We can then fit a density function to these judgements.

In higher dimensions, we consider each dimension in turn. If $\mathbf{x}^T = (x_1, \ldots, x_n)$, then for $i = 1, \ldots, n$, we allow x_i to vary and fix the other $n - 1$ inputs at their mean values. $\eta(\mathbf{x})$ is then treated as a one-dimensional function and we proceed as before.

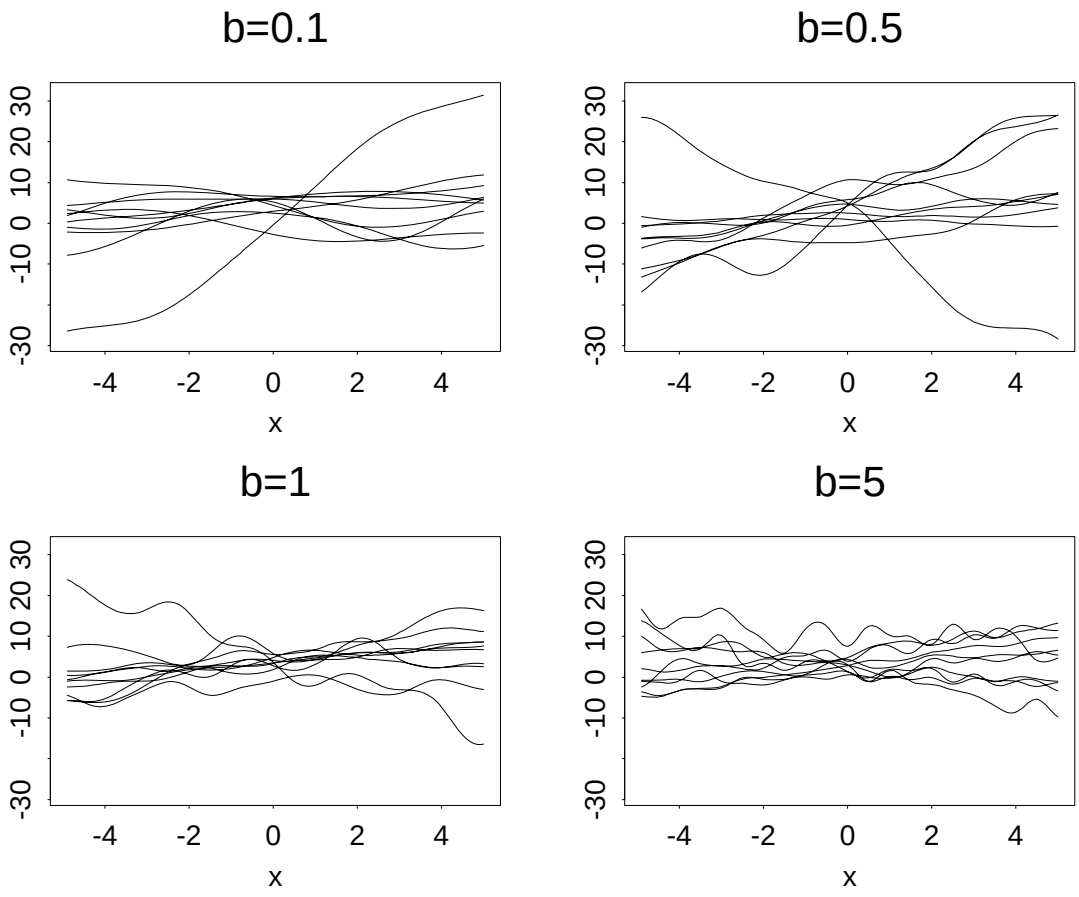

Figure 7.5: Generated functions for eliciting beliefs about smoothing parameters

Determining E($\boldsymbol{\beta}$)

Once we have obtained a prior value for B, we can elicit a prior mean for $\boldsymbol{\beta}$. We estimate $E(\boldsymbol{\beta})$ by

$$E(\boldsymbol{\beta}) = (H^T A^{-1} H)^{-1} H^T A^{-1} \mathbf{f}_{0.5}. \tag{7.23}$$

We can also compare $\mathbf{f}_{0.5}$ with $HE(\boldsymbol{\beta})$, to see if any single median is far from the fitted mean.

7.5 Example: the FARMLAND model

We now give an example of an elicitation involving a submodel of the FARMLAND (Food Activity from Radionuclide Movement on LAND) model. The FARMLAND model is designed to simulate the passage of radionuclides through the foodchain

following deposition of substances on the ground. It consists of a set of submodels which each represent a different part of the foodchain. A full description of the model is given in Brown and Simmonds (1995). Here we consider one such submodel which describes the movement of radionuclides in grain. A diagram representing the grain model is shown in figure 7.6. The plant is represented by a series of compartments, shown by the boxes. Radionuclides are deposited onto compartments 1,2 and 7. The arrows between the compartments show the possible transfers within the plant. The transfer rates are given by the constants k_{ij}, and we consider these to be the input parameters of the model. The output of the model is the concentration at harvest of radionuclides in the internal grain 1 and 2 compartments, and these are highlighted in bold in the diagram. Two scenarios for the deposition can be considered; either continuous deposition over a period of time, or a single deposit at a certain time before harvest.

In this example we will fix all but two of the transfer rates at their default values. The two parameters that we will allow to vary are k_{23}, the transfer rate from the external plant compartment to the internal plant compartment, and k_{15}, the transfer rate from the soil to the internal grain output compartment. We consider a single deposit of a unit quantity of strontium occuring thirty days before harvest.

We elicit the beliefs about the output of the grain model from one of the authors of Brown and Simmonds (1995). The expert states that their $100\alpha\%$ intervals for $\eta(\mathbf{x})$ will be symmetrical about the median for the range of inputs that we will consider. In addition, the function is known to be continuous everywhere, and smooth (so that small values of the smoothing parameters are likely to be appropriate). The expert also believes that over the range of inputs of interest, the output will be approximately linear in the two inputs. Examining figure 7.6, it is immediately obvious that the output must be zero if both k_{23} and k_{15} are zero. Consequently, we assume that $\eta(\mathbf{0}) = 0$ is known when fitting the normal inverse gamma distribution, and set

$$h(\mathbf{x})^T = (k_{23} \quad k_{15}). \tag{7.24}$$

In table 7.1, we show the initial 50th, 75th and 95th percentiles of the output

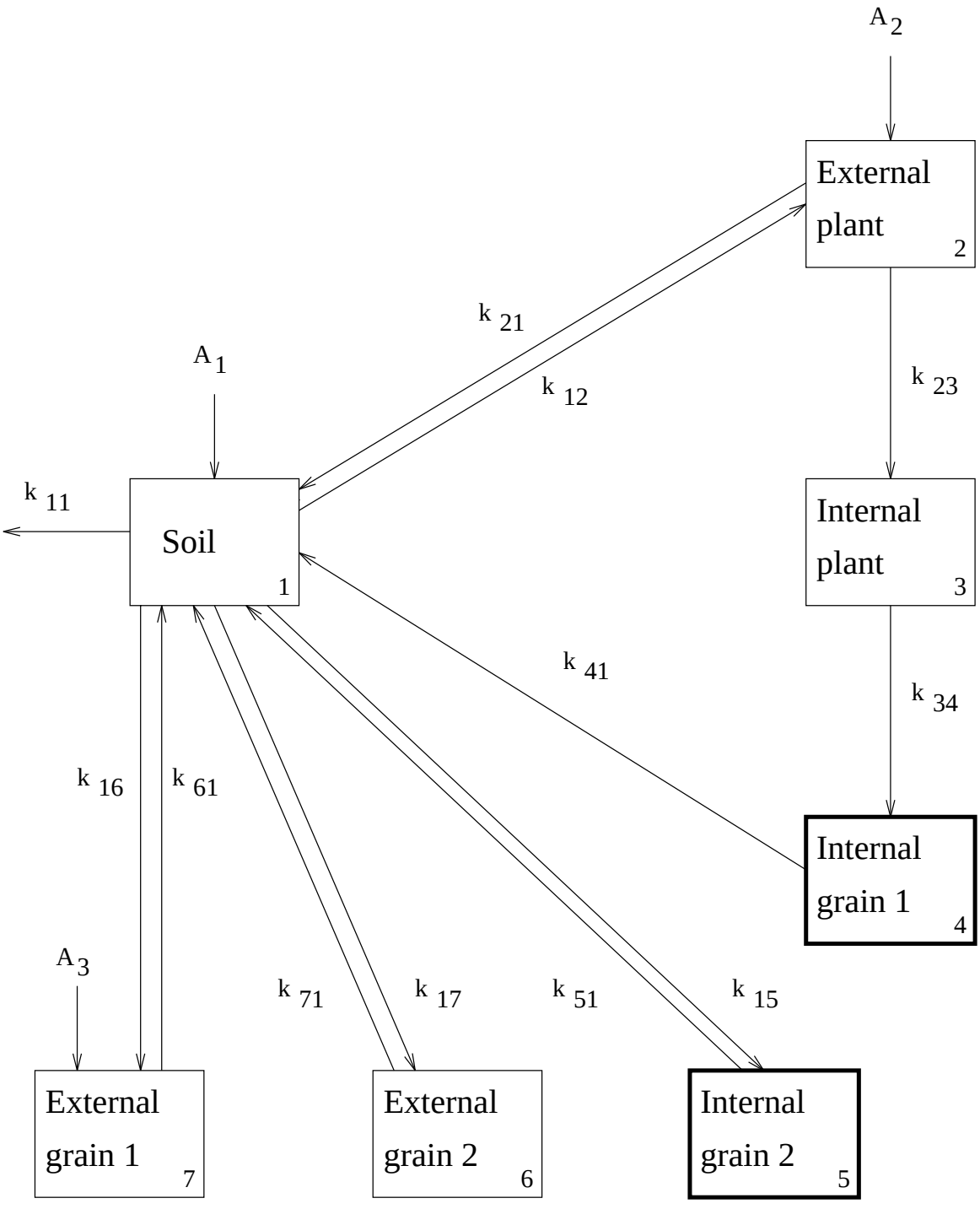

Figure 7.6: A graphical representation of the grain model

provided by the expert for various inputs. We now find a suitable value to use for the degrees of freedom parameter, d. In table 7.2 we give the expert's tail ratio's for each output.

The first and sixth ratios are noticeably larger than the others, and the expert revises their judgements for outputs 1 and 6. The revised judgements are given in table 7.3. Examining outputs 7, 8 and 9 we see that the ratios are in fact smaller than $\frac{t_\infty(0.95)}{t_\infty(0.75)}$, which may suggest over confidence from the expert. However, if we choose d to be 3, calculate the variance of each output using the 50th and 95th percentiles, and then calculate the fitted 75th percentile using this variance, the difference between the fitted and observed percentiles are small, as can be seen in table 7.4. Given that there is a certain amount of imprecision in the expert's assessments, these differences should be acceptable. The fitted standard deviations are given in table 7.5. We notice another irregularity in table 7.5. Since $\boldsymbol{\beta}$ is unknown, the variance of $\eta(\mathbf{x})$ should increase as k_{23} and k_{15} increase, but this has not happened with these judgements. Either the set of outputs 4,5 and 6 or the set 7,8 and 9 need reconsidering. If we fit $E(\sigma^2)V$ to observations 1 to 6, then we find that significant probability is given to negative outputs for most inputs, which suggests that the expert's variances for outputs 4,5 and 6 are too large. Consequently, we instead fit $E(\sigma^2)V$ to observations 1,2,3,7,8 and 9.

We now consider a hypothetical observation, and elicit a value for $E(\sigma^2)$. Since the expert believes that $\eta(.)$ is very smooth, we fix b_{23} and b_{15} at small values initially. We propose the observation $\eta(0.46, 40.5) = 0.09$. In table 7.6, we give the experts suggested 50th and 95th percentiles for the output at various inputs, and our fitted 50th and 95th percentiles using a suitable value of $E(\sigma^2)$.

In figure 7.7 we show the fitted 50th, 75th and 95th percentiles of the output at various inputs. In the top three graphs, k_{15} is fixed at three different values, and in the bottom three graphs, k_{23} is fixed. The dots show the original percentiles provided by the expert, and the squares indicate the three outputs that were not considered when fitting the normal inverse gamma distribution. In figures 7.8 and 7.9 we show realisations of $\eta(\mathbf{x})$ for various values of b_{23} and b_{15}. The expert believes the function $\eta(.)$ to be very smooth, and so fixing both b_{23} and b_{15} to be 0.1 is appropriate here.

n	k_{23}	k_{15}	$\eta_{0.5}(\mathbf{x})$	$\eta_{0.75}(\mathbf{x})$	$\eta_{0.95}(\mathbf{x})$
1	0.02	1	0.007	0.008	0.02
2	0.02	40.5	0.01	0.015	0.025
3	0.02	80	0.015	0.02	0.03
4	0.46	1	0.08	0.1	0.2
5	0.46	40.5	0.08	0.1	0.2
6	0.46	80	0.08	0.1	0.25
7	0.9	1	0.15	0.2	0.25
8	0.9	40.5	0.15	0.2	0.25
9	0.9	80	0.15	0.2	0.25

Table 7.1: The elicited percentiles of the output at various inputs

n	k_{23}	k_{15}	$\frac{\eta_{0.95}(\mathbf{X})-\eta_{0.5}(\mathbf{X})}{\eta_{0.75}(\mathbf{X})-\eta_{0.5}(\mathbf{X})}$
1	0.02	1	13
2	0.02	40.5	3
3	0.02	80	3
4	0.46	1	6
5	0.46	40.5	6
6	0.46	80	8.5
7	0.9	1	2
8	0.9	40.5	2
9	0.9	80	2

Table 7.2: The expert's tail ratios

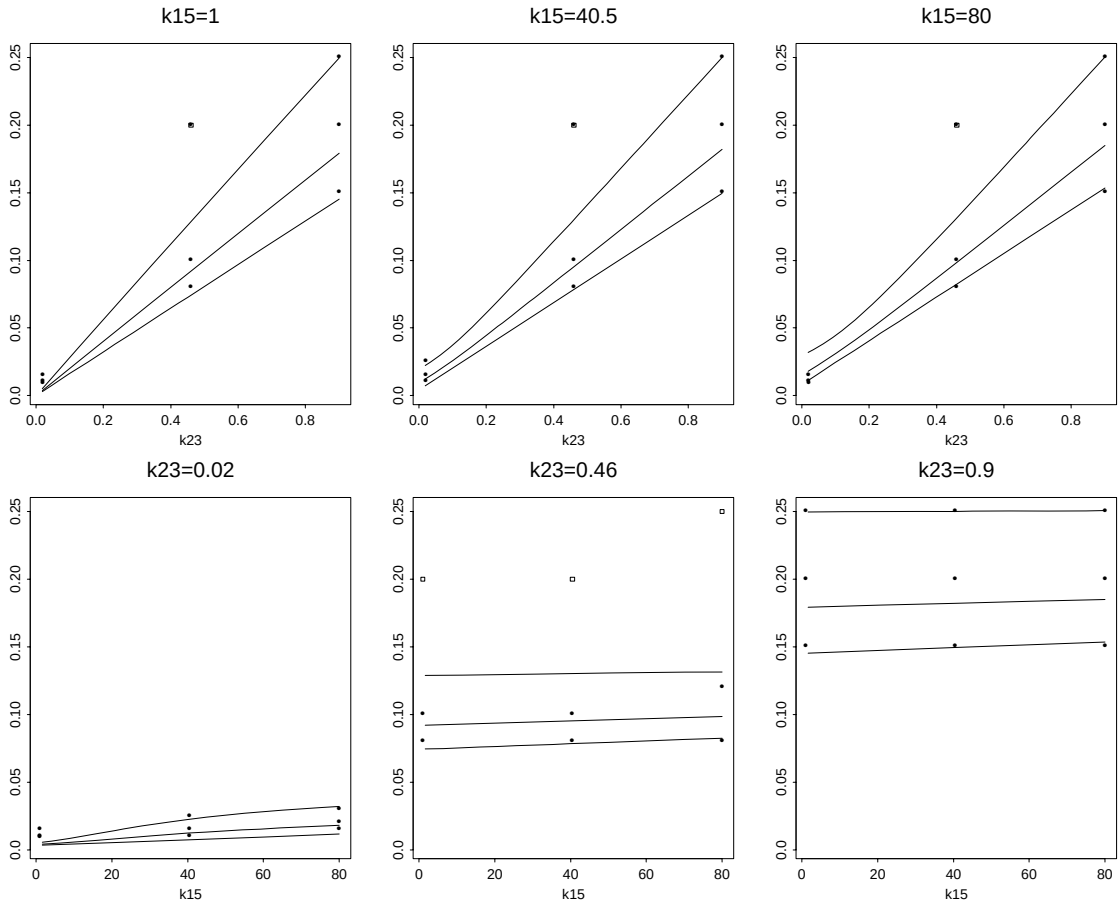

Figure 7.7: Fitted percentiles of the output at various inputs

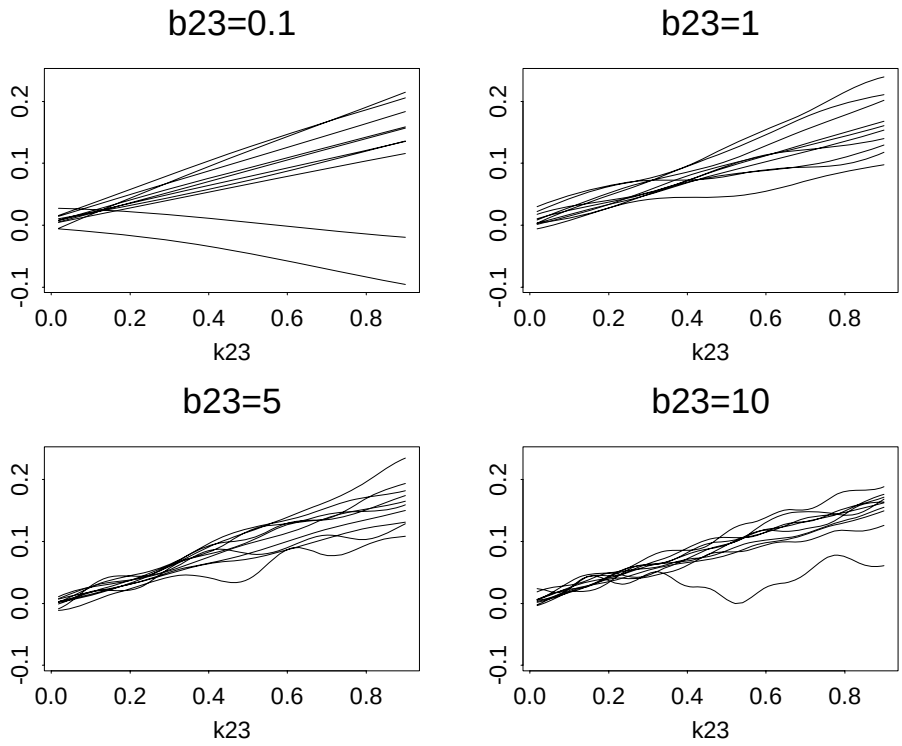

Figure 7.8: Simulated functions at four values of b_{23}

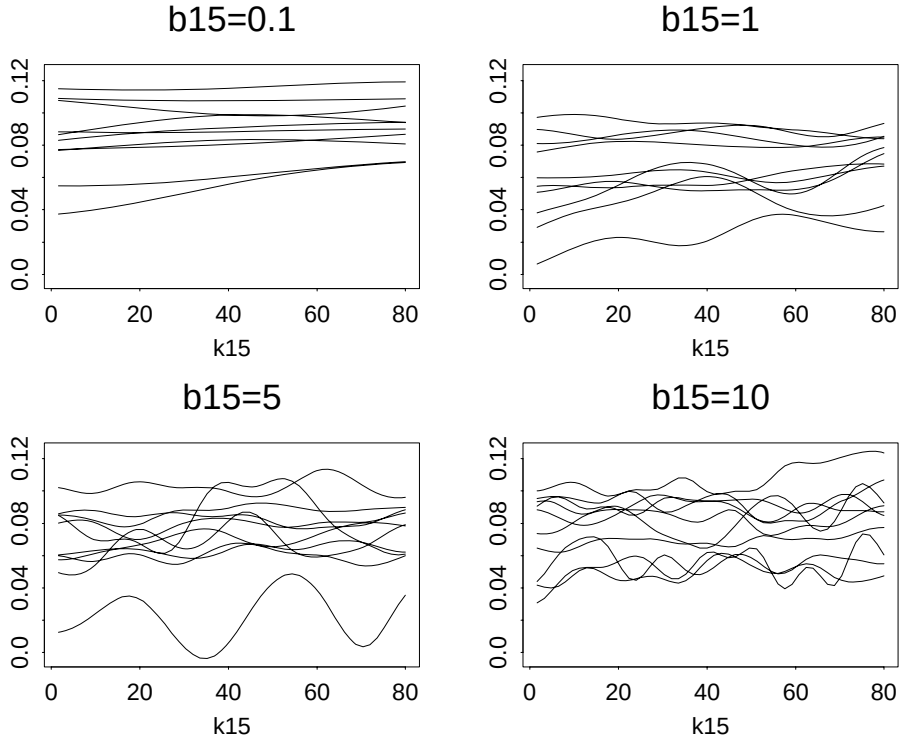

Figure 7.9: Simulated functions at four values of b_{15}

n	k_{23}	k_{15}	$\eta_{0.5}(\mathbf{x})$	$\eta_{0.75}(\mathbf{x})$	$\eta_{0.95}(\mathbf{x})$
1	0.02	1	0.009	0.01	0.15
6	0.46	80	0.08	0.12	0.25

Table 7.3: Revised judgements following examination of tail ratios

n	k_{23}	k_{15}	$\eta_{0.75}(\mathbf{x})$	$\hat{\eta}_{0.75}(\mathbf{x})$
1	0.02	1	0.01	0.011
2	0.02	40.5	0.015	0.015
3	0.02	80	0.02	0.020
4	0.46	1	0.1	0.119
5	0.46	40.5	0.1	0.119
6	0.46	80	0.1	0.135
7	0.9	1	0.2	0.183
8	0.9	40.5	0.2	0.183
9	0.9	80	0.2	0.183

Table 7.4: Elicited and fitted 75th percentiles assuming three degrees of freedom

7.5.1 Combining the prior distribution with runs from the code

We first compare the fitted prior distribution with runs from the grain model. In figure 7.10 we show six plots of the output with one of the input parameters fixed. The solid lines show the prior mean, and a central ninety percent interval for $\eta(\mathbf{x})$. The dots indicate the true outputs. We can see that the expert has provided a fairly accurate estimate of the true output of the model. We now see if using this proper prior distribution offers any improvements in an uncertainty analysis. We consider estimating the distribution function of the output. For this exercise we assume independent normal distributions for the two inputs, both truncated at zero. We first choose four outputs to evaluate the model at, and then combine the data with the expert prior distribution to derive the posterior. We estimate the distribution

n	k_{23}	k_{15}	fitted sd
1	0.02	1	0.003
2	0.02	40.5	0.006
3	0.02	80	0.006
4	0.46	1	0.051
5	0.46	40.5	0.051
6	0.46	80	0.072
7	0.9	1	0.042
8	0.9	40.5	0.042
9	0.9	80	0.042

Table 7.5: Fitted standard deviations

k_{23}	k_{15}	$\eta_{0.5}(\mathbf{x})$	$\hat{\eta}_{0.5}(\mathbf{x})$	$\eta_{0.95}(\mathbf{x})$	$\hat{\eta}_{0.5}(\mathbf{x})$
0.46	50	0.09	0.0908	0.11	0.0912
0.6	40.5	0.1	0.116	0.13	0.118
0.6	50	0.1	0.117	0.13	0.120

Table 7.6: Elicited and fitted percentiles conditional on a hypothetical observation

function of the true output using the simulation procedure. We now estimate the distribution function using a weak prior distribution for $\eta(.)$. The same values of the smoothing parameters are used, and the same form is chosen for $h(.)$. There is no constant term in $h(.)$, since the expert knew that the output would be zero if both inputs were zero. In the weak prior case, we also include the observation $\eta(\mathbf{0}) = 0$, so that we have five observations in total.

In figure 7.11 we plot the median distribution function using the proper prior (as a solid line), using a weak prior (as the dotted line), and the true distribution function (as the squares). We can see that the two estimates are very similar, and give a good estimate of the true distribution function. In figure 7.12, we show a ninety five percent interval for the distribution function using the proper prior (as the solid lines), using the a weak prior (as the dotted lines), and the true distribution function

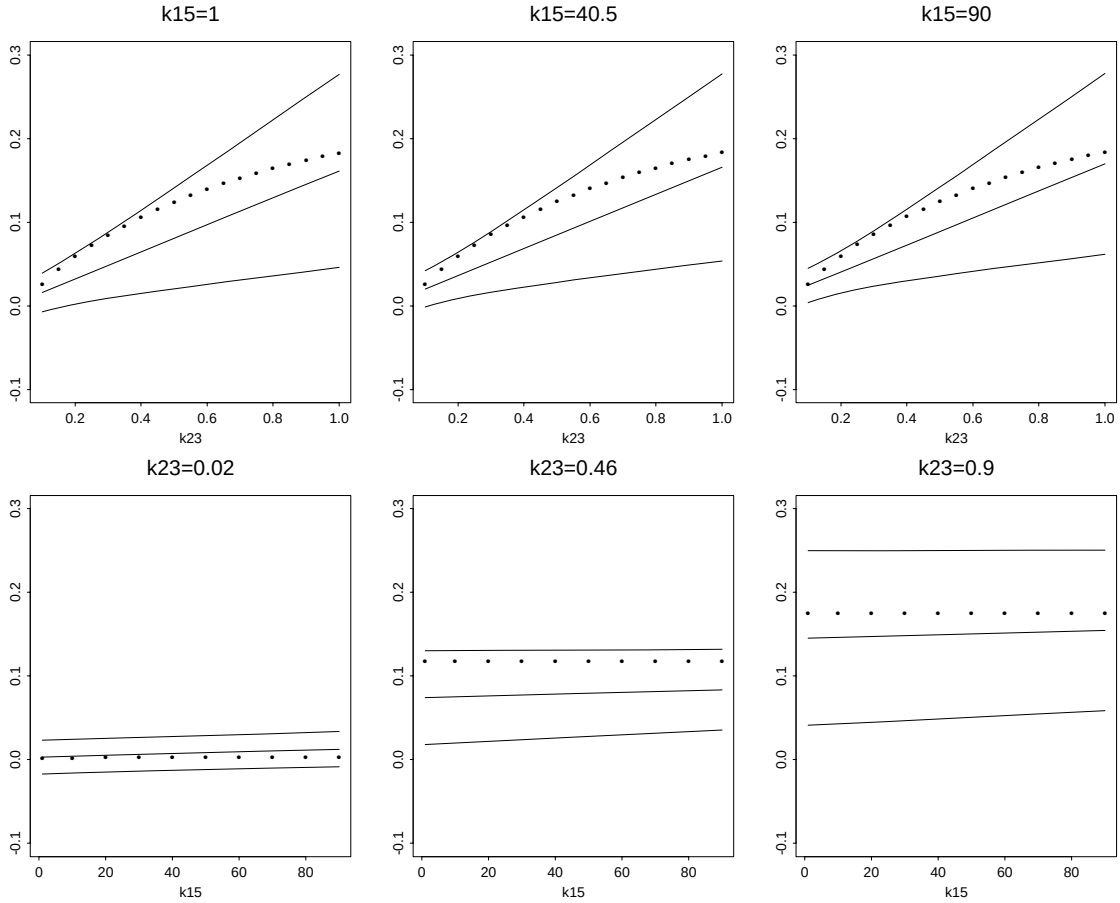

Figure 7.10: True outputs of the grain model

(as the squares). We can now see that there is clear reduction in the uncertainty about the true distribution function when expert prior knowledge is used.

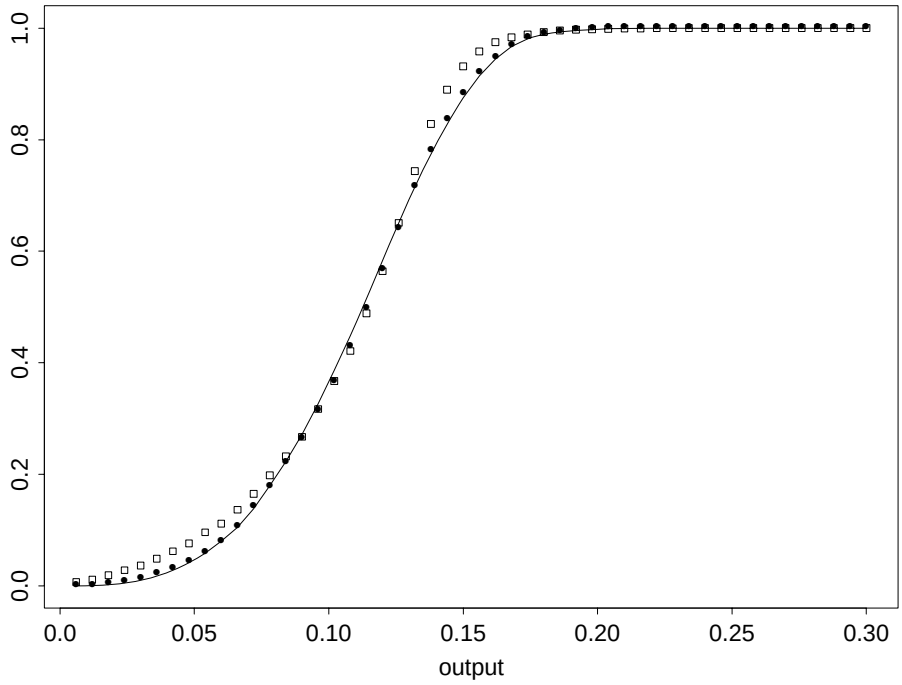

Figure 7.11: Estimates of the distribution function

7.6 Conclusions

The objective in this chapter was to provide a means of including expert prior knowledge about the unknown function $\eta(.)$ in an uncertainty analysis. We have shown that this can be done whilst only asking the expert questions about observable quantities. Furthermore, in the example we considered, it was demonstrated that if the quantity of available data is small, then the benefits of using proper prior beliefs are substantial.

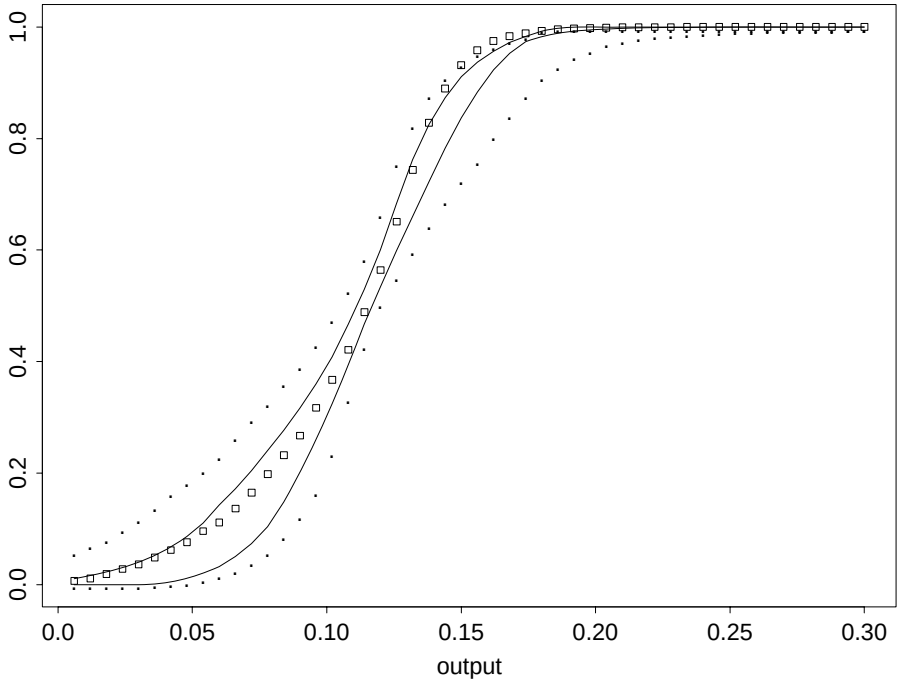

Figure 7.12: Uncertainty about the distribution function

Chapter 8

Discussion

We have considered the problem of an individual wishing to use a computer model to make a prediction, but having uncertainty about the true values of some or all of the model inputs, and being restricted to running the model at a relatively small number of distinct input values. After describing their uncertainty about the true inputs $\mathbf{X}$ through a distribution $G(\mathbf{x})$, the user then wishes to know about the distribution of the true output Y. In particular, they will wish to know if their uncertainty about Y is too great for the model to be effective, or if their knowledge about $\mathbf{X}$ is sufficient for answering the questions of interest. A standard method for learning about Y is to use a simple Monte Carlo approach to obtain a random sample of outputs from the distribution of Y. This is not viable in this scenario, as it will be impractical to obtain the large sample that will typically be required, due to the computing times involved. The goal of this thesis has been to offer an improvement over the Monte Carlo approach in terms of computing times required to make accurate inferences about Y.

The user may have to make a decision based on whether or not Y will exceed some critical value c. For this purpose, we can estimate the distribution function $F_Y(y)$. Once the user has decided how certain they need to be that $Y < c$, they can see from the estimate of the distribution function if the model can be used as an aid for making the decision without knowing the true value of $\mathbf{X}$. Alternatively, the user might simply wish to know what values of Y are plausible given their distribution

$G(\mathbf{x})$ for $\mathbf{X}$. We can provide the user with an estimate of the density function, which will provide a good graphical summary of the uncertainty in Y. Again, from this the user can decide whether it is necessary to learn the true value of $\mathbf{X}$ before the model can be informative.

In chapters four and five we developed means of estimating both these summaries based on a small sample of data. This was achieved by using a Bayesian approach to learn about the model itself. Provided the function $\eta(.)$ is fairly smooth, each time we run the model at an input $\mathbf{x}$, we gain information about $\eta(\mathbf{x}')$ for $\mathbf{x}'$ in the neighbourhood of $\mathbf{x}$. This has enabled us to obtain accurate estimates more efficiently than by using Monte Carlo methods, as the Monte Carlo approach ignores the information that $\eta(\mathbf{x})$ gives us about $\eta(\mathbf{x}')$. We do not expect the improvement to be so large for rough functions $\eta(.)$. The rougher the function $\eta(.)$ is, the less we will be able to learn about $\eta(.)$ from each run. We have used the Gaussian process model to describe our uncertainty about $\eta(.)$. This has placed restrictions on the cases where our methods can be applied. In particular, we have assumed $\eta(.)$ to be continuous. Applying our approach to functions with discontinuities is an area for future research.

We have also shown that expert knowledge about the computer code itself can be included in the analysis, and this can reduce further the number of runs of the code needed. When eliciting expert beliefs, the expert will only have to make statements about observable quantities, i.e., their judgements about the output at various inputs. There can be substantial benefits in using expert prior knowledge here, since the quantity of data available, the number of runs of the code, will typically be small.

Haylock and O'Hagan (1996) and Haylock (1997) pioneered the Bayesian approach to uncertainty analysis and presented a technique for efficiently making inferences about the mean and variance of Y. We have been successful in adding to this a method for learning about two further summaries of Y, the distribution and density functions, a means of choosing design points to make inferences about percentiles of Y, and a method for including expert prior knowledge about the computer code. We now discuss some of the areas in uncertainty anaylsis that have not

been investigated here, and highlight some the difficulties we have encountered that have not been fully resolved.

A general issue in uncertainty analysis that we have not considered is the sensitivity of any summary of the true output Y to the choice of the input distribution $G(\mathbf{x})$. This is an issue that is relevent to both the Bayesian and Monte Carlo approaches. When a distribution $G(\mathbf{x})$ is elicited from the expert, there is likely to be some imprecision in the elicitation. For example, in the SIMPOL model, an expert specified their 5th and 95th percentiles for each unknown input parameter, and a log normal distribution was fitted to these judgements. Suppose for example that the expert believes that the 95th percentile of an input paramter x is approximately $p_{0.95}$. It is unlikely that given the two statements

$$P(X \leq p_{0.95}) = 0.95, \tag{8.1}$$

and

$$P(X \leq p_{0.95} + \delta) = 0.95, \tag{8.2}$$

for some small value of δ, the expert could state which of these two statements they believe to be true. Consequently, there are likely to be a range of lognormal distributions that could be fitted to $\mathbf{X}$ that the expert would believe to be appropriate. Secondly, the choice of a lognormal distribution may not be the only distribution that gives a suitable description of the uncertainty about $\mathbf{X}$. A complete uncertainty analysis should investigate the robustness of any inference to the exact choice of $G(\mathbf{x})$.

The Bayesian approach to uncertainty analysis is centered on modelling an unknown function as a Gaussian process. However, we have seen that the use of the Gaussian process model in this context is not entirely straightforward. We have encountered two complications; the choice of $\mathbf{h}(.)$ in the prior mean, and dealing with the uncertainty in the smoothing parameters B.

In chapter two we saw that when specifying the prior mean of $\eta(\mathbf{x})$, which is given by $\mathbf{h}(\mathbf{x})^T\boldsymbol{\beta}$, including additional terms in the vector of functions $\mathbf{h}(.)$ does not necessarily give better results, even if intuitively we would expect them to. In the example we considered, the function $\eta(x)$ had a linear term in x, and yet including

a linear term in $h(.)$ could lead to greater uncertainty about $\eta(.)$, depending on the data. This is an area that needs further investigation. In the ^{131}I model and SIMPOL examples, we obtained good results without considering the best choice of $\mathbf{h}(.)$. With more careful thought about what function $\mathbf{h}(.)$ to use, it might be possible to improve the results further. This issue also complicates the process of eliciting prior beliefs. The expert may know from their knowledge about $\eta(.)$ that a particular form for $\mathbf{h}(.)$ is appropriate, and yet an alternative form may still give better results once the data has been observed. Finally, if there is uncertainty about the exact choice of $\mathbf{h}(.)$, then we should investigate any uncertainty that this induces in the posterior distribution of $\eta(.)$.

The other difficulty encountered has been in dealing with the smoothing parameters B. In our model for $\eta(.)$, we have conditioned on a posterior estimate of B, and so we have not allowed for the uncertainty we have about B. In addition, we have seen that in some cases, the data does not give much information about B. This suggests that eliciting expert beliefs about B might prove particularly useful here. Given that we have kept B fixed, we should also determine how robust our inferences about Y are to changes in B.

Throughout this thesis we have restricted our attention to computer models with scalar outputs. The extension to vector outputs is another area for future research, and a means of describing the correlations between individual elements in a vector output will need to be considered.

The methods for uncertainty analysis given here could be improved by considering the sensitivity of the output to the individual elements of the input vector $\mathbf{x}$. Understanding which inputs are the most influential could lead to a better choice of design points, both in choosing the initial design points to derive the posterior distribution of $\eta(.)$, and in choosing the simulation design points when using the simulation method described in chapter three.

We have given effective demonstrations of the value of the Bayesian approach in examples where the input is low-dimensional. It remains to be seen how effective our methods are for models with a high-dimensional input. It is these models which are more likely to be computationally expensive. With a high-dimensional input,

it will be necessary to evaluate the model at a larger number of distinct inputs to obtain good information about $\eta(\mathbf{x})$ for the values of $\mathbf{x}$ of interest. This may result in problems when trying to invert the correlation matrix A. It will also be necessary to use far more simulation design points when using the simulation method. This is another reason for investigating means of determining influential inputs.

Another issue that has not been considered is the contribution to the uncertainty in Y from the individual elements in $\mathbf{X}$. Note that this is not the same as performing a sensitivity analysis. Considering a scalar function $y = \eta(x)$, the output y may be sensitive to the value of x, but depending on $G(x)$, our posterior probabilty $P(\eta(X) < c)$ for some critical value of c may still be very small. In general, there may be interest in ascertaining which elements in $\mathbf{X}$ if learnt would most reduce the uncertainty in Y.

Finally, there are more complex scenarios which will require uncertainty analyses. Kennedy and O'Hagan (1998) consider a situation where there are real observations available in addition to the computer code with unknown inputs. In this case, there is interest in making the prediction using both the real data and the information from the computer code. The uncertainty in the computer inputs will be propagated through to the final predictions, and so it will be necessary to quantify the uncertainty that results from these unknown inputs.

Bibliography

Adams, N. and Fell, T. P. (1988). Recycling and metabolic models for internal dosimetry: with special reference to iodine, *Radiation Protection Dosimetry*, **22(3)**.

Brown, J. and Simmonds, J. R. (1995). FARMLAND: a dynamic model for the transfer of radionuclides through terrestrial foodchains, Tech. Rep. NRPB-R273, National Radiological Protection Board.

Campbell, M. J. and Gardner, M. J. (1988). Calculating confidence intervals for some non-parametric analyses, *British Medical Journal*, **296**: 1454–1456.

Clarke, R. H. (1979). The first report of a working group on atmospheric dispersion: a model for short and medium range dispersion of radionuclides released to the atmosphere, Tech. Rep. NRPB-R91, National Radiological Protection Board.

Craig, P., Goldstein, M., Seheult, A. H. and Smith, J. A. (1996). Bayes linear strategies for matching hydrocarbon reservoir history, in *Bayesian Statistics 5*, edited by Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., pp. 69–95, Oxford: University Press.

Crick, M. J., Hofer, E., Jones, J. A. and Haywood, S. M. (1988). Uncertainty analysis of the foodchain and atmospheric dispersion models of marc, Tech. Rep. NRPB-R184, National Radiological Protection Board.

Currin, C., Mitchell, T. J., Morris, M. and Ylvisaker, D. (1991). Bayesian prediction of deterministic functions with applications to the design and analysis of computer experiments, *Journal of the American Statistical Association*, **86**: 953–963.

Diaconis, P. (1988). Bayesian numerical analysis, in *Statistical Decision Theory and Related Topics IV*, edited by Gupta, S. S. and Berger, J. O., vol. 1, pp. 163–175, New York: Springer-Verlag.

Dunning, D. E., Schwarz, J. R. and Schwarz, G. (1981). Variability of human thyroid characteristics and estimates of dose from ^{131}I, *Health Physics*, **40**.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association*, **90**: 577–588.

Feller, W. (1966). *An Introduction to Probability Theory and its Applications*, vol. 2, New York: Wiley.

Haylock, R. (1997). *Bayesian Inference about Outputs of Computationally Expensive Algorithms with Uncertainty on the Inputs*, Ph.D. thesis, Department of Mathematics, University of Nottingham.

Haylock, R. G. and O'Hagan, A. (1996). On inference for outputs of computationally expensive algorithms with uncertainty on the inputs, in *Bayesian Statistics 5*, edited by Bernardo J. M., Berger J. O., D. A. P. and M., S. A. F., pp. 629–637, Oxford: University Press.

Helton, J. C., Garner, J. W., McCurley, R. D. and Rudeen, D. K. (1991). Sensitivity analysis techniques and results for performance assessment at the waste isolation pilot plant, Tech. Rep. SAND90-7103, Sandia National Laboratories Albuquerque, New Mexico.

Hjort, N. L. (1996). Bayesian approaches to non- and semiparametric density estimation, in *Bayesian Statistics 5*, edited by Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., pp. 223–254, Oxford: University Press.

Kadane, J. B. and Wolfson, L. J. (1998). Experiences in elicitation (with discussion), *The Statistician*, **47**: 3–19.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**: 457–481.

Kennedy, M. C. and O'Hagan, A. (1998). Bayesian callibration of complex computer models, Tech. Rep. 98-10, Nottingham Statistics Group.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondance between Bayesian estimation on stochastic processes and smoothing by splines, *Ann. Math. Statist.*, **41**: 495–502.

Krzanowski, W. J. (1988). *Principles of Multivariate Analysis, a User's Perspective*, Oxford: University Press.

Lancaster, P. and Tismenetsky, M. (1985). *The Theory of Matrices*, London: Academic Press.

Lindley, D. V. (1956). On a measure of the information provided by an experiment, *Ann. Statist.*, **27**: 986–1005.

Matheron, G. (1963). Principles of geostatistics, *Economic Geol.*, **58**: 1246–1266.

McKay, M. D., Conover, W. J. and Beckman, R. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, **21**: 239–245.

Mitchell, T. J. and Morris, M. D. (1995). Exploratory designs for computational experiments, *J. Statist. Planning and Inference*, **43**: 381–402.

Mitchell, T. J., Morris, M. D. and Ylvisaker, D. (1993). Bayesian design and analysis of computer experiments: use of derivatives in surface prediction, *Technometrics*, **35**: 243–255.

Neal, R. (1999). Regression and classification using gaussian process priors, in *Bayesian Statistics 6*, edited by Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., pp. 69–95, Oxford: University Press.

Nelder, J. and Mead, R. (1965). An iterative method for finding stationary values of a function of several variables, *Computer Journal*, **7**: 308–313.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction (with discussion), *J. Roy. Statist. Soc. Ser. B*, **40**: 1–42.

O'Hagan, A. (1991). Bayes-hermite quadrature, *J. Statist. Planning and Inference*, **91**: 245–260.

O'Hagan, A. (1992). Some Bayesian numerical analysis, in *Bayesian Statistics 4*, edited by Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., pp. 345–363, Oxford: University Press.

O'Hagan, A. (1993). *Kendall's Advanced Theory of Statistics, Volume 2B, Bayesian Inference*, London: Edward Arnold.

O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications (with discussion), *The Statistician*, **47**: 21–35.

O'Hagan, A., Kennedy, M. and Oakley, J. E. (1999). Uncertainty analysis and other inference tools for complex computer codes (with discussion), in *Bayesian Statistics 6*, edited by Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., pp. 503–524, Oxford: University Press.

Poincarré, H. (1989). *Calcul des Probabilitiés*, Paris: Georges Carré.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals, *Journal of the American Statistical Association*, **92**: 894–902.

Sacks, J., Schiller, S. and Welch, W. J. (1989a). Designs for computer experiments, *Technometrics*, **31(1)**: 41–47.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989). Design and analysis of computer experiments, *Statistical Science*, **4**: 409–435.

Saltelli, A., Chan, K. and Scott, M., eds. (1999). *Mathematical and Statistical Methods for Sensitivity Analysis*, New York: Wiley.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Technical Journal*, **27**: 379–423,623–656.

Shewry, M. C. and Wynn, H. P. (1987). Maximum entropy sampling, *J. Appl Statist*, **14**: 165–170.

Silverman, B. W. (1988). *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion), *J. Roy. Statist. Soc. Ser. B*, **61**: 485–527.