Uncertainty in Financial Models of Large and Complex Government Projects

Kevin McNally

Thesis submitted to the University of Sheffield for the degree of Doctor of Philosophy

Department of Probability and Statistics

School of Mathematics and Statistics

September, 2005

Summary

Government financial models, a particular type of deterministic computer model, are created in order to estimate the cost of expensive projects with large time frames. The model is a function of many inputs, most of which are taken to be known. However the value of a small number of inputs $\mathbf{X}$ is unknown. Whilst the precise value of $\mathbf{X}$ is unknown, subjective knowledge about $\mathbf{X}$ can be represented by a joint probability distribution $\mathbf{G}(\mathbf{x})$. As a result of the uncertainty in $\mathbf{X}$, the scalar output of the financial model is the random variable, Y. The main focus of this thesis is in learning about the uncertainty in Y that results from uncertainty in $\mathbf{X}$ (uncertainty analysis), and in determining which elements of $\mathbf{X}$ are most (and least) important in driving the uncertainty in Y (sensitivity analysis).

In principle both uncertainty and sensitivity analyses can be conducted using Monte Carlo. This method requires a large number of model evaluations. We are interested in the case where the computer model is too computationally expensive to make Monte Carlo practical. We consider a Bayesian approach, which uses the Gaussian Process prior for unknown functions in order to make inference about the computer model itself, using a small number of model evaluations. We then use this information about the structure of the computer model in order to perform uncertainty and sensitivity analyses using relatively few runs of the model.

In this thesis, we adapt the standard Gaussian Process prior in order to utilize the additional information we have about the structure of government financial models. We develop methodology for calculating measures of uncertainty and sensitivity based upon a Gaussian Process model. The methodology also utilizes the additional structural information within government financial models. Finally, we develop elicitation methodology for use in determining the joint probability distribution $\mathbf{G}(\mathbf{x})$. We provide an example from the Private Finance Initiative.

Contents

Chapter 1

Introduction to Government Financial Models

In this introductory chapter we discuss the creation of Government financial models, and the associated statistical issues. To understand the statistical issues we need to motivate the need for financial models. We begin this chapter by discussing the evolution of private investment in public services, and the impact of the Private Finance Initiative (PFI). We then introduce the concept of Value For Money (VFM), and finally the need for the Public Sector Comparator (PSC) – a detailed Government financial model. We will then introduce the statistical issues in Government financial models that we will tackle in this thesis.

1.1 Private Finance, Public Services

British governments have for many years sought to move activities away from the public to the private sector. This began back in 1979 after the election of a Conservative Government under the leadership of Margaret Thatcher. The Government of the time harbored strong beliefs that private financing and ownership could increase efficiency, as highlighted in a recent speech given by former chancellor *Lord Lawson*(38). However, this has been the subject of debate, with the *Public Services International Research Unit*(55) stating that the use of private finance allowed the Government to implement their 'neo-liberal' economic policies of reducing the role of the state, and reducing public sector borrowing.

The most high profile of Thatcher's reforms of the public sector was without doubt the privatization of the utilities, the first programme of its kind in the world. The programme began with privatization of the gas industry in 1986, followed by water in 1989, and electricity, which began in 1991. However, a wider reform of the public services was also evolving, with the private sector beginning to play a role in the provision of health, education, transport infrastructure, prisons and the administration of the functions of the state.

Early privately financed projects were designed mainly to evade Government imposed expenditure controls. They made use of off-budget finance (locally raised extra budgetary and self raised funds) which meant that if a service was contracted out it did not count against the body's capital budget. The loophole allowing such abuses was quickly closed after a report by Sir William Ryrie, second permanent secretary to the treasury, which lead to the "Ryrie Rules". These stated that a project funded by the private sector

1. should go ahead only if it could be demonstrated as more cost effective than a comparable publicly funded project;

2. should result in a corresponding reduction of public spending (although this rule was subject to individual exceptions by Ministers).

The Ryrie Rules are generally held to have provided little incentive to seek private funding (see for example *Heald*(24)). However, small scale local services such as cleaning, catering and refuse collection were a notable exception – these projects were successful.

In an attempt to attract more private investment the second of the Ryrie Rules was abolished in 1989. The belief in Government was that public bodies needed the additional incentive of off-budget financing in order to consider private finance. However, this still did not not stimulate any new flow of privately financed projects. The Private Finance Initiative was launched in 1992, and relaxed the first of the Ryrie Rules. In the words of the then chancellor Norman Lamont as quoted from the *House of Commons Treasury Committee* report(30), "any privately financed project which can be operated profitably will be allowed to proceed". The initiative aimed to encourage projects funded directly by the public through charges, such as the recently finished M6 Toll motorway, which bypasses a busy section of the M6 motorway near Birmingham. However, public sector comparisons were still expected for most other types of project.

The principle of PFI is that a public sector body obtains a service rather than an asset. A private sector contractor funds any asset required and is then paid for the service provided. This translates as (see *House of Commons Treasury Committee* report(30)) *"Government no longer builds roads, it purchases miles of maintained highway ... it no longer builds prisons, it buys custodial services ... it no longer always buys computers and software, but pays for managed IT services"*. Normally, the commissioning body will avoid the need for capital expenditure at the beginning of the project in exchange for making payments for the service as it is delivered, often over a period of up to thirty years. The private finance is temporary: the public sector still pays in the end.

However, the initiative did not have the effect the Government anticipated, with few large projects (over £5 million in value) commissioned in 1993-1994. This lead to a final major effort to force project managers to consider alternative methods of funding. The 1994 "universal testing rule" required public sector project managers to consider private finance for every project. This final push had the desired effect, with a glut of projects commissioned in subsequent years. In *Table*(1.1) we show figures (reproduced from the *House of Commons Treasury Committee* report(30)) detailing estimated expenditure (£m) and the major PFI projects signed off during each year for the period 1986 – 1999.

Year	Total (£m)	Notable projects (included in total for year)
1986	150	Dartford Bridge
1990	330	Second Severn Crossing
1992	324	Birmingham Northern Relief Road. Skye Bridge
1993	42	Royal Armouries Museum
1994	11	Lothian Forth Health Board. Northern NHS Trust
1995	862	London Underground Northern Line Trains
1996	6,064	Channel Tunnel Rail Link (4,300m)
1997	1,500	Manchester Metrolink. Ministry of Defence projects
1998	2,679	Several hospitals
1999	804	National Savings IT. Almond Valley and Seafield Sewage

Table 1.1: PFI projects: 1986-1999

By early 2002 about 500 *PFI* contracts had been signed, of which about half are operational, and at present approximately 15 percent of all publicly sponsored gross capital spending is provided by the private sector. The size of projects signed off has been varied, but the trend has been toward larger projects with massive budgets. Our interest lies in the larger projects, with large timescales.

1.2 Government Financial Models

The PFI bidding process is complex and can take a period of years . In the first instance the project manager has to investigate if the project is suitable for PFI, before bids from the private sector are analysed, and a preferred bidder selected. See HM Treasury publication *How to construct a public sector comparator*(28) for a detailed analysis of the bidding process. The project manager has to decide whether to accept the PFI bid outright, to try to negotiate additional services or a discount, or to reject the PFI bid in favor of conventional funding. This decision is based upon Value For Money (VFM).

1.2.1 Value For Money

The Private Finance Initiative was introduced because the thinking in Government was that the private sector was more efficient than Government agencies. Therefore, the private sector could provide services at a lower cost than conventional means, thus providing the public with Value For Money (VFM). However *Ball et al.*(3) and *Froud*(17) have found PFI may not offer the advantages that its proponents suggest, whilst *Monbiot*(42) suggests the entire PFI process is riddled with corruption. *Heald*(25) provides an excellent overview of PFI and the VFM case.

Value for money is the single most important factor in the decision on whether or not to accept a PFI deal. Since the use of private finance is no longer constrained by the Ryrie Rules, VFM is no longer exclusively based on cost – the additional benefits that PFI funding may or may not provide can strongly influence the public body's decision. However, the additional benefits that PFI may provide will usually only influence the funding decision if the PFI bid is very close to the cost at which the public body estimates they could provide the service themselves. Achieving value for money is especially important in very large projects that go before Parliament in the form of the Public Accounts Committee (PAC).

In order to assess if the proposed PFI deal offers value for money, the public sector body needs to assess how much the project would cost if they were to do it themselves. They do so by producing a detailed financial model, known as the Public Sector Comparator (PSC).

1.2.2 Public Sector Comparators

The PSC is the public sector's *risk adjusted estimate* of the total cost of the project, or informally how much it will cost to provide the service using traditional funding. It is a hypothetical costing, and not to be confused with a genuine Government bid. The PSC is complex, consisting of a detailed timetable of works and a series of costs. Due to the time frame of projects it is typically produced in spreadsheet format, which minimizes the risk of error and allows the PSC to be more easily audited. In the projects that interest us the PSC returns a single output, the Net Present Value (NPV) of the project – the total cost of the project in terms of current prices. Hence, VFM can be assessed by comparing the NPV of the PFI bid with the NPV from the PSC.

1.2.3 Risk

The statistical issues concerned with Public Sector Comparators arise from the word risk in the definition of the PSC. In a project with a lifespan of decades we will have many inherent uncertainties, and the PSC attempts to quantify these in terms of additional costs.

Risks can be sub-divided into two main categories

1. Overoptimism

   When making cost assessments the assessors involved tend to consider the best case scenario. They make assessments of costs based on the project running smoothly and on time, with no unanticipated problems. Experience from past Government projects shows this is rarely the case.

2. Financial Indices

   Because projects will have a lifespan of years, often decades, quantities such as the rate of inflation in future years need to be taken into account in producing an estimate of the NPV of the procurement. Figures such as the rate of inflation cannot be known for certain and so have to be estimated in the model. We discuss the motivation behind the modelling of financial indices, and the associated problems, in section 1.4.

In some Public Sector Comparators risk is treated in a very basic manner, and estimated as a percentage of the total cost of the project. The choice of this multiplier is subjective and of great importance; it will almost certainly have a bearing on whether or not to use private finance. *Monbiot*(42) suggests that in some projects this multiplier is chosen in order to ensure private finance is used.

However, in some of the larger projects that are audited by the National Audit Office (NAO), one of which is the Ministry of Defence (MOD) Main Building Redevelopment (studied in depth later on in this thesis), a more ambitious approach to estimating risks is used. Adjustments for overoptimism can by made by examining the over-run from approval cost on previous relevant projects. The magnitude of the adjustment is unknown, but previous projects will provide information about the bounds of this multiplier. This task that is becoming increasing difficult since a report by the *House of Commons Treasury Committee*(30) notes that the increase in PFI has led to a narrowing scope of reference; that is, there are fewer Government funded projects with which to compare potential PFI projects, and care has to be taken if a comparison is with a project from the distant past since lessons are usually learnt from budget overruns – large overruns from budget will usually be investigated by the PAC.

Uncertainty in financial indices can be taken into account by using past data and the subjective knowledge of financial experts.

A Public Sector Comparator may contain many risks, with very complicated models such as the London Underground PSC containing thousands. It is the uncertainty in these risks that causes problems in decision making

1.3 Uncertainty in Computer Models

It is the uncertainty in the risks that interests us, and the statistical issues associated with this uncertainty are the focus of this thesis. The statistical issues arise from the simple fact that if we are uncertain about some of the inputs into the PSC, then we have uncertainty about the model output, the NPV. We also have the property that if we input the same series of risks into the PSC we get the same answer; if we input different series of risks into the model, we will in general get different answers. That is, given known risks, the NPV is deterministic. However, given the uncertainty in the risks the NPV is stochastic.

Government financial models are an example of a computer model. The field of computer models has been well studied, with an ever growing literature. A deterministic computer model is used to represent a complex system, physical or otherwise. Frequently the system is too costly, difficult or impossible to observe directly. A financial system, which represents a future cost, is impossible to observe. The system is represented by a computer model, and studied by a computer experiment, a process which involves running the computer code at various different input configurations, with the purpose of learning something about the real system. The model is an imperfect representation of the system, and resultantly in error, however if we repeatedly run the model at the same set of inputs we obtain identical output(s).

Computer experiments are often expensive to run, even with modern computational power and supercomputers. As a result we only have a small number of model runs available – the precise number frequently dictated by resources rather than requirements. Inference about the system, therefore, needs to make efficient use of the available runs of the model.

One inference that is often of interest is to use the small amount of data in order to make inference about the whole function, i.e use the data to make inference about the output at the infinite (for continuous inputs) number of untried inputs. We shall develop methodology in this thesis that allows us to do this efficiently, although this is only a secondary aim of this thesis.

Our main interest lies in investigating the uncertainty in the model output that arises due to the uncertainty in the model inputs. This area is known as uncertainty analysis. Since the inputs are random variables, then resultantly the output(s) are also random variables. If we are able to provide upper and lower bounds on each model input, then these provide bounds on the model output(s). For a complex model, this would need to be explored numerically. A more thorough analysis requires the (usually) subjective information about the model inputs to be represented via a joint probability distribution. Not only can we provide more accurate bounds on the output, but we can assess how probable a specific value of the output, or a range of values for the output are. This will also require numerical methods. For our financial models we have to compare the distribution of the NPV with a bid price from PFI. We discuss the area of uncertainty analysis in detail in chapter 2. We discuss both summaries of the output that we wish to calculate and we describe common methods of evaluating these summaries.

Our second primary aim is to quantify which of the model inputs are most and least influential in driving the variation in the model output(s). This is important since it may allow us to address failings of the model with better or more detailed modelling, or allow us to simplify the model if the magnitude of some of the inputs has little effect on the output(s). This area is known as global sensitivity analysis and we introduce this, and methods for assessing sensitivity in chapter 2.

For complex models, uncertainty and sensitivity analysis require numerical methods. Monte Carlo is one such method, and this typically requires many evaluations of the model. This approach is not practical for a computationally expensive model. A method that we review in chapter 3 of this thesis, non-parametric regression using Gaussian Processes, has sought to use features of the model in order to improve efficiency. For some problems, this method is able to reduce the required number of evaluations by an order of magnitude. We will use the special features of financial models in order to further improve the efficiency of this method.

In this thesis we shall develop methodology for:

1. function approximation;
2. uncertainty analysis;
3. sensitivity analysis.

We will develop methodology that can achieve accurate results for high dimensional functions, using relatively few model evaluations.

1.4 Uncertainty in Financial Indices

One of the major factors that induces uncertainty in financial models is the long timescale of the project. Even if all other inputs within the model were known, financial indices would induce substantial uncertainty. This is since although prices may be known when the model is developed, the cost of component x in year y of the project is unknown. The prices of commodities change over time, with prices usually increasing – a property known as inflation. The price in year y will be cx for some unknown multiplier c.

If all prices increased at the same rate over time, inflation would not be a problem since the relative increase in price, the rate at which component x increases in price relative to prices in general, would be zero. However this is not the case (consider for example current oil prices which have increased in price at far above the rate of inflation). This illustrates the need to model how the prices of a PFI project will change relative to a general measure of inflation. We do so by considering two financial time-series, the general measure, which in this thesis we will measure using the GDP deflator, and Tender Price Inflation (TPI), which measures the price changes of the project. We will need to model these in the long term – over the full period of a project (which is frequently decades). The series can be taken to be independent of each other, but serially correlated with themselves. We will only have a small amount of data on these series, however the subjective information of an expert is available.

The final aim of this thesis is to

4 develop methodology for modelling inflation in the long term, when we have only a small amount of data on these series.

The methodology that we develop in order to achieve this final aim is independent of the methodology developed to achieve aims 1-3. Resultantly, this work will be in a self contained chapter within the thesis. We bring together all the methodology in an example chapter at the end of the thesis.

1.5 Application: MOD Main Building

This work is supported by the National Audit Office (NAO). The NAO are responsible for auditing the Public Sector Comparators arising from projects commissioned by central Government, and reporting to Parliaments Public Accounts Committee. Work on the MOD Main Building PSC began in 1998, and the project was signed off in 2001.

We also consider the PSC that was developed in order to estimate the cost of redeveloping the London Underground. This represents another 30 year project, but with a far larger scope and huge budget. The PSC developed for this project contains thousands of unknown inputs. Although we will not study this PSC in detail as an example, we will discuss how the methodology we have developed in the thesis could be used for a project of this magnitude.

Chapter 2

Introduction to Uncertainty and Sensitivity Analysis

In chapter 1 we gave a general introduction to Government financial models, and why they are created. In this chapter we consider the mathematical issues concerning inference that we encounter with computer models, and the associated literature that has addressed these issues.

We first describe the problem of uncertainty analysis in the context of computer models, before considering the more complex issue of sensitivity analysis.

2.1 Uncertainty Analysis

Computer models are a mathematical representation of a complex system (physical system or otherwise). For a deterministic computer model, we have two sources of uncertainty that we may wish to quantify.

- *Analysis of model inadequacy*

  The model usually represents a simplification of the system, and this typically induces systematic errors between the model output and reality. We may wish to quantify these differences.

- *Parameter value uncertainty*

  Often the inputs to the computer model are not fixed, and can take many different values. We may have uncertainty about what the true values of the inputs should be. A parameter value uncertainty analysis is concerned with how uncertainty on the model inputs propagates through the model to the model outputs.

Government financial models are concerned with estimating a hypothetical future cost, and as a result we will never have useful data to compare with the output of our computer model. Therefore, we cannot hope to thoroughly assess model inadequacy. We may, however, be able to identify some obvious flaws or inadequacies in the model.

In Government financial models, the model inputs all represent future values, which are usually individual costs or inflations. Since we are dealing with the future, we have considerable uncertainty about many of our model inputs. As a result of this, we may have considerable uncertainty about the model output(s). Our interest therefore lies in parameter value uncertainty analysis.

We now introduce some notation, and provide a formal definition of what is known in the computer models literature, as the uncertainty distribution.

2.1.1 Notation

We define the inputs, which are the parameters in our financial model to be the p dimensional vector $\mathbf{x}$, and the deterministic scalar output of the model to be y. We represent the computer code by $\eta(.)$ and the relationship between the inputs and the output is given by

$$y = \eta(\mathbf{x}). \tag{2.1}$$

The true input configuration is the random variable $\mathbf{X}$, and as noted in (2.1), the uncertainty in the model inputs induces uncertainty in the model output, so the corresponding output is the random variable Y, where

$$Y = \eta(\mathbf{X}). \tag{2.2}$$

We now need to quantify the uncertainty about the true value of each of the inputs in the model. We assume that our knowledge about $\mathbf{X}$ is represented by the joint probability distribution $G(\mathbf{x})$. In the absence of any data, $G(\mathbf{x})$ represents our subjective beliefs about $\mathbf{X}$ and is formed using probability statements from an expert. For now we presume $G(\mathbf{x})$ is known from objective or subjective information, but we consider the (partial) elicitation of $G(\mathbf{x})$ in chapter 6. Finally we denote the sample space of $\mathbf{X}$ by χ.

For some computer model represented by $\eta(.)$ and some unknown true input $\mathbf{X}$ with distribution $G(\mathbf{x})$, the distribution of Y is the uncertainty distribution.

In uncertainty analysis we want to make inferences about Y.

2.2 Classical uncertainty analysis

For a very simple computer model, $\eta(.)$, it may be possible to obtain summaries of Y analytically by integrating over the joint distribution $G(\mathbf{x})$. Some useful measures for expressing our uncertainty about Y are the expectation, variance, and distribution function.

These summaries require us to calculate the integrals

$$E(Y) = \int_{\mathcal{X}} \eta(\mathbf{x})\, dG(\mathbf{x}), \tag{2.3}$$

$$E(Y^2) = \int_{\mathcal{X}} \eta(\mathbf{x})^2\, dG(\mathbf{x}), \tag{2.4}$$

$$F_Y(s) = \int_{\mathcal{X}} I\{\eta(\mathbf{x}) \leq s\}\, dG(\mathbf{x}), \tag{2.5}$$

where $I\{.\}$ denotes the indicator function.

However, for all but the most trivial of problems, an analytical approach to summaries of Y is not feasible. The function, $\eta(.)$, is sufficiently complex in many models from engineering, chemistry, physical science and geography for us to regard $\eta(.)$ as a black box. We supply the model inputs, $\mathbf{x}$, and after computationally intensive calculations, performed by a computer, y is returned as the output.

Government financial models are more transparent than models of physical systems. The output is a complex function of costs and financial indices, and we may well have information about groups of inputs which cannot possibly interact. However, these models are still far too complex for us to be about to calculate summaries of Y analytically. We can think about the computations within a financial model as comprising of a series of black boxes.

The classical approach to uncertainty analysis uses brute force in order to calculate summaries of Y, using Monte Carlo techniques. We draw a random sample of inputs $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$ from $G(\mathbf{x})$ and evaluate the function at each of these points in order to obtain corresponding outputs $y_1 = \eta(\mathbf{x}_1), y_2 = \eta(\mathbf{x}_2), \ldots y_n = \eta(\mathbf{x}_n)$. We estimate summaries of the distribution of Y from this sample of model evaluations. For example our integrals, (2.3)-(2.5), are estimated by their sample equivalents

$$\hat{E}(Y) = \frac{1}{n}\sum_{i=1}^{n}\eta(\mathbf{x}_i), \tag{2.6}$$

$$\hat{E}(Y^2) = \frac{1}{n}\sum_{i=1}^{n}\eta(\mathbf{x}_i)^2, \tag{2.7}$$

$$\hat{F}_Y(s) = \frac{1}{n}\sum_{i=1}^{n}I\{\eta(\mathbf{x}_i) \le s\}. \tag{2.8}$$

However, for a deterministic model, classical Monte Carlo techniques can be inefficient. Often we find that the variability in Y is dominated by only a few of the components of $\mathbf{X}$. If we only have a few active inputs, and m of our inputs are dormant, the design points are projected onto a $p-m$ dimensional hyperplane. In a random design, if this projection induces some clustering of design points, then some parts of the design space may be sparsely covered. Resultantly, more design points may be required for precise inferences than with more efficient designs.

An alternative to random sampling was proposed by *McKay et al.*(40) for use in deterministic computer models. Latin Hypercube Sampling (LHS) is a stratified sampling technique that ensures the margins are well covered. In its simplest form, where the inputs are all independent or we have independent vectors of inputs, we partition the range of each of the marginal distributions x_j into n intervals of equal probability. We then randomly sample one value from each interval. The x_1 values are paired at random and without replacement with the x_2 values in order to gives us n pairs. These pairs are combined randomly and without replacement with x_3, and we continue in this manner until we have our set of n design points. In the more complex case when the inputs or input vectors are correlated, the restrictive pairing technique of *Iman and Conover* (31) can be used to induce the correct rank correlation structure.

We illustrate the advantages of a LHS by way of a simple two dimensional example:

$$y = \eta(x, z), \tag{2.9}$$

where the trues values of x and z are unknown. We assign $N(0, 1)$ distributions to the inputs and generate 20 design points using LHS and random samples.

We show the design points generated using these two methods in *Figure* (2.1).

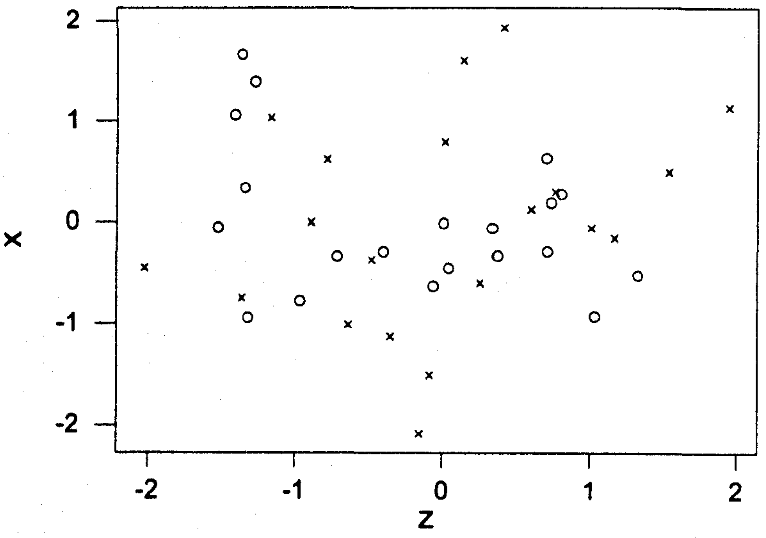

Figure 2.1: Design points under LHS (crosses) and random (circles) designs

Now if we suppose that the model output is insensitive to z, then (2.9) reduces to a function of one uncertain input,

$$y = \eta(x). \tag{2.10}$$

Our design points are now projected down onto the x-plane, which we show in *Figure* (2.2). When the points from the LHS are projected from the 2 dimensional sample space onto the x space, the $n = 20$ intervals of equal probability are retained. The random sample covers the design space well 1 standard deviation

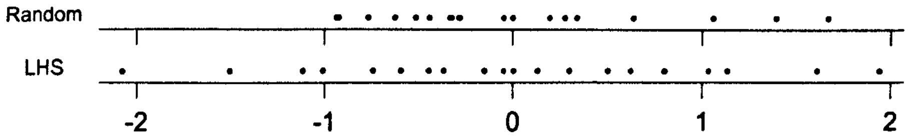

Figure 2.2: Marginal Samples

either side of the mean, but only covers the tails sparsely. Some of the design points from the random sample provide us with very little information.

Suppose that for a problem of p uncertain inputs we chose a p dimensional Latin hypercube design. If the output is insensitive to m of these inputs, then the design points are projected onto an $p - m$ dimensional hyperplane. Resultantly each equal probability interval now has many points, rather than a single point.

When we have input sparsity, we find that in general a Latin hypercube sample contains more information than a random sample of the same size. In their empirical work *McKay et al.*(40) found that $Var\{\hat{E}(Y)\}$ and $Var\{\hat{Var}(Y)\}$ were smaller when using LHS compared with a random sample. It is this reduction in variance that makes LHS preferable to a random sample. *Saliby and Pacheco*(60) and *Helton and Davis*(27) have also undertaken numerical work in this area.

2.3 Sensitivity Analysis

Sensitivity analysis, in the context of computer models, is concerned with understanding how changes in the inputs, $\mathbf{x}$, influence the output, y. There are two types of sensitivity analysis: *local* and *global* analyses.

- *Local* sensitivity analysis is concerned with small changes about some central case of interest, $\mathbf{x}_0$. Local sensitivity about one location may be completely different from that about another location. A local sensitivity analysis is based upon partial derivatives of $\eta(.)$, with respect to the inputs, evaluated at $\mathbf{x_0}$.

- *Global* sensitivity analysis considers more substantial changes in the inputs, and is useful when our knowledge about the 'true' input configuration is more vague. We regard the input vector as a random variable, with distribution $G(\mathbf{x})$. A global analysis attempts to answer the question; 'how important are the individual elements in $\mathbf{X}$ with respect to the uncertainty in $\eta(\mathbf{X})$'?

We only consider measures of global sensitivity analysis.

Measures of global sensitivity analysis vary in both computational burden, and in the quality of information they provide. Unsurprisingly, the more sophisticated measures of global sensitivity analysis require many more evaluations of $\eta(.)$ than the more basic techniques.

The most basic measures of global sensitivity analysis are screening techniques, discussed in detail in *Campolongo, Kleijnen and Andres*(7). These measures are able to attach a qualitative measure of importance to each of the inputs. *Morris*(43) devised methodology that was able to rank the inputs by importance with just $r(p+1)$ evaluations of $\eta(.)$, where $5 < r < 15$. However, the method is based on just the main effects, where the main effect of input x_i is the variability in $\eta(.)$ due to input x_i after averaging over all other variables. As a result the method may rank the inputs, in terms of their total importance, incorrectly. The method was extended by *Campolongo and Braddock*(6) in order to take into account main effects and first order interactions, but at an increased computational burden ($n = O(p^2)$).

The relatively small computational burden of screening techniques is an attractive feature. However, knowing just that input X_i has more importance than X_j, with respect to the variability of Y, is of limited value. For this reason, screening methods are often used as the first stage of a two staged analysis (see for example *Campolongo, Tarantola and Saltelli* (8)), when the dimension of $\mathbf{X}$ is too large for more sophisticated methods. The second stage of the sensitivity analysis comprises of a more advanced technique on the reduced set of model inputs.

Quantitative measures of sensitivity analysis are more useful for a meaningful sensitivity analysis. These methods are able to inform us of how much more importance X_i has than X_j with respect to the variability of Y. However, these require many more function evaluations than screening methods. We consider two of these measures further.

2.3.1 Main Effects and Interactions

One of the widely used measures of sensitivity (*Sobol'*(62)(63); *Jansen, Rossing and Daamen*(32); *Saltelli, Tarantola and Chan*(9)(61); *Chan et al.*(10)) is based on a decomposition of $\eta(.)$ into main effects and interactions (2.11). *Sobol'* termed this a decomposition into summands of different dimensions. We write

$$y = \eta(\mathrm{x}) = E(Y) + \sum_{i=1}^{p} z_i(x_i) + \sum_{1\leq i\leq j} z_{ij}(x_i, x_j) + \ldots + z_{1,\ldots,p}(\mathrm{x}), \tag{2.11}$$

where

$$\begin{aligned}
z_i(x_i) &= E(Y\,|\,x_i) - E(Y), && (2.12)\\
z_{ij}(x_i, x_j) &= E(Y\,|\,x_i,\, x_j) - z_i(x_i) - z_j(x_j) - E(Y), && (2.13)\\
z_{ijk}(x_i, x_j, x_k) &= E(Y\,|\,x_i,\, x_j,\, x_k) - z_{ij}(x_i, x_j) - z_{ik}(x_i, x_k) \\
&- z_{jk}(x_j, x_k) - z_i(x_i) - z_j(x_j) - z_k(x_k) - E(Y), && (2.14)\\
&\vdots \\
z_{1,\ldots,p}(\mathrm{x}) &= E(Y\,|\mathrm{x}) - \sum_{i=1}^{p} z_i(x_i) - \sum_{1\leq i\leq j} z_{ij}(x_i, x_j) - \ldots - E(Y). && (2.15)
\end{aligned}$$

The term $z_i(x_i)$ is referred to as the main effect of variable x_i, $z_{ij}(x_i, x_j)$ as the interaction between variables x_i and x_j and so on.

Oakley and O'Hagan(51) show that we find the terms in (2.11) by integration. Defining $\mathbf{x}_{-\mathbf{i}}$ and $\mathbf{x}_{-\mathbf{ij}}$ as the vector containing all inputs but the i^{th} and all except the i^{th} and j^{th} respectively, then

$$E(Y) = \int_{\chi} \eta(\mathbf{x})\, dG(\mathbf{x}), \tag{2.16}$$

$$z_i(x_i) = \int_{\chi_{-i}} \eta(\mathbf{x})\, dG_{\mathbf{x}_{-\mathbf{i}}|x_i}(\mathbf{x}_{-\mathbf{i}}|x_i) - E(Y), \tag{2.17}$$

$$z_{ij}(x_i, x_j) = \int_{\chi_{-ij}} \eta(\mathbf{x})\, dG_{\mathbf{x}_{-\mathbf{ij}}|x_{ij}}(\mathbf{x}_{-\mathbf{ij}}|x_{ij}) - z_i(x_i) - z_j(x_j) - E(Y), \tag{2.18}$$

where following *Oakley and O'Hagan*, we use χ_{-i} to denote the space of possible values for $\mathbf{x}_{-\mathbf{i}}$, and $G_{\mathbf{x}_{-\mathbf{i}}|x_i}(\mathbf{x}_{-\mathbf{i}}|x_i)$ denotes the conditional distribution of $\mathbf{X}_{-\mathbf{i}}$ given X_i. The higher order terms of (2.11) follow similarly. *Chan et al.*(9) note that typically as the order of the integral increases, then $z_{1,\ldots,r}(x_1, \ldots x_r) \rightarrow 0$. That is, the high order interactions are often negligible compared with the main effects and low order interactions.

If we first scale our inputs so that they have the same range (we re-scale our inputs to $[0, 1]$), plots of main effects and first order interactions over the range of their marginal distributions provide a powerful graphical tool for assessing the influence of our inputs with respect to the magnitude of y. For unbounded inputs we might need to consider an $\alpha\%$ (e.g $\alpha = 99\%$) interval and scale these to $[0, 1]$.

2.3.2 Variance Based Methods

Whilst the decomposition into main effect and interactions, (2.11), is able to identify the role of x_i in the function $\eta(.)$, it is unable to assess the importance of the uncertain quantity X_i with respect to the uncertainty in Y. The importance of X_i depends on both the distribution of X_i, and the role of x_i in the function $\eta(.)$.

We illustrate with the function

$$\eta(x_1, x_2) = x_1 + 1.3x_2^2, \tag{2.19}$$

where we have independent inputs, and $x_1 \sim U(-1,1)$ and $x_2 \sim N(0,1)$. We re-scale the inputs and plot the main effects in *Figure* (2.3).

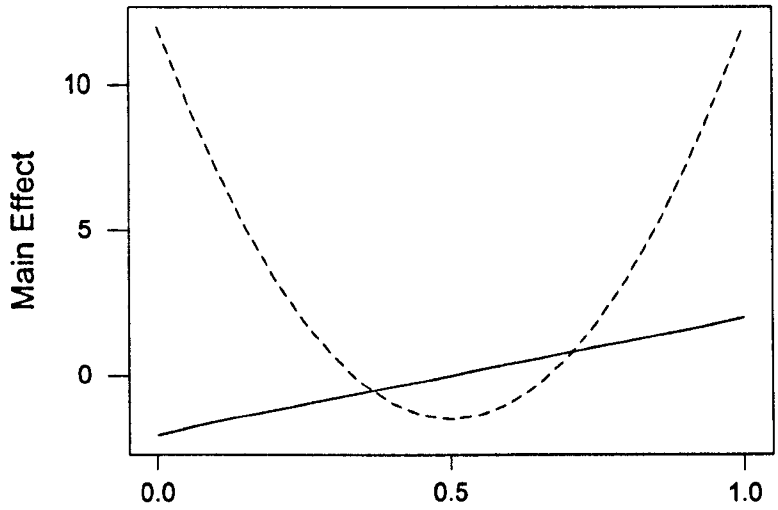

Figure 2.3: Main effects of x_1 (bold) and x_2 (dash)

From *Figure* (2.3) we can see that input x_2 has most influence on the magnitude of the output. However, $z_2(x_2)$ shows the most rapid rate of change in the tails, where X_2 has little probability. In order to assess the importance of X_1 and X_2 it is clear that we need to take into account their respective marginal distributions. Our second quantitative measure does so, using variance in order to assess the importance of the uncertain quantity X_i with respect to the uncertainty in Y.

Variance based methods of sensitivity analysis as recently reviewed by *Chan* *et al.* (10) quantify the sensitivity of the output Y to the model inputs in terms of the reduction in the variance of Y. *Oakley and O'Hagan* (51) have formally justified this approach in a Bayesian setting in terms of quadratic loss.

For independent inputs, we can decompose the variance (*Sobol'*(63)) as

$$V = Var[Y] = \int \eta(\mathbf{x})^2 \, dG(\mathbf{x}) - [E(Y)]^2 = \sum_{i=1}^{q} V_i + \sum_{1 \leq i \leq j} V_{ij} + \ldots + V_{1,\ldots,q}, \quad (2.20)$$

where

$$V_i = Var[z_i(x_i)], \quad (2.21)$$

$$V_{ij} = Var[z_{ij}(x_{ij})], \quad (2.22)$$

are known as the partial variances of the main effect $z_i(x_i)$, the interaction $z_{ij}(x_{ij})$ and so on. With some dependence between the inputs we can achieve a similar decomposition to (2.20) for independent sub-vectors.

For a variance based approach we have two principal measures of sensitivity:

$$V_i = Var\left[E[Y \mid X_i]\right], \quad (2.23)$$

is the expected amount by which the uncertainty about Y would be reduced if we learnt the true value of X_i. It is referred to in the literature as the main effect variance. Since $V_i = Var\{z_i(x_i)\}$, this measure of sensitivity is especially useful when the main effects explain most of the variance.

Our second measure

$$V_{Ti} = Var[Y] - Var\left[E[Y \mid \mathbf{X}_{-\mathrm{i}}]\right], \quad (2.24)$$

is the expectation of the variance that remains if we knew everything but the value of input X_i. This measure, first proposed by *Homma and Saltelli*(29), is referred to as the total effect variance of input X_i. The measure can be thought of as a cheap surrogate for the interaction variances, which some methods for assessing these measures (such as FAST, discussed later) are unable to calculate. Total effect variances are useful when the interactions are non negligible. Main effects approximately equal to total effects suggests little interaction between inputs.

It is usual to scale (2.23) and (2.24) by $Var[Y]$ in order to obtain main effect and total effect indices

$$\begin{aligned} S_i &= V_i/Var[Y], \qquad (2.25) \\ S_{Ti} &= V_{Ti}/Var[Y]. \qquad (2.26) \end{aligned}$$

We have the relation

$$0 \leq \sum S_i \leq 1 \leq \sum S_{Ti}, \qquad (2.27)$$

since any interaction between inputs i and j contributes to the total effect of both of these inputs, an interaction between i, j and k contributes to the total effect of all three inputs and so on.

For independent inputs (see *Chan, Saltelli and Tarantola*(9)), V_i is given by the integral

$$V_i = \int_{\mathcal{X}_i} \left\{ \int_{\mathcal{X}_{-i}} \eta(\mathbf{x})\, dG_{\mathbf{x}_{-i}|x_i}(\mathbf{x}_{-i}|x_i) \right\}^2 dG(x_i) - E(Y)^2, \qquad (2.28)$$

with interaction variances requiring similar integrals.

A variance based sensitivity analysis, in conjunction with plots of main effects and interactions allows us a good insight into the 'workings' of a complex computer model, even though we treat the model as a 'black box'. However, an analysis of this form is very computationally expensive, since all our integrals must be evaluated numerically and require many evaluations of $\eta(.)$. In the next section we discuss some of the classical computational methods for this analysis.

2.4 Classical Sensitivity Analysis

In practice the computation of main effects and interactions is a time consuming process. We see that the main effect (2.17) and first order interaction (2.18) both require the evaluation of a multidimensional integral (with $p-1$ and $p-2$) respective dimensions. Sensitivity indices require additional integrals to be calculated, although these are computationally cheap given the main effects and interactions.

As with calculations for uncertainty analysis, we can apply a brute force approach to evaluate these integrals. We find $z_i(x_i)$ by fixing x_i at various values over its range and evaluate $E(Y|x_i)$ by sampling from $\mathbf{X_{-i}}$ for each value of x_i. The precision with which we estimate $E(Y|x_i)$ for each x_i is determined by the size of our sample. Given $z_i(x_i)$ evaluated uniformly along X_i, we can evaluate V_i fairly cheaply using Simpson's rule.

However, if our vector of inputs is large, this is a computationally expensive process even if $\eta(.)$ is a cheap function to evaluate. Various authors have addressed this, by implementing more efficient procedures, however the methods which we now discuss are only able to calculate variance based measures.

2.4.1 Sobol' Indices

Sobol'(62) (63) provided a computationally cheaper solution for calculating sensitivity indices. We consider a subset of m of the inputs which we denote $\mathbf{x_1}$ and let the complementary set of $p-m$ inputs be denoted by $\mathbf{x_2}$, where $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2})$.

The main effect variance corresponding to subset $\mathbf{x_1}$ and total effect variance of $\mathbf{x_1}$ are then denoted as $V_{\mathbf{x_1}}$ and $V_{T\mathbf{x_1}} = V - V_{\mathbf{x_2}}$ respectively, and are the subset analogues of (2.23) and (2.24).

Sobol'(63) proves that for independent $U(0,1)$ inputs

$$V_{\mathbf{x_1}} = \int \eta(\mathbf{x})\eta(\mathbf{x_1}, \mathbf{x}'_2)\, d\mathbf{x}\, d\mathbf{x}'_2 \; - [E(Y)]^2 \tag{2.29}$$

$$V_{\mathbf{x_2}} = \int \eta(\mathbf{x})\eta(\mathbf{x}'_1, \mathbf{x_2})\, d\mathbf{x}\, d\mathbf{x}'_1 \; - [E(Y)]^2 \tag{2.30}$$

where in this notation $\mathbf{x}'_1$ denotes a different point in the sample space of our subset of m values from $\mathbf{x_1}$, similarly for $\mathbf{x}'_2$ and $\mathbf{x_2}$.

Now we consider 2 random points ξ and ξ' from the sample space of $\mathbf{X}$ and let $\xi = (\mathbf{x_1}, \mathbf{x_2})$ and $\xi' = (\mathbf{x_1}', \mathbf{x_2}')$. Each run of our sampling algorithm then requires 3 computations, $\eta(\mathbf{x_1}, \mathbf{x_2})$, $\eta(\mathbf{x_1}', \mathbf{x_2})$ and $\eta(\mathbf{x_1}, \mathbf{x_2}')$. *Sobol'* shows that

$$\frac{1}{N_s}\sum_{j=1}^{N_s}\eta(\xi_j) \quad \xrightarrow{p} \quad E(Y) \tag{2.31}$$

$$\frac{1}{N_s}\sum_{j=1}^{N_s}\eta(\xi)\eta(\mathbf{x_1}, \mathbf{x_2}') \quad \xrightarrow{p} \quad V_{\mathbf{x_1}} + [E(Y)]^2 \tag{2.32}$$

$$\frac{1}{N_s}\sum_{j=1}^{N_s}\eta^2(\xi_j) \quad \xrightarrow{p} \quad V + [E(Y)]^2 \tag{2.33}$$

$$\frac{1}{N_s}\sum_{j=1}^{N_s}\eta(\xi)\eta(\mathbf{x_1}', \mathbf{x_2}) \quad \xrightarrow{p} \quad V_{\mathbf{x_2}} + [E(Y)]^2 \tag{2.34}$$

where $\xrightarrow{p}$ denotes convergence in probability as $N_s \rightarrow \infty$

If we let $\eta(\mathbf{x_1}, \mathbf{x_2}') = \eta(x_i, \mathbf{x_{-i}})$ and $\eta(\mathbf{x_1}', \mathbf{x_2}) = \eta(\mathbf{x_{-i}}, x_i)$ we can compute the main and total effects of input x_i. We can see from the form of (2.31)-(2.34) the method easily generalizes to any set of independent inputs, regardless of their marginal distributions. We use N_s in this notation to denote the very large sample required using this procedure.

Jansen, Rossing and Daamen (32) also considered partitioning the parameters into two subsets and found their own 'Top Marginal Variances'. We give the expressions for the main and total effect variances below. We again have to evaluate the expectations via Monte-Carlo, requiring a total of N_J model evaluations.

$$\begin{aligned} V_{\mathbf{x_1}} &= V - 1/2\, E\{\eta(\mathbf{x_1}, \mathbf{x_2}) - \eta(\mathbf{x_1}, \mathbf{x_2'})\}^2, & (2.35) \\ V_{T\mathbf{x_1}} &= 1/2\, E\{\eta(\mathbf{x_1}, \mathbf{x_2}) - \eta(\mathbf{x_1'}, \mathbf{x_2})\}^2. & (2.36) \end{aligned}$$

The two methods are closely related and from a computational point of view *Chan, Saltelli and Tarantola*(9) state that they are equivalent. However *Chan et al.* (10) show that the method of *Jansen at al.* is more efficient for calculating total effect indices whilst the method of *Sobol'*(62)(63) is more efficient when calculating main effects.

The winding stairs sampling scheme considered in *Chan, Saltelli and Tarantola* is an efficient method of calculating both these sets of sensitivity indices. It can find main effect, total effect and indeed all interaction variances with a total of pN_{WS} model evaluations, where p is the number of model inputs. The method can also be easily extended to handle some dependence between inputs.

2.4.2 Fast indices

An alternative method of estimating variance based sensitivity indices is the Fourier Amplitude Sensitivity Test (FAST), which dates back to the early 70's. FAST was devised by *Cukier et al.* (13), who wanted to learn about the sensitivity of systems of coupled reactions described by a series of differential equations. However, a classical sensitivity analysis was impractical due to a restriction on the number of runs of their code.

The method as devised by *Cukier et al.*(13) and further developed by *Cukier and Schuler*(14) *Cukier, Levine and Schuler* (14) and *Koda et al.* (37), involves a mono-dimensional Fourier decomposition along a curve exploring the sample space, χ. The curve is described by the series of parametric equations

$$x_i = G_i(\sin \omega_i s) \qquad i = 1 \dots p, \tag{2.37}$$

where s is a scalar varying over $-\infty \leq s \leq \infty$ and for an appropriate set of transformation functions G_i, and integer frequencies ω_i.

As s varies the model parameters all change simultaneously along a curve that systematically explores χ. Each x_i oscillates periodically at the corresponding frequency ω_i and the output y shows different periodicities combined with the different frequencies ω_i. If parameter i has a strong influence on the output the oscillations of y at ω_i are of high amplitude, and this forms the basis of the sensitivity measure. The details of the method are complex and not given here, see *Saltelli, Tarantola and Chan*(61) and the references therein for a detailed description.

The literature on FAST is sparse until recent years, probably owing to its complexity and inability to compute interaction variances in its original form. This restriction meant that it was of limited use unless (2.11) simplified to

$$y = \eta(\mathbf{x}) \approx E\left(Y\right) + \sum_{i=1}^{p} z_i(x_i), \tag{2.38}$$

and for this form, simpler methods of sensitivity analysis are available.

However, in the recent past two developments have occurred that have greatly increased the literature on FAST: *Saltelli, Tarantola and Chan*(61) extended the methodology so that it can now compute total sensitivity indices; online software for sensitivity analysis using FAST has become available.

The original version of FAST required N_{FAST} runs of the model, where $N_{FAST} < N_S, N_J, N_{WS}$, which are roughly equivalent. However, the computation of total sensitivity indices increases the required sample size by a factor of p to pN_{FAST}, although this is still more efficient than the other methods discussed. The main disadvantages of FAST are its inability to calculate interaction variances, and an inability to handle some dependence between the inputs.

2.4.3 Two Stage Approaches

FAST and the methods of *Sobol'* and *Jansen et al.* require too many evaluations to make a full variance based sensitivity analysis practical when the number of parameters, p, is large. For problems with many inputs, a quantitative sensitivity analysis has to be conducted in two stages.

Firstly, a screening method, as described earlier in this chapter, ranks the inputs in order of importance and the less important variables are set at some nominal level. The number of inputs to eliminate in the first phase of the analysis seems to be somewhat arbitrary, and influenced more by computational resources than genuine subjective information about the number of active inputs. The second stage uses a variance based method to assess the sensitivity of the model output to the reduced set of parameters.

The results from the quantitative sensitivity analysis of the simplified model are only an approximation to the full model. Obviously some care needs to be used if we wish to use these results to infer properties of the full model. *Campolongo, Tarantola and Saltelli* (8) found that in their two stage analysis, the ranking method placed the inputs in a different order of importance to the variance based method (based on the reduced set of inputs).

2.4.4 Other methods

Alternative methods for global sensitivity analysis are available that require many fewer model evaluations than Sobol' and FAST, but they require strong assumptions about the form of $\eta(.)$. These can be classified as regression based measures, and are described in detail in *Helton and Davis*(26).

Regression based methods such as Standardized Regression Coefficients (SRC), Spearman Correlation Coefficients (SCC) and Partial Correlation Coefficients (PCC) have all been used to assess sensitivity. All these measures are based on the strong assumption that the computer code is well approximated by a linear model. They produce reliable results provided that the assumed linear model approximates the computer model well, with a model coefficient determination, R^2, close to 1. This is a more restrictive form of (2.38), where not only are interactions assumed to be small, but $\eta(.)$ is assumed to be well approximated by a linear function of the model inputs. The proportion, $1 - R^2$, of the variance is not explained by the regression, so the sensitivity analysis is only approximate.

In practice we will often find our complex model is not well approximated by a linear model. However, a parallel case exists for non-linear models, which can be assessed using similar methods based on the rank transform. These methods require a high R^2 on the rank scale. In using the rank transformation the restrictive assumptions under the linear model are relaxed somewhat however in order for R^2 to be large, these methods require the relationship between model inputs and output to be monotonic.

The final measure we review is the Correlation Ratio (CR), which has been used extensively to assess global sensitivity (see for example *McKay*(39) and

McKay, Morrison and Upton(41)).

The CR between Y and $\mathbf{X_1}$ is defined as

$$CR(Y, \mathbf{X_1}) = \frac{Var[E\{Y|\mathbf{X_1}\}]}{Var[Y]}, \tag{2.39}$$

where $\mathbf{X_1}$ denotes a subset of inputs.

The CR is closely related to Sobol' indices, for example S_i given in (2.25) is equal to $CR(Y, X_i)$ and S_{ij} given in (2.26) is equal to $CR(Y, \{X_i, X_j\})$. *McKay*(39) also shows that the CR is closely related to regression based methods, with regression based methods corresponding to a special form for the expectation $E\{Y|\mathbf{X_1}\}$. However, the CR requires many observations in order to evaluate each expectation, $E\{Y|\mathbf{X_1}\}$, and resultantly requires more observations of $\eta(.)$ than Sobol' indices or FAST.

2.5 Computationally Expensive Models

Complex computer models, which take minutes or even hours to compute a single model evaluation require a novel approach. *Sacks et al.*(57)(58)(59) and then *Currin et al.*(15) and *O'Hagan*(53) in a Bayesian context, noted that although the relationship between model inputs and the output(s) is complex, and $\eta(.)$ is regarded as a 'black box', the function may well be smooth. As such the output $\eta(\mathbf{x})$ and some adjacent output $\eta(\mathbf{x}')$, will be correlated. Therefore the evaluations, $\{\eta(\mathbf{x}_1), \ldots \eta(\mathbf{x}_n)\}$, convey some information about $\eta(.)$ as a whole, which a conventional analysis does not exploit.

The approach in these (and other) papers is to build a statistical model which emulates the computer model. The approach is similar in spirit to the parametric regression based methods discussed earlier, since it also uses a parametric approximation. However, the approach differs since it allows the local correlation structure to modify the parametric approximation, such that the statistical model smoothly interpolates the evaluations, $\{\eta(\mathbf{x}_1), \dots \eta(\mathbf{x}_n)\}$. With enough data observed the statistical model 'becomes' the computer model. Furthermore, the statistical model is of a simple enough form for us to be able to calculate measures of uncertainty and sensitivity. We are able make inferences about the computer model based upon these measures.

2.5.1 Overview of Remaining Chapters

In chapter 3 we develop a Bayesian model of the form described above, based upon the work of *O'Hagan*(53). We also examine the extensions developed by *O'Hagan*(54), *Haylock and O'Hagan*(23), *Oakley*(47) and *Oakley and O'Hagan*(49) (51) that allow us to calculate uncertainty and sensitivity measures.

In chapter 4 we consider the special kind of structural prior information that we have in Government financial models. We consider extensions to the methodology of chapter 3 for when we have non interacting groups of inputs, for the cases of both known and unknown groups of inputs.

In chapter 5 we develop the extensions that are required in order to calculate uncertainty and sensitivity measures for our models of chapter 4.

In chapter 6 we consider the elicitation of an autoregressive model, in order to model inflation rates in the future. This work is motivated by our application.

Finally in chapter 7 we examine the Ministry of Defence main building redevelopment project. We exploit the special structure of the model in order to calculate measures of uncertainty using the methodology of chapter 5.

Chapter 3

Uncertainty and Sensitivity Analysis for Expensive Functions

3.1 Introduction

In this chapter we discuss a hierarchical Bayesian stochastic process model that can be used to perform uncertainty and sensitivity analysis for computationally expensive computer models. The model formulation dates back to the late 70's when *Blight and Ott*(5) and *O'Hagan*(52) first applied Gaussian Process modelling to regression problems. The technique was modified and applied to computationally expensive, deterministic computer algorithms by *Sacks et al.*(59) and in a Bayesian setting by *Currin et al.*(15). Further developments by *O'Hagan*(54), *Haylock and O'Hagan*(23) and *Oakley and O'Hagan*(49) extended the Gaussian Process model to perform uncertainty analysis and *Oakley and O'Hagan*(51) developed methodology to allow probabilistic sensitivity analysis.

The contribution that these and other papers made to the methodology in this area is significant and relates heavily to the content of subsequent chapters of this thesis. We will take time in the remainder of this chapter to develop the Gaussian process model and the extensions that allow uncertainty and sensitivity analysis.

We develop the Bayesian model for approximating expensive functions in section 3.2. We go through the full prior to posterior analysis using the methodology developed by *O'Hagan*(52) in sections 3.2.1-3.2.2. We discuss alternative derivations in section 3.2.3. We develop a small section of original work that examines a flexible correlation function in 3.2.4, and we discuss the choice of design points in 3.2.5. We go on to examine how this approach can be used to calculate measures of uncertainty in section 3.3 and measures of sensitivity in section 3.4.

3.2 The Bayesian Model

Following on from the previous chapter we use $\eta(.)$ to represent our deterministic complex computer code, and $\eta(\mathbf{x})$ to denote the output at input configuration $\mathbf{x}$, where $\mathbf{x}$ is a p dimensional vector of inputs.

3.2.1 Specification of the Prior Distribution

We first consider the specification of the prior distribution, that represents our knowledge about the function, $\eta(.)$, before we make any observations of the function. Our beliefs about $\eta(.)$ will be expressed using a hierarchical stochastic process model. This requires us to formulate our beliefs about expectations, variances and covariances, along with some distributional assumptions. In all our prior specification contains four key elements and we address these in turn.

Our function, $\eta(.)$, is complex and whilst it is not transparent how the inputs, $\mathbf{x}$, affect the output, $\eta(\mathbf{x})$, it may be reasonable to suppose *a priori* that we can crudely approximate $\eta(.)$ at any point within the input space χ by some simple parametric form. We specify a regression model for this parametric form. We further we suppose that we can quantify the variability about our parametric fit.

Formally we express these statements as

$$\begin{aligned} E[\eta(\mathbf{x})|\boldsymbol{\beta},\sigma^2] &= \mathbf{h}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta}, \qquad (3.1)\\ Var[\eta(\mathbf{x})|\boldsymbol{\beta},\sigma^2] &= \sigma^2, \qquad (3.2)\end{aligned}$$

where $\mathbf{h}(\mathbf{x})^T$ is a vector of q regressor variables, $\boldsymbol{\beta}$ is a vector of parameters, and σ^2 quantifies the uncertainty surrounding the parametric approximation.

The second part of our model sets us apart from standard methods, and it is herein that the power of our approach is evident. Rather than treating our outputs as independent, we utilize the dependence between adjacent outputs and assume a structured form of covariance.

We define the covariance between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ as

$$Cov[\eta(\mathbf{x})\,,\,\eta(\mathbf{x}')\,|\boldsymbol{\beta},\sigma^2] = \sigma^2\mathbf{c}(\mathbf{x},\mathbf{x}'), \qquad (3.3)$$

where $\mathbf{c}(.,.)$ is a correlation function.

The correlation function describes the extent to which learning about $\eta(\mathbf{x})$ aids learning about $\eta(\mathbf{x}')$. We can choose from many different correlation functions, (see for example *Currin et al.*(15)). Our correlation function has the properties:

1. $\mathbf{c}(\mathbf{x},\mathbf{x}) = 1$;

2. $\mathbf{c}(\mathbf{x},\mathbf{x}')$ is stationary; it is a monotonically decreasing function of $|\,\mathbf{x}-\mathbf{x}'\,|$ for some distance measure $|\,.\,|$, and hence $\mathbf{c}(\mathbf{x},\mathbf{x}') = \mathbf{c}(\mathbf{x}+\boldsymbol{\delta},\mathbf{x}'+\boldsymbol{\delta})\,\forall\,\boldsymbol{\delta}$.

3. the correlation matrix of any finite set of m points, $\{\eta(\mathbf{x}_1),\ldots\eta(\mathbf{x}_m)\}$, is positive semi definite. This requires $\mathbf{c}(|\mathbf{x}-\mathbf{x}'|)$ to be the characteristic func-

tion of a random variable whose distribution function is symmetric about the origin (see *Feller*(16) for a fuller discussion).

Specifying a correlation function is not a simple task; we may have very little knowledge about how the function, $\eta(.)$, interacts with each of the model inputs, much less how smooth the output is relative to each of the model inputs. We solve this problem in part by defining the covariance in terms of additional hyperparameters.

The precise choice of correlation function, $\mathbf{c}(.,.)$, is ultimately down to experience and application driven, although the inferences we wish to derive about $\eta(.)$ may also influence this choice. In our application we expect the output, $\eta(.)$, to be smooth with respect to the inputs, and with no discontinuities. A correlation function with derivatives models these beliefs.

We adopt the exponential form

$$\mathrm{c}(\mathbf{x},\mathbf{x}') = \exp\{-(\mathbf{x}-\mathbf{x}')^{\mathbf{T}}\boldsymbol{\Omega}(\mathbf{x}-\mathbf{x}')\}, \tag{3.4}$$

for some positive semi-definite matrix, $\boldsymbol{\Omega}$, of (unknown) parameters. For now we take $\boldsymbol{\Omega}$ to be a diagonal matrix, although we consider the more general matrix form later in the chapter.

The third element of our model is the distributional assumptions. We take the joint distribution of any m outputs, $\{\eta(\mathbf{x}_1),\ldots\eta(\mathbf{x}_m)\}$, conditional on $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\Omega}$, to be m-dimensional multivariate normal. This holds for any m and defines the joint distribution for the entire function $\eta(.)$. This is known in the literature as a Gaussian Process, and denoted

$$\eta(.)\,|\,\boldsymbol{\beta},\sigma^2,\boldsymbol{\Omega} \sim GP(\,\mathbf{h}(.)^{\mathbf{T}}\boldsymbol{\beta},\sigma^2\mathrm{c}(.,.)\,), \tag{3.5}$$

where $\mathbf{h}(.)^{\mathbf{T}}\boldsymbol{\beta}$ is the prior mean function, and $\sigma^2\mathbf{c}(.,.)$ the prior covariance function.

The final element of our prior distribution requires us to consider our beliefs about the hyperparameters, $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\Omega}$. Assuming independence of $p(\boldsymbol{\Omega})$ and $p(\boldsymbol{\beta}, \sigma^2)$, our prior takes the form

$$p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Omega}) = p(\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\Omega}). \tag{3.6}$$

We adopt improper uniform priors on the (diagonal) elements of $\boldsymbol{\Omega}$ and the conjugate prior distribution

$$f(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-(d+q+2)/2} \exp\{-\frac{(\boldsymbol{\beta} - \mathbf{m})^{\mathbf{T}}\mathbf{V}^{-1}(\boldsymbol{\beta} - \mathbf{m}) + \mathbf{a}}{2\sigma^2}\}, \tag{3.7}$$

to represent our beliefs about $\boldsymbol{\beta}$ and σ^2.

Beliefs about $\boldsymbol{\beta}$ and σ^2 are elicited from expert knowledge. This is a difficult task requiring detailed questioning, and has recently been addressed by *Oakley*(48). Given the difficulties in effectively undertaking an elicitation, it is not unusual in practice to resort to the non informative prior $f(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$.

3.2.2 Prior to Posterior Analysis

Suppose that we are able to make n runs of the expensive computer code and we obtain the data vector, $\mathbf{y} = \{\eta(\mathbf{x}_1), \eta(\mathbf{x}_2), \ldots, \eta(\mathbf{x}_n)\}$, at inputs $\{\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n\}$. We are able to choose these design points in order to maximize the information on $\eta(.)$. We consider the choice of design in more detail later.

Given these data we wish to update our beliefs about the model parameters $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\Omega}$ and the function $\eta(.)$ itself. We begin with the distribution of $\mathbf{y}$, which conditional on $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\Omega}$ has a multivariate normal distribution:

$$\mathbf{y}\,|\,\boldsymbol{\beta},\,\sigma^2,\,\boldsymbol{\Omega} \sim N\,(\,\mathbf{H}\boldsymbol{\beta},\,\sigma^2\mathbf{A}\,), \tag{3.8}$$

where

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

$$\mathbf{A} = \begin{pmatrix} \mathbf{c}(\mathbf{x}_1,\mathbf{x}_1) & \mathbf{c}(\mathbf{x}_1,\mathbf{x}_2) & \ldots & \mathbf{c}(\mathbf{x}_1,\mathbf{x}_i) & \ldots & \mathbf{c}(\mathbf{x}_1,\mathbf{x}_n) \\ \mathbf{c}(\mathbf{x}_2,\mathbf{x}_1) & \mathbf{c}(\mathbf{x}_2,\mathbf{x}_2) & & & & \\ \vdots & & \ddots & \mathbf{c}(\mathbf{x}_j,\mathbf{x}_i) & & \\ & & \ldots & \mathbf{c}(\mathbf{x}_i,\mathbf{x}_i) & \ldots & \\ \mathbf{c}(\mathbf{x}_n,\mathbf{x}_1) & \ldots & & & & \mathbf{c}(\mathbf{x}_n,\mathbf{x}_n) \end{pmatrix}.$$

We now calculate the joint posterior of $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\Omega}$, which from (3.6) and (3.8) gives us

$$\begin{aligned} f(\boldsymbol{\beta},\sigma^2,\boldsymbol{\Omega}\,|\,\mathbf{y}) &\propto f(\,\mathbf{y}\,|\,\boldsymbol{\beta},\,\sigma^2,\boldsymbol{\Omega}) \times f(\boldsymbol{\beta},\sigma^2,\boldsymbol{\Omega}) & (3.9) \\ &\propto |\mathbf{A}|^{-1/2}(\sigma^2)^{-(d+n+q+2)/2}\exp\{-L/(2\sigma^2)\}, & (3.10) \end{aligned}$$

where

$$\begin{aligned} L &= (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\mathbf{V}^{*-1})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + a^*, & (3.11) \\ a^* &= a + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} + \mathbf{y}^T\mathbf{A}^{-1}\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{V}^{*-1}\hat{\boldsymbol{\beta}}, & (3.12) \\ \hat{\boldsymbol{\beta}} &= (\mathbf{V}^{-1} + \mathbf{H}^T\mathbf{A}^{-1}\mathbf{H})^{-1}(\mathbf{V}^{-1}\mathbf{m} + \mathbf{H}^T\mathbf{A}^{-1}\mathbf{y}), & (3.13) \\ \mathbf{V}^* &= (\mathbf{V}^{-1} + \mathbf{H}^T\mathbf{A}^{-1}\mathbf{H})^{-1}, & (3.14) \\ d^* &= d + n. & (3.15) \end{aligned}$$

We find it convenient to use the relation

$$f(\boldsymbol{\beta},\sigma^2,\boldsymbol{\Omega}|\,\mathbf{y}) = f(\boldsymbol{\beta}\,|\,\sigma^2,\,\boldsymbol{\Omega},\mathbf{y}) \times f(\sigma^2\,|\,\boldsymbol{\Omega},\mathbf{y}) \times f(\boldsymbol{\Omega}\,|\,\mathbf{y}), \tag{3.16}$$

where (see for example *Raiffa and Schlaifer*(56))

$$\boldsymbol{\beta}\,|\,\sigma^2,\,\boldsymbol{\Omega},\,\mathbf{y} \;\sim\; N(\hat{\boldsymbol{\beta}},\sigma^2\mathbf{V}^*), \tag{3.17}$$

$$\sigma^2\,|\,\boldsymbol{\Omega},\,\mathbf{y} \;\sim\; a^*\chi^{-2}_{d^*}, \tag{3.18}$$

and $f(\boldsymbol{\Omega}\,|\,\mathbf{y})$ has the non standard form

$$f(\boldsymbol{\Omega}\,|\,\mathbf{y}) \propto \hat{\sigma}^{-d^*}|\mathbf{A}|^{-1/2}|\mathbf{V}^*|^{-1/2}, \tag{3.19}$$

with

$$\hat{\sigma} = (\frac{a^*}{d^*-2})^{1/2}. \tag{3.20}$$

We now move onto the task of updating our beliefs about the function $\eta(.)$. We do so by exploiting properties of multivariate normal distributions. We note that any finite set of m outputs, $\{\eta(\mathbf{x}_1),\ldots\eta(\mathbf{x}_m)\}$, and our vector of n observations, $\mathbf{y}$, have, conditional on $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\Omega}$, a multivariate normal distribution.

It is simple to show that the distribution of $\{\eta(\mathbf{x}_1),\ldots\eta(\mathbf{x}_m)\}$, conditional on $\boldsymbol{\beta},\sigma^2$, $\boldsymbol{\Omega}$ and $\mathbf{y}$, is also multivariate normal. If, rather than considering a finite collection of random variables, we consider the joint distribution of the entire function, $\eta(.)$, then we have the result

$$\eta(.)\,|\,\mathbf{y},\boldsymbol{\beta},\sigma^2,\boldsymbol{\Omega} \sim GP(\mathbf{m}^*(.),\,\sigma^2\mathbf{c}^*(.,.)), \tag{3.21}$$

where

$$\begin{aligned} \mathbf{m}^*(\mathbf{x}) &= \mathbf{h}(\mathbf{x})^T\boldsymbol{\beta} + \mathbf{t}(\mathbf{x})^{\mathbf{T}}\mathbf{A}^{-1}(\mathbf{y} - \mathbf{H}\boldsymbol{\beta}), & (3.22)\\ \mathbf{t}(\mathbf{x}) &= [\mathbf{c}(\mathbf{x},\mathbf{x}_1)\ldots\mathbf{c}(\mathbf{x},\mathbf{x}_n)]^T, & (3.23)\\ \mathbf{c}^*(\mathbf{x},\mathbf{x}') &= \mathbf{c}(\mathbf{x},\mathbf{x}') - \mathbf{t}(\mathbf{x})^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{t}(\mathbf{x}'). & (3.24) \end{aligned}$$

We remove the conditioning on hyperparameters $\boldsymbol{\beta}$ and σ^2 in two stages. First we note that the product of (3.17) and (3.21) gives us the joint posterior of $\boldsymbol{\beta}$ and $\eta(.)$ conditional on hyperparameters σ^2 and $\boldsymbol{\Omega}$ and the data $\mathbf{y}$. We then integrate over $\boldsymbol{\beta}$ to obtain

$$\eta(.)\,|\,\mathbf{y}\,,\sigma^2\,,\boldsymbol{\Omega} \sim GP\,(\mathbf{m}^{**}(.),\, \sigma^2\mathbf{c}^{**}(.,.)), \qquad (3.25)$$

where

$$\mathbf{m}^{**}(\mathbf{x}) = \mathbf{h}(\mathbf{x})^{\mathbf{T}}\hat{\boldsymbol{\beta}} + \mathbf{t}(\mathbf{x})^{\mathbf{T}}\mathbf{A}^{-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}), \qquad (3.26)$$

and

$$\begin{aligned} \mathbf{c}^{**}(\mathbf{x},\mathbf{x}') = \mathbf{c}^*(\mathbf{x},\mathbf{x}') &+ (\mathbf{h}(\mathbf{x})^{\mathbf{T}} - \mathbf{t}(\mathbf{x})^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{H})\\ &\times (\mathbf{H}^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{H})^{-1}(\mathbf{h}(\mathbf{x}')^{\mathbf{T}} - \mathbf{t}(\mathbf{x}')^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{H})^{\mathbf{T}}. \quad (3.27) \end{aligned}$$

We now remove the conditioning on σ^2. We take the product of (3.25) and (3.18), which gives us the joint posterior of $\eta(.)$ and σ^2 conditional on $\boldsymbol{\Omega}$ and $\mathbf{y}$. Integrating over σ^2 leaves us with the posterior, $\eta(.)\,|\,\mathbf{y}\,,\boldsymbol{\Omega}$. The posterior distribution is a student process, with a description analogous to the Gaussian Process.

In particular for a given input configuration, $\mathbf{x}$, (see *Gosling*(21))

$$\frac{\eta(\mathbf{x}) - \mathbf{m}^{**}(\mathbf{x})}{\hat{\sigma}\sqrt{\frac{d^*-2}{d^*}\,\mathbf{c}^{**}(\mathbf{x},\mathbf{x})}}\;|\;\mathbf{y},\,\boldsymbol{\Omega} \sim t_{d^*}. \qquad (3.28)$$

However, we still have to remove the conditioning on $\mathbf{\Omega}$. Unfortunately, we are unable to remove the conditioning analytically, and MCMC is required, as used in *Neal*(45) and *Bayarri et al.*(4).

The solution proposed in *Kennedy and O'Hagan*(36), is to derive plausible estimates for the components of $\mathbf{\Omega}$, and act as if these were fixed. For inference about $\eta(.)$, we use the conditional posterior given the data, $\mathbf{y}$, and the estimated value of $\mathbf{\Omega}$. This is no longer a fully Bayesian analysis, however *Kennedy and O'Hagan* claim that it is only a 'second order' effect that is neglected, and such an analysis captures the major part of the uncertainty. By adopting this simplification, it is possible to calculate uncertainty and sensitivity measures analytically.

We adopt *Kennedy and O'Hagan's* methodology, and derive plausible estimates for $\mathbf{\Omega}$, proceeding as if these were known. We estimate the elements of $\mathbf{\Omega}$ from their joint posterior mode. We have to numerically maximize (3.19), although it is numerically better to work with the logarithm

$$\log f(\mathbf{\Omega} \,|\, \mathbf{y}) \propto -d^* \log \hat{\sigma} - 1/2 \log |\mathbf{A}| - 1/2 \log |\mathbf{V}^*|. \tag{3.29}$$

Inference for $\eta(.)$ is based upon (3.28), where $\mathbf{\Omega}$ is replaced by the posterior mode.

3.2.3 Alternative Derivations

In deriving the posterior (3.28), we followed the approach first proposed by *O'Hagan*(52) in the context of regression (although not for deterministic models). *O'Hagan's* approach is unique in that it takes into account the uncertainty in $\boldsymbol{\beta}$ and σ^2, although the 'second order' uncertainty in $\mathbf{\Omega}$ is ignored. However, other authors have proposed a similar method in both frequentist and Bayesian settings.

Sacks et al.(59) were the first to tackle computationally expensive computer codes in this context. They modelled the deterministic computer code output by (3.30), which they interpret as *Response = Linear Model + Departures*:

$$Y(\mathbf{x}) = \sum_{j=1}^{k} \beta_j f_j(x) + Z(\mathbf{x}). \tag{3.30}$$

They treated $Z(\mathbf{x})$ as a systematic departure, and it is modelled as a realization of a stochastic process. The covariance structure of $Z(\mathbf{x})$ relates to the smoothness of the response.

In order to construct an estimator for future values of the complex computer code, *Sacks et al.*(59) use the criterion of best linear predictor of $Y(\mathbf{x})$. Letting $\mathbf{y}^T = [Y(\mathbf{x}_1), Y(\mathbf{x}_2), \ldots, Y(\mathbf{x}_n)]$ denote the vector of responses from design points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, if $\mathbf{c}^\mathbf{T}\mathbf{y}$ is a linear predictor of $Y(\mathbf{x})$, then its mean squared error is given by

$$E[\mathbf{c}^\mathbf{T}\mathbf{Y} - Y(x)]^2, \tag{3.31}$$

where the expectation is with respect to the random process $\mathbf{Y}$.

The best linear predictor of $Y(\mathbf{x})$ is found by minimizing (3.31). However, *Sacks et al.* add an additional unbiasedness constraint to ensure the predictor interpolates the computer code output at the design points.

The best linear unbiased predictor, as derived explicitly in *Sacks et al.*(59), is identical to our posterior mean (3.26). However, the coefficients, $\hat{\boldsymbol{\beta}}$, are the generalized least squares estimates, which will not generally be the same as our Bayesian estimates, (3.13) (unless we adopt the non informative prior $f(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$). In addition, *Sacks et al.* only consider a predictor of $Y(\mathbf{x})$, and don't consider the uncertainty about their predictor. However, if they did consider the uncertainty about their predictor, $Y(\mathbf{x})$, their inference would be based on a

Gaussian Process.

Currin et al.(15), were the first to use a Bayesian approach. They treat $\boldsymbol{\beta}$ and σ^2 as known constants, which they later estimate using empirical Bayes methods, rather than explicitly modelling the uncertainty in these parameters. As a result *Currin et al.* have a posterior Gaussian Process (identical to *Sacks et al.'s* predictor, albeit with a different interpretation).

3.2.4 A Flexible Correlation Function

We now consider a more general form of correlation function. The correlation between outputs $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ is given by

$$\mathbf{c}(\mathbf{x},\mathbf{x}') = \exp\{-(\mathbf{x}-\mathbf{x}')^{\mathrm{T}}\boldsymbol{\Omega}(\mathbf{x}-\mathbf{x}')\}, \tag{3.32}$$

for some positive semi-definite matrix, $\boldsymbol{\Omega}$, of (unknown) parameters. This is in fact an identical structure to (3.4) however we no longer constrain the parameter matrix $\boldsymbol{\Omega}$ to be diagonal. This non-diagonal form has been suggested before by *Kennedy and O'Hagan*(36), although to our knowledge it has never been applied.

We begin with our prior specification. We once more assume *a priori* independence of ($\boldsymbol{\beta}$, σ^2) and $\boldsymbol{\Omega}$;

$$p(\boldsymbol{\beta},\sigma^2,\boldsymbol{\Omega}) = p(\boldsymbol{\beta},\sigma^2) \times p(\boldsymbol{\Omega}). \tag{3.33}$$

However, we now represent our beliefs about $\boldsymbol{\Omega}$ with the Inverse Wishart prior

$$p(\boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{-(c+p+1)/2}\exp\{-(tr\boldsymbol{\Omega}^{-1}\mathrm{B})/2\}, \tag{3.34}$$

for $c > p$ and symmetric positive definite matrix $\mathbf{B}$. However, eliciting beliefs about $\mathbf{\Omega}$ will be impractical, so the limiting case (with c = 0 and $\mathbf{B} = \mathbf{0}$), of $p(\mathbf{\Omega}) \propto |\mathbf{\Omega}|^{-(p+1)/2}$ is appropriate.

Inference for $\eta(.)$ when using this correlation function follows the same prior to posterior analysis we developed in sections 3.2.1-3.2.2. Our posterior distribution, (3.28), is conditional on the data $\mathbf{y}$ and $\mathbf{\Omega}$. We estimate $\mathbf{\Omega}$ by maximizing

$$f(\mathbf{\Omega} \,|\, \mathbf{y}) \propto |\mathbf{\Omega}|^{-(p+1)/2} \hat{\sigma}^{-d^*} |\mathbf{A}|^{-1/2} |\mathbf{V}^*|^{-1/2}. \tag{3.35}$$

or alternatively maximizing the logarithm

$$\log f(\mathbf{\Omega} \,|\, \mathbf{y}) \propto -(p+1)/2 \log |\mathbf{\Omega}| - d^* \log \hat{\sigma} - 1/2 \log |\mathbf{A}| - 1/2 \log |\mathbf{V}^*|. \tag{3.36}$$

Computation

In both the diagonal and non-diagonal forms for $\mathbf{\Omega}$, we have to maximize the log posteriors, (3.29) and (3.36) respectively, using numerical methods. We opt for the downhill simplex algorithm of *Nelder and Mead*(46). However, in the non-diagonal form we have introduced an additional $p(p-1)/2$ dimensions to maximize over, at significant computational expense. We only want to use this more complex form if the increased flexibility it offers is worth the additional computational expense.

Our numerical work has shown that the non diagonal form can lead to significant improvements in prediction in some problems. The potential improvements depend on the form of the function, $\eta(.)$, and the various interactions between the model inputs. When $\eta(.)$ is an additive function of the inputs, or the main effects (which we defined in chapter 2) are large compared with any interactions, our numerical work has found that the correlation can be modelled well by a diagonal form. In this scenario, when using the non-diagonal form we usually find that the off-diagonal elements of $\mathbf{\Omega}$ are very small compared with the diagonal elements. When $\eta(.)$ contains interactions between the inputs that are not small compared with the main effects, our numerical work has shown that there may be considerable improvement when we use the non-diagonal form.

We demonstrate with the 6 dimensional function (3.37), where the true values of the inputs all have $N(0,1)$ distributions. The example contains non-negligible first and second order interactions, and the output ranges from around -10 to $+10$. The example is structured such that we have no interactions between the 2 groups x_1, x_2, x_3 and x_4, x_5, x_6.

$$\begin{aligned} y &= 1.5x_1 + 0.95x_2 - 0.25x_3 + 1.3x_4 + 1.3x_5 - 0.3x_6 \\ &+ \cos(0.8x_1 + 0.75x_2 + 0.65x_3) + \cos(0.7x_1 + 0.2x_2 - 0.9x_3) \\ &+ \cos(0.7x_4 + 0.8x_5 + 0.55x_6) + \cos(0.6x_4 + 0.5x_5 - 0.85x_6) \\ &+ \sin(0.8x_1 + 0.75x_2 + 0.65x_3) + \sin(0.7x_1 + 0.2x_2 - 0.9x_3) \\ &+ \sin(0.7x_4 + 0.8x_5 + 0.55x_6) + \sin(0.6x_4 + 0.5x_5 - 0.85x_6). \end{aligned} \tag{3.37}$$

We observe the function at 50 design points. We take $\mathbf{h}(\mathbf{x}) = (1, \mathbf{x})$ and we use the correlation function (3.32). We maximize (3.36) in order to estimate the matrix $\mathbf{\Omega}$. For this example, we find that the log posterior is maximized at

$$\hat{\mathbf{\Omega}} = \begin{pmatrix} 0.091 & 0.061 & -0.002 & 0.000 & 0.000 & 0.000 \\ 0.061 & 0.051 & 0.028 & 0.000 & 0.000 & 0.000 \\ -0.002 & 0.028 & 0.096 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.065 & 0.056 & -0.01 \\ 0.000 & 0.000 & 0.000 & 0.056 & 0.056 & 0.014 \\ 0.000 & 0.000 & 0.000 & -0.01 & 0.014 & 0.078 \end{pmatrix}.$$

The first thing we note is the block diagonal structure of $\hat{\mathbf{\Omega}}$, which mirrors the structure of the example. The 2 groups x_1, x_2, x_3 and x_4, x_5, x_6 are non interacting and the corresponding elements of $\hat{\mathbf{\Omega}}$ are zero. In general we would not expect to find such an extreme result, but we would expect these elements to be close to zero, since the corresponding inputs are non-interacting. If we look within the two blocks, we note the off diagonal elements are not small compared with the diagonal elements. Again this mirrors the structure of the example since we have interactions within x_1, x_2, x_3 and x_4, x_5, x_6.

A formal comparison of the diagonal and non-diagonal matrix forms is possible using the criterion of Expected Root Mean Squared Error (ERMSE). We fit Gaussian Process models using both correlation functions ((3.4) and (3.32) respectively), and predict the output at a further 200 randomly selected points. The diagonal form yields an ERMSE of 0.667166, whilst the non-diagonal form has ERMSE of 0.242424; less than half the error.

The non-diagonal form for $\mathbf{\Omega}$ contains an additional 15 parameters (21 parameters in (3.32) and 6 parameters in (3.4)). Since numerical maximization routines are an $O(m^2)$ operation (for an m dimensional maximization), the improvements in prediction need to be balanced against the computational burden in maximizing (3.36). For a cheap function such as (3.37), the most efficient option would be to improve the accuracy of predictions (and reduce the variance of the predictions) by making more observations of $\eta(.)$ and use the diagonal form (3.4). However, for a computationally expensive function, using (3.32) with a smaller number of observations may be a more efficient use of resources.

Transformations

It may be possible to model the correlation using the function (3.32) with no greater computational burden than when using the diagonal form (3.4). We can write the parameter matrix $\mathbf{\Omega}$ as

$$\mathbf{\Omega} = \mathbf{C}^{\mathbf{T}}\mathbf{\Omega}^{*}\mathbf{C}, \tag{3.38}$$

where $\mathbf{\Omega}^{*}$ is an $r \times r$ diagonal matrix, and $\mathbf{C}$ a $r \times p$ transformation matrix, with $r \leq p$. The diagonal elements of $\mathbf{\Omega}^{*}$ are the r non zero eigenvalues of $\mathbf{\Omega}$ and the rows of $\mathbf{C}$ are the corresponding eigenvectors.

Thus, we can write the correlation between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ as

$$\begin{aligned} c(\mathbf{x}, \mathbf{x}') &= \exp\{(\mathbf{x} - \mathbf{x}')^{\mathbf{T}}\mathbf{\Omega}(\mathbf{x} - \mathbf{x}')\} \\ &= \exp\{(\mathbf{C}\mathbf{x} - \mathbf{C}\mathbf{x}')^{\mathbf{T}}\mathbf{\Omega}^{*}(\mathbf{C}\mathbf{x} - \mathbf{C}\mathbf{x}')\}. \end{aligned} \tag{3.39}$$

From (3.39) we can see that a non diagonal matrix of parameters corresponds to a diagonal matrix on a linearly transformed scale, $\mathbf{z} = \mathbf{C}\mathbf{x}$. Moreover the transformed scale is of dimension $r \leq p$.

An efficient method for estimating the $p(p+1)/2$ components of $\mathbf{\Omega}$ would be to specify $\mathbf{C}$ such that $\mathbf{\Omega}^{*}$ is approx diagonal. We would only need to maximize (3.29) over r dimensions in order to estimate the $p(p+1)/2$ components of $\mathbf{\Omega}$. However, since $\eta(.)$ is an unknown function, in practice it is not obvious how to choose $\mathbf{C}$. Our attempts at specifying a transformation have been unsuccessful.

3.2.5 Design

We now consider our choice of design points that we wish to observe the function at. We have a fixed number, n, of design points, and we wish to select these in order to maximize the information, in some sense, about $\eta(.)$ at the infinite collection of unobserved points. One approach is to define some criterion which describes what a good design is, and then find the design which best fulfils this.

Criteria for selecting good designs in order to maximize the information about $\eta(.)$ have been proposed by various authors. The various criteria developed have sought to exploit the smoothness of the output, $\eta(.)$, in order to improve on the Latin Hypercube methodology that we discussed in chapter 2, and which takes no account of the smoothness of the output.

For a fixed number of model evaluations, n, and a specified correlation function $\mathbf{c}(.,.)$, *Sacks et al.*(59) considered 3 different criteria, although they only implemented the first of these. The first criterion was *Integrated Mean Square Error* (IMSE) for their estimator $\hat{\eta}(\mathbf{x})$, which chooses the design, D, to minimize

$$\begin{aligned} & \int_{\chi} MSE[\hat{\eta}(\mathbf{x})]\phi(\mathbf{x})d\mathbf{x} \\ = & \int_{\chi} E[\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})]^2\phi(\mathbf{x})d\mathbf{x}, \end{aligned} \tag{3.40}$$

where the expectation is taken with respect to $\eta(\mathbf{x})$, and for a given weight function $\phi(\mathbf{x})$. A general weight function causes no difficulties, but the authors take $\phi(\mathbf{x})$ to be uniform over the whole of χ in their applications.

The second criterion considered by *Sacks et al.* is *Maximum Mean Squared Error* (MMSE). This is a minimax criterion, which seeks to minimize the maximum prediction error.

The design is chosen to minimize

$$\begin{aligned} & max_{\chi} MSE[\hat{\eta}(\mathbf{x})], \\ = & max_{\chi} E[\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})]^2. \end{aligned} \tag{3.41}$$

However, the criterion is not implemented in their examples due to the computational expense. The method has been compared with IMSE for discrete regions by *Sacks and Schiller*(57), but for continuous regions the implementation is far more difficult and as a result computationally expensive.

The final criterion they consider is choosing a design, D, to minimize the *posterior entropy*, which is a measure, in some sense, of the 'amount of uncertainty'. This criterion is studied in more detail, and applied to examples in *Currin et al.*(15). We define the entropy of a continuous random variable Z as:

$$H(Z) = E[-\log p_Z(Z)], \tag{3.42}$$

where $p_Z(Z)$ is the density of Z.

In our case, Z represents the infinite collection of untried outputs, $\eta(.)$, given the output at the n evaluated outputs $\mathbf{y}$. In the case considered by *Currin et al.*, the regression parameters, $\boldsymbol{\beta}$, are regarded as fixed (and estimated in the analysis by empirical bayes), and the posterior entropy is minimized when $|\mathbf{A}|$ is maximized.

Choice of design in a computer models context has also been considered by *Haylock*(22). He considered a loss function of the form

$$L\{D,\mathbf{y},\mathbf{X},\eta(\mathbf{X})\} = \int_{\mathcal{X}} \{\mathrm{m}^{**}(\mathbf{X}) - \eta(\mathbf{X})\}^2 d\mathrm{G}(\mathbf{x}), \tag{3.43}$$

where $\mathbf{y}$, $\mathbf{X}$ and $\eta(\mathbf{X})$ are unknown. *Haylock* takes the expectation over the unknown parameters of his loss function, and finds the loss as a function of the design alone can be written as

$$L\{D\} \propto \int_{\mathcal{X}} \mathbf{c}^{**}(\mathbf{x},\mathbf{x})\, d\mathrm{G}(\mathbf{x}). \tag{3.44}$$

A problem common to all four of the above criteria is that the parameters of the correlation function, $\mathbf{c}(.,.)$, are unknown. *Currin et al.*(15) state that in their experience, the correlation function itself is often unknown prior to analysis.

Unless the parameters of $\mathbf{c}(.,.)$ are known or at least estimated using expert knowledge (see *Oakley*(48)), then an optimality is difficult to achieve. The literature has instead focussed on design robustness. *Sacks et al.*(59) consider the robustness of designs to mis-specification of the parameters of $\mathbf{c}(.,.)$, and provide a good review of empirical work in the area. However, the various authors they cite have differing conclusions, with robust solutions dependent on both the form of $\mathbf{c}(.,.)$, and the magnitude of the (unknown) parameters.

One resolution that has been proposed in the literature is a two phased approach to the selection of design points (*Currin et al.*(15), *Sacks et al.*(58), (59)). An initial design is chosen subject to some criterion, before the correlation function, $\mathbf{c}(.,.)$, is chosen and the parameters estimated. The remaining design points are then selected, again using some design criterion. A sequential design of this form cannot be optimal under any of the above criteria however it seems to provide a reasonable solution. In addition, this approach results in a lower computational burden (a one at a time search, rather than a global search) in the numerical searches that are required in order to select the design points.

In the next chapter we go on to consider decompositions of $\eta(.)$ into lower dimensional functions. We could in principle extend any of the above criteria for this situation, provided that the decomposition of $\eta(.)$ is known. For an unknown decomposition we could consider a 2 stage approach; initially a small Latin Hypercube design to identify the decomposition of $\eta(.)$, followed by a design exploiting the known structure of $\eta(.)$. However, these criteria are not considered further. We use Latin Hypercube Designs in the remainder of the thesis.

3.3 Uncertainty Analysis

In chapter 2 we stated that three measures of interest for expressing our uncertainty about Y are the expectation, variance, and distribution function. These summaries require us to calculate the integrals

$$\begin{aligned} E(Y|\eta(.)) &= \int_{\mathcal{X}} \eta(\mathbf{x})\, dG(\mathbf{x}), && (3.45) \\ E(Y^2|\eta(.)) &= \int_{\mathcal{X}} \eta(\mathbf{x})^2\, dG(\mathbf{x}), && (3.46) \\ F_Y(s)|\eta(.) &= \int_{\mathcal{X}} I\{\eta(\mathbf{x}) \leq s\}\, dG(\mathbf{x}), && (3.47) \end{aligned}$$

where we now condition explicitly on the functional relationship between $\mathbf{x}$ and $\eta(\mathbf{x})$.

In chapter 2 we stated that for a complex function these summaries will not be available analytically. In principle we can find (3.45)-(3.47) by integrating numerically. However, the function $\eta(.)$ is computationally expensive, so numerical methods are impractical.

Using our Bayesian method we are able to make more effective use of the data in order to estimate these summaries. In the previous section we found the posterior distribution of $\eta(.)$ given data $\mathbf{y}$ and the estimated values of $\mathbf{\Omega}$. If we knew $\eta(\mathbf{x})$ for every $\mathbf{x}$ we could calculate our summaries exactly. However, we only have the posterior distribution of $\eta(\mathbf{x})$ for any $\mathbf{x}$. The uncertainty about $\eta(.)$ means we also have uncertainty about our summaries. Therefore, the summaries we wish to calculate are random variables (see for example *Haylock and O'Hagan*(23), *Oakley and O'Hagan*(51)). In this section we calculate the posterior distribution of $E(Y)$, and posterior summaries of $Var(Y)$ and $F_Y(y)$, since their posterior distributions have no closed form. In the remainder of this chapter we estimate $\mathbf{\Omega}$ by $\hat{\mathbf{\Omega}}$, and treat it as fixed and known.

Expectation

We begin with inference about $K_1 = E(Y|\eta(.)) = \int_{\mathcal{X}} \eta(\mathbf{x})dG(\mathbf{x})$, which was first tackled by *Haylock and O'Hagan*(23). They showed that

$$K_1|\sigma^2, \mathbf{\Omega}, \mathbf{y} \sim N(\hat{k}, \sigma^2 W), \tag{3.48}$$

where

$$E^*\left[K_1\,|\sigma^2, \mathbf{\Omega},\, \mathbf{y}\right] = \hat{k} = \int_{\mathcal{X}} \mathbf{m}^{**}(\mathbf{x})\; dG(\mathbf{x}) = \mathbf{R}\hat{\boldsymbol{\beta}} + \mathbf{T}\mathbf{A}^{-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}), \tag{3.49}$$

and

$$\begin{aligned} Var^*\left[K_1\,|\sigma^2, \mathbf{\Omega},\, \mathbf{y}\right] &= \sigma^2 W = \sigma^2 \int_{\mathcal{X}}\int_{\mathcal{X}} \mathbf{c}^{**}(\mathbf{x}, \mathbf{x}')\; dG(\mathbf{x})\; dG(\mathbf{x}'), \\ &= \sigma^2\{U - \mathbf{T}\mathbf{A}^{-1}\mathbf{T}^{\mathbf{T}} + (\mathbf{R} - \mathbf{T}\mathbf{A}^{-1}\mathbf{H})(\mathbf{H}^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{H})^{-1} \\ &\qquad \times (\mathbf{R} - \mathbf{T}\mathbf{A}^{-1}\mathbf{H})^{\mathbf{T}}\}. \end{aligned} \tag{3.50}$$

We use the notation E^* and Var^* to denote the expectation and variance with respect to the posterior distribution of $\eta(.)$.

The quantities $\mathbf{R}, \mathbf{T}$ and U are themselves expressed in terms of integrals

$$\mathbf{R} = \int_{\mathcal{X}} \mathbf{h}(\mathbf{x})^T dG(\mathbf{x}), \tag{3.51}$$

$$\mathbf{T} = \int_{\mathcal{X}} \mathbf{t}(\mathbf{x})^T dG(\mathbf{x}), \tag{3.52}$$

$$U = \int_{\mathcal{X}}\int_{\mathcal{X}} \mathbf{c}(\mathbf{x}, \mathbf{x}')\; dG(\mathbf{x})\; dG(\mathbf{x}'). \tag{3.53}$$

The conditioning on σ^2 is removed by taking the product of (3.48) and (3.18) and integrating over σ^2.

The expectation, K_1, has a t-distribution:

$$\frac{K_1 - \hat{k}}{\hat{\sigma}\sqrt{\frac{d^*-2}{d^*}W}} \mid \boldsymbol{\Omega}, \mathbf{y} \sim t_{d^*}. \tag{3.54}$$

We now have a point estimate for $E(Y|\eta(.))$ in the form of $\hat{k}$, but we also have a measure of our uncertainty about this estimate, as measured by (3.54).

Variance

Haylock and O'Hagan(23) also considered the variance of $\eta(.)$ however their calculation was corrected by *Oakley and O'Hagan*(51). For this we require the posterior distribution of $K_2 = E(Y^2|\eta(.)) = \int_{\mathcal{X}} \eta(\mathbf{x})^2 dG(\mathbf{x})$, which is intractable. *Haylock and O'Hagan* calculated posterior moments of K_2. We just show the expectation calculation here (see *Haylock and O'Hagan* for the variance calculation)

$$\begin{aligned} E^*[K_2|\sigma^2, \boldsymbol{\Omega}, \mathbf{y},] &= E\left[\int_{\mathcal{X}} \eta^2(\mathbf{x}) dG(\mathbf{x}) | \sigma^2, \boldsymbol{\Omega}, \mathbf{y}\right], \\ &= E\left[\int_{\mathcal{X}} \eta^2(\mathbf{x}) | \sigma^2, \boldsymbol{\Omega}, \mathbf{y}\, dG(\mathbf{x})\right], \end{aligned} \tag{3.55}$$

and

$$E^*\left[\int_{\mathcal{X}} \eta^2(\mathbf{x}) | \sigma^2, \boldsymbol{\Omega}, \mathbf{y}\right] = \int_{\mathcal{X}} \mathbf{m}^{**}(\mathbf{x})^2 + \sigma^2 \mathbf{c}^{**}(\mathbf{x}, \mathbf{x})\, dG(\mathbf{x}). \tag{3.56}$$

Substituting the expressions for $\mathbf{m}^{**}(\mathbf{x})$ and $\mathbf{c}^{**}(\mathbf{x}, \mathbf{x})$ as given in (3.26) and (3.27) respectively into (3.56) we can expand this expression as

$$\begin{aligned} E^*[K_2|\sigma^2, \boldsymbol{\Omega}, \mathbf{y}] &= tr(\hat{\beta}^{\mathbf{T}} \mathbf{Q} \hat{\beta}) + tr((\mathbf{y} - \mathbf{H}\hat{\beta})^{\mathbf{T}} \mathbf{A}^{-1} \mathbf{P} \mathbf{A}^{-1} (\mathbf{y} - \mathbf{H}\hat{\beta})) \\ &+ 2tr(\hat{\beta} \mathbf{T}^{\mathbf{T}} \mathbf{R} \mathbf{A}^{-1} (\mathbf{y} - \mathbf{H}\hat{\beta}) + \sigma^2[1 - tr(\mathbf{A}^{-1}\mathbf{P}) \\ &+ tr((\mathbf{H}^{\mathbf{T}} \mathbf{A}^{-1} \mathbf{H})^{-1} \mathbf{Q}) - 2tr((\mathbf{H}^{\mathbf{T}} \mathbf{A}^{-1} \mathbf{H})^{-1} \mathbf{S} \mathbf{A}^{-1} \mathbf{H}) \\ &+ tr((\mathbf{H}^{\mathbf{T}} \mathbf{A}^{-1} \mathbf{H})^{-1} \mathbf{H}^{\mathbf{T}} \mathbf{A}^{-1} \mathbf{P} \mathbf{A}^{-1} \mathbf{H})]. \end{aligned} \tag{3.57}$$

where once again the expressions for $\mathbf{P}, \mathbf{Q}$ and $\mathbf{S}$ are integrals that we need to evaluate:

$$\mathbf{P} = \int_{\mathcal{X}} \mathbf{t}(\mathbf{x})\mathbf{t}(\mathbf{x})^{\mathbf{T}}\, d\,G(\mathbf{x}); \tag{3.58}$$

$$\mathbf{Q} = \int_{\mathcal{X}} \mathbf{h}(\mathbf{x})\mathbf{h}(\mathbf{x})^{\mathbf{T}}\, d\,G(\mathbf{x}); \tag{3.59}$$

$$\mathbf{S} = \int_{\mathcal{X}} \mathbf{t}(\mathbf{x})\mathbf{h}(\mathbf{x})^{\mathbf{T}}\, d\,G(\mathbf{x}). \tag{3.60}$$

By taking the product of $E^{*}[K_2|\sigma^2, \mathbf{\Omega}, \mathbf{y}]$ and (3.18) and integrating over σ^2 we can find $E^{*}[K_2|\mathbf{\Omega}, \mathbf{y}]$. This expression is identical to (3.57), but with σ^2 replaced by $\hat{\sigma}^2$. Finally, we calculate the expectation of $Var(Y)$ using

$$\begin{aligned} E^{*}[Var\{Y\}|\mathbf{y}] &= E^{*}[(K_2 - K_1^2)|\mathbf{\Omega}, \mathbf{y}] \\ &= E^{*}[K_2|\mathbf{\Omega}, \mathbf{y}] - (Var^{*}[K_1|\mathbf{\Omega}, \mathbf{y}] + E^{*}[K_1|\mathbf{\Omega}, \mathbf{y}]^2). \end{aligned} \tag{3.61}$$

Haylock and O'Hagan stated that the posterior distribution of $E(Y)$ and the expectation and variance of $[Var(Y)]$ could be found analytically for common choices of $\mathbf{h}(.)$, $\mathbf{c}(.,.)$ and $G(\mathbf{x})$. *Haylock*(22) gave explicit calculations for the case where $\mathbf{h}(.)$ takes the form of polynomials of the elements of $\mathbf{x}$, $\mathbf{c}(.,.)$ has the diagonal form (3.4) and $G(\mathbf{x})$ is product normal. Given the same $\mathbf{h}(.)$ and $\mathbf{c}(.,.)$ as *Haylock* assumed we can consider much more general forms of $G(\mathbf{x})$. In our example in chapter 7 we adopt the same forms for $\mathbf{h}(.)$ and $\mathbf{c}(.,.)$, but $G(\mathbf{x})$ contains uniform (discrete and continuous), triangular, and multivariate normal distributions. We are able to to perform these calculations analytically.

Oakley and O'Hagan(49) developed a simulation based method that could be used to calculate the expectation and variance (amongst other summaries) of Y, for when the integrals $\mathbf{R}, \mathbf{T}, \mathbf{S}, \mathbf{P}$ and $\mathbf{Q}$ were not tractable.

Distribution Function

The final summary we require is the distribution function of Y, $F_{Y|\eta(.)}(s)$. The posterior distribution requires us to calculate the the integral

$$F_{Y|\eta(.)}(s) = \int_{\mathcal{X}} I\{\eta(\mathbf{x}) \leq s\}\, dG(\mathbf{x}). \tag{3.62}$$

However, as we can see, the calculation for the distribution function involves the indicator function, $I\{.\}$, and as a result the posterior distribution of $F_Y|\eta(.)(s)$ is intractable.

Oakley and O'Hagan(49) derived the first two posterior moments of $F_{Y|\eta(.)}(s)$.

$$\begin{aligned} E^*\{F_{Y|\eta(.)}(s)|\mathbf{\Omega},\mathbf{y}\} &= \int_{\mathcal{X}} E^*[I\{\eta(\mathbf{x}) \leq s\}\,|\,\mathbf{\Omega},\mathbf{y}]\, dG(\mathbf{x}) \\ &= \int_{\mathcal{X}} P[\{\eta(\mathbf{x}) \leq s\}\,|\,\mathbf{\Omega},\mathbf{y}]\, dG(\mathbf{x}) \\ &= \int_{\mathcal{X}} P[\{\frac{\eta(\mathbf{x}) - \mathbf{m}^{**}(\mathbf{x})}{\hat{\sigma}\sqrt{\frac{d^*-2}{d^*}\mathbf{c}^{**}(\mathbf{x},\mathbf{x})}} \leq \frac{s - \mathbf{m}^{**}(\mathbf{x})}{\hat{\sigma}\sqrt{\frac{d^*-2}{d^*}\mathbf{c}^{**}(\mathbf{x},\mathbf{x})}}\}\,|\,\mathbf{\Omega},\mathbf{y}]\, dG(\mathbf{x}) \\ &= \int_{\mathcal{X}} \int_{-\infty}^{\frac{s-\mathbf{m}^{**}(\mathbf{x})}{\hat{\sigma}\sqrt{\frac{d^*-2}{d^*}\mathbf{c}^{**}(\mathbf{x},\mathbf{x})}}} f_{T_{d^*}}\, dt\, dG(\mathbf{x}), \end{aligned} \tag{3.63}$$

where $f_{T_{d^*}}$ is the density of a t-distribution with d^* degrees of freedom.

The posterior covariance requires

$$E^*\{F_{Y|\eta(.)}(s_1)F_{Y|\eta(.)}(s_2)|\mathbf{\Omega},\mathbf{y}\} = \int_{\mathcal{X}} E^*[I\{\eta(\mathbf{x}) \leq s_1\}I\{\eta(\mathbf{z}) \leq s_2\}\,|\,\mathbf{\Omega},\mathbf{y}]\, dG(\mathbf{x})\; dG(\mathbf{z}).$$

Oakley and O'Hagan show that

$$P\{\eta(\mathbf{z}) \leq s_2\}P\{\eta(\mathbf{x}) \leq s_1|\eta(\mathbf{z}) \leq s_2\} = \int_{-\infty}^{s_2} P\{\eta(\mathbf{x}) \leq s_1|\eta(\mathbf{z}) = k\} f_{\eta(\mathbf{z})}(k)\, dk, \tag{3.64}$$

where $f_{\eta(\mathbf{z})}(k)$ is the density function of $\eta(\mathbf{z})$.

Now $\eta(\mathbf{x})|\eta(\mathbf{z}) = k$ also has a t-distribution, but with the additional point $\eta(\mathbf{z}) = k$. Hence, it follows that

$$E^*\{F_{Y|\eta(.)}(s_1)F_{Y|\eta(.)}(s_2)|\boldsymbol{\Omega},\mathbf{y}\} = \int_\chi\int_\chi\int_{-\infty}^{s_2}\int_{-\infty}^{\frac{s_1-\mathbf{m}_\mathbf{k}^{**}(\mathbf{X})}{\hat{\sigma}\sqrt{\frac{d^*-2}{d^*}\mathbf{c}_\mathbf{k}^{**}(\mathbf{X},\mathbf{X})}}} f_{T_{d^*+1}}\, f_{\eta(\mathbf{z})}(k) dt\, dk\, dG(\mathbf{x})\, dG(\mathbf{z}). \tag{3.65}$$

Both (3.63) and (3.65) have to be evaluated numerically however these integrals are cheap calculations. Inference for the distribution function can also be made using the simulation method described in *Oakley and O'Hagan*(49).

3.4 Sensitivity Analysis

We now move onto the extensions for sensitivity analysis, recently developed by *Oakley and O'Hagan*(51). In chapter 2 we considered the decomposition of $\eta(\mathbf{x})$ into main effects and interactions

$$y = \eta(\mathbf{x}) = E(Y) + \sum_{i=1}^{p} z_i(x_i) + \sum_{1\leq i\leq j} z_{ij}(x_i, x_j) + \ldots + z_{1,\ldots,p}(\mathbf{x}), \tag{3.66}$$

where we explicitly defined $z_i(x_i)$, $z_{ij}(x_i, x_j)$ etc in chapter 2 (see equations 2.13-2.15).

These expressions require us to calculate expectations, $E(Y|\mathbf{X_r} = \mathbf{x_r}, \eta(.))$, where $\mathbf{X_r}$ is a sub vector of $\mathbf{X}$. The expectation, $K_{1,r} = E(Y|\mathbf{X_r} = \mathbf{x_r}, \eta(.))$, where subscript r identifies the expectation is conditional on $\mathbf{X_r} = \mathbf{x_r}$, can be written as

$$K_{1,r} = \int_{\chi_{-r}} \eta(\mathbf{x})\, dG_{\mathbf{x_{-r}}|\mathbf{x_r}}(\mathbf{x_{-r}}|\mathbf{x_r}), \tag{3.67}$$

where in this notation (consistent with chapter 2), χ_{-r} denotes the space of possible values for $\mathbf{x_{-r}}$, and $G_{\mathbf{x_{-r}}|\mathbf{x_r}}$ denotes the conditional distribution, $\mathbf{X_{-r}}|\mathbf{X_r}$.

Oakley and O'Hagan(51) extended the work of *Haylock and O'Hagan*(23) in order to calculate the posterior distribution of conditional expectation $K_{1,r}$. The expectation is given by

$$\hat{k}_{1,r} = E^*\{K_{1,r}|\sigma^2, \mathbf{\Omega}, \mathbf{y}\} = \mathbf{R_r}(\mathbf{x_r})\hat{\beta} + \mathbf{T_r}(\mathbf{x_r})\mathbf{A}^{-1}(\mathbf{y} - \mathbf{H}\hat{\beta}), \qquad (3.68)$$

where $\mathbf{R_r}(\mathbf{x_r})$ and $\mathbf{T_r}(\mathbf{x_r})$ are given by

$$\begin{aligned} \mathbf{R_r}(\mathbf{x_r}) &= \int_{\mathcal{X}_{-r}} \mathbf{h}(\mathbf{x})^T dG_{-r\,|\,r}(\mathbf{x_{-r}}\,|\,\mathbf{x_r}), & (3.69)\\ \mathbf{T_r}(\mathbf{x_r}) &= \int_{\mathcal{X}_{-r}} \mathbf{t}(\mathbf{x})^T dG_{-r\,|\,r}(\mathbf{x_{-r}}\,|\,\mathbf{x_r}). & (3.70) \end{aligned}$$

Oakley and O'Hagan also provide the following result for covariances

$$\begin{aligned} & Cov^*\{K_{1,r},\ K_{1,s}|\sigma^2, \mathbf{\Omega}, \mathbf{y}\} \\ = \ & \sigma^2 \int_{\mathcal{X}_{-r}} \int_{\mathcal{X}_{-s}} \mathbf{c}^{**}(\mathbf{x}, \mathbf{x}')\ dG_{-r\,|\,r}(\mathbf{x_{-r}}\,|\,\mathbf{x_r})\, dG_{-s\,|\,s}(\mathbf{x}'_{\mathbf{-s}}\,|\,\mathbf{x}'_{\mathbf{s}}) \\ = \ & \sigma^2 W_{r,s} = \sigma^2\{U_{r;s}(\mathbf{x_r}, \mathbf{x}'_{\mathbf{s}}) - \mathbf{T_r}(\mathbf{x_r})\mathbf{A}^{-1}\mathbf{T_s}(\mathbf{x_s})^{\mathbf{T}} + (\mathbf{R_r}(\mathbf{x_r}) - \mathbf{T_r}(\mathbf{x_r})\mathbf{A}^{-1}\mathbf{H}) \\ & \times (\mathbf{H}^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{H})^{-1}(\mathbf{R_s}(\mathbf{x_s}) - \mathbf{T_s}(\mathbf{x_s})\mathbf{A}^{-1}\mathbf{H})^{\mathbf{T}}\}. \quad (3.71) \end{aligned}$$

where

$$U_{r;s}(\mathbf{x_r}, \mathbf{x}'_{\mathbf{s}}) = \int_{\mathcal{X}_{-r}} \int_{\mathcal{X}_{-s}} \mathbf{c}(\mathbf{x}, \mathbf{x}')\ dG_{-r\,|\,r}(\mathbf{x_{-r}}\,|\,\mathbf{x_r})\, dG_{-s\,|\,s}(\mathbf{x}'_{\mathbf{-s}}\,|\,\mathbf{x}'_{\mathbf{s}}). \qquad (3.72)$$

From the general result on covariances, (3.71), we can calculate the posterior distribution of any expectation, $K_{1,r}$, conditional on $\mathbf{y}$, σ^2 and $\mathbf{\Omega}$. After removing the conditioning on σ^2, the posteriors are t-distributions with d^* degrees of freedom:

$$\frac{K_{1,r} - \hat{k}_{1,r}}{\hat{\sigma}\sqrt{\frac{d^*-2}{d^*}W_{r,r}}}\ |\,\mathbf{\Omega},\ \mathbf{y} \sim t_{d^*}. \qquad (3.73)$$

Main Effects and Interactions

We now consider the decomposition, (3.66), of $\eta(\mathbf{x})$ into main effects and interactions. The main effect, $z_i(x_i)$, and first order interaction $z_{i,j}(x_i, x_j)$ are defined as

$$z_i(x_i) = E(Y|x_i) - E(Y), \tag{3.74}$$

$$z_{i,j}(x_i, x_j) = E(Y|x_i, x_j) - z_i(x_i) - z_j(x_j) - E(Y), \tag{3.75}$$

with higher order terms following similarly. These expressions are conditional on $\eta(.)$, but for ease of notation we don't show this explicitly.

Since $K_{1,r}|\sigma^2, \mathbf{\Omega}, \mathbf{y}$ is normally distributed for any r (including $r = 0$, the null set), it is simple to note that conditional on $\mathbf{y}$, σ^2 and $\mathbf{\Omega}$, main effects and interactions are functions of correlated normal distributions and therefore normally distributed. The expectations all follow from (3.68) and the variances can be calculated from (3.71). After removing the conditioning on σ^2, main effects and interactions have t-distributions with d^* degrees of freedom.

In particular the expectations of main effects and first order interaction are

$$\begin{aligned} E^*\{z_i(x_i)|\mathbf{y}\} &= \{\mathbf{R}_i(x_i) - \mathbf{R}\}\hat{\beta} + \{\mathbf{T}_i(x_i) - \mathbf{T}\}\mathbf{A}^{-1}(\mathbf{y} - \mathbf{H}\hat{\beta}), \\ E^*\{z_{ij}(x_i, x_j)|\mathbf{y}\} &= \{\mathbf{R}_{ij}(x_i, x_j) - \mathbf{R}_i(x_i) - \mathbf{R}_j(x_j) + \mathbf{R}\}\hat{\beta} \\ &+ \{\mathbf{T}_{ij}(x_i, x_j) - \mathbf{T}_i(x_i) - \mathbf{T}_j(x_j) + \mathbf{T}\}\mathbf{A}^{-1}(\mathbf{y} - \mathbf{H}\hat{\beta}), \end{aligned}$$

with higher order terms following similarly.

In chapter 2 we quoted the result from *Chan et al.*(9), that typically, as the order of the interaction increases, then $z_{1,\ldots,r}(x_1, \ldots x_r) \rightarrow 0$. We would expect the same result to hold for the expectations (with respect to the posterior distribution of $\eta(.)$). However, the posterior variances of these terms may be large.

Variances

Finally we consider the decomposition of the variance. The variance of the output conditional on sub vector $\mathbf{X_r}$, is given by

$$\begin{aligned} Var\{E(Y\,|\mathbf{X_r},\eta(.))\} &= E\{E(Y\,|\mathbf{X_r},\eta(.))^2\} - E\{E(Y\,|\mathbf{X_r},\eta(.))\}^2 \\ &= E\{E(Y\,|\mathbf{X_r},\eta(.))^2\} - E(Y|\eta(.))^2. \end{aligned} \qquad (3.76)$$

Oakley and O'Hagan found these to be intractable, and just calculated expectations. We found the posterior distribution of $E(Y)$ earlier, and hence $E^*\{E(Y)^2|\mathbf{\Omega},\mathbf{y}\}$ is known. We now have to calculate $E^*[E\{E(Y\,|\mathbf{X_r})^2\}|\mathbf{\Omega},\mathbf{y}]$.

$$\begin{aligned} E^* \quad & [E\{E(Y\,|\mathbf{X_r})^2\}|\mathbf{\Omega},\mathbf{y}] \\ = \quad & \int_{\mathcal{X}_r}\int_{\mathcal{X}_{-r}}\int_{\mathcal{X}_{-r}} E^*\{\eta(\mathbf{x})\eta(\mathbf{x}^*)\}\,dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x_{-r}}\,|\mathbf{x_r})\,dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x'_{-r}}\,|\mathbf{x_r})\,dG_{\mathbf{r}}(\mathbf{x_r}) \\ = \quad & \int_{\mathcal{X}_r}\int_{\mathcal{X}_{-r}}\int_{\mathcal{X}_{-r}} \hat{\sigma}^2\ \mathbf{c}^{**}(\mathbf{x},\mathbf{x}^*)dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x_{-r}}\,|\mathbf{x_r})\,dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x'_{-r}}\,|\mathbf{x_r})\,dG_{\mathbf{r}}(\mathbf{x_r}) \\ + \quad & \int_{\mathcal{X}_r}\int_{\mathcal{X}_{-r}}\int_{\mathcal{X}_{-r}} \mathbf{m}^{**}(\mathbf{x})\mathbf{m}^{**}(\mathbf{x}^*)dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x_{-r}}\,|\mathbf{x_r})\,dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x'_{-r}}\,|\mathbf{x_r})\,dG_{\mathbf{r}}(\mathbf{x_r}).(3.77) \end{aligned}$$

We use $\mathbf{x}^*$ to denote the vector $\mathbf{x}^* = (\mathbf{x_r},\mathbf{x'_{-r}})$, whilst $\mathbf{x} = (\mathbf{x_r},\mathbf{x_{-r}})$, and $G_{\mathbf{r}}(.)$ denotes the marginal distribution of $\mathbf{X_r}$. We can see that the form of (3.77) is similar to the equations *Sobol'*(63) derived (see equations (2.31)-(2.34) of chapter 2).

The first term of (3.77) can be expanded as

$$\begin{aligned} &\hat{\sigma}^2[U_r - tr(\mathbf{A}^{-1}\mathbf{P_r}) + tr((\mathbf{H^T}\mathbf{A}^{-1}\mathbf{H})^{-1} \times \\ &\qquad\qquad (\mathbf{Q_r} - \mathbf{S_r}\mathbf{A}^{-1}\mathbf{H} - \mathbf{H^T}\mathbf{A}^{-1}\mathbf{S_r^T} + \mathbf{H^T}\mathbf{A}^{-1}\mathbf{P_r}\mathbf{A}^{-1}\mathbf{H}))], \end{aligned}$$

and the second term is expanded as

$$tr((\mathbf{y}-\mathbf{H}\hat{\beta})^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{P}_{\mathbf{r}}\mathbf{A}^{-1}(\mathbf{y}-\mathbf{H}\hat{\beta}))+2tr(\hat{\beta}\mathbf{S}_{\mathbf{r}}\mathbf{A}^{-1}(\mathbf{y}-\mathbf{H}\hat{\beta}))+tr(\hat{\beta}^{\mathbf{T}}\mathbf{Q}_{\mathbf{r}}\hat{\beta}),$$

where

$$\begin{aligned}
U_r &= \int_{\mathcal{X}_r}\int_{\mathcal{X}_{-r}}\int_{\mathcal{X}_{-r}} \mathbf{c}(\mathbf{x},\mathbf{x}^*)\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}'_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{\mathbf{r}}(\mathbf{x}_{\mathbf{r}}),\\
\mathbf{P}_{\mathbf{r}} &= \int_{\mathcal{X}_r}\int_{\mathcal{X}_{-r}}\int_{\mathcal{X}_{-r}} \mathbf{t}(\mathbf{x})\mathbf{t}(\mathbf{x}^*)^{\mathbf{T}}\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}'_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{\mathbf{r}}(\mathbf{x}_{\mathbf{r}}),\\
\mathbf{Q}_{\mathbf{r}} &= \int_{\mathcal{X}_r}\int_{\mathcal{X}_{-r}}\int_{\mathcal{X}_{-r}} \mathbf{h}(\mathbf{x})\mathbf{h}(\mathbf{x}^*)^{\mathbf{T}}\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}'_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{\mathbf{r}}(\mathbf{x}_{\mathbf{r}}),\\
\mathbf{S}_{\mathbf{r}} &= \int_{\mathcal{X}_r}\int_{\mathcal{X}_{-r}}\int_{\mathcal{X}_{-r}} \mathbf{h}(\mathbf{x})\mathbf{t}(\mathbf{x}^*)^{\mathbf{T}}\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}'_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{\mathbf{r}}(\mathbf{x}_{\mathbf{r}}).
\end{aligned}$$

The above integrals can be calculated analytically for common choices of $\mathbf{h}(.)$, $\mathbf{c}(.,.)$ and $G(\mathbf{x})$. However, the integrals are all computationally cheap functions, so numerical methods do not take an excessive amount of time.

Thus, from (3.76) the expectation (with respect to the posterior distribution of $\eta(.)$) of the main effect and total effect variances, as defined in equations 2.23 and 2.24 of chapter 2, can be calculated. *Oakley and O'Hagan*(51) estimated main and total effect indices by dividing the variances by $E^*\{Var(Y)|\mathbf{\Omega},\mathbf{y}\}$. *Oakley and O'Hagan* noted that these estimates are not the same as the posterior expectations of the main and total effects, which would be intractable. However, the approximation appears to be a good one, and results in a lower computational burden.

3.5 Conclusions

In this chapter we have reviewed a Bayesian method for making inference about computationally expensive functions using Gaussian Processes. We examined a previously proposed but untried correlation function and noted we can make significant improvements over standard product correlation forms when large interactions are present. We also noted a connection between the geometry of the inputs and the correlation function, which could be utilized in order to improve computational efficiency. Finally, we reviewed how Gaussian Processes could be used in order to calculate measures of uncertainty and sensitivity. These calculations represented small corrections to those in the cited papers.

Chapter 4

Decomposable Models

4.1 Introduction

We now consider some modifications to the methodology of chapter 3, that are appropriate when we have additional information about the function, $\eta(.)$.

In some computer models (one of which we discuss in chapter 7), although it is not known how the output, $\eta(\mathbf{x})$, varies as we vary the inputs, $\mathbf{x}$, there is additional information about the structure of $\eta(.)$. For example in a computer model representing a physical system we might know that it is impossible for two groups of inputs to interact. As a result of this information we can simplify $\eta(.)$ to

$$\eta(.) = \eta_1(.) + \eta_2(.), \tag{4.1}$$

where $\eta_1(.)$ and $\eta_2(.)$ are functions of lower dimensional input vectors $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$ and $\mathbf{x}$ partitions as $\mathbf{x} = \{\mathbf{x}_{(1)}, \mathbf{x}_{(2)}\}$. Note that the vector $\mathbf{x}_{(i)}$ and the design point $\mathbf{x}_i$ are quite distinct.

The decomposition, (4.1), is a special case of a more general form of decomposition which we now define. In general we have a p dimensional vector of inputs, $\mathbf{x} = \{x_1, x_2, \ldots, x_p\}$. We let $S = \{1, 2, \ldots p\}$ denote the set of integers, and we let $S_1, \ldots, S_r$ denote subsets of S.

We can write $\eta(\mathbf{x})$ as the sum

$$\eta(\mathbf{x}) = \eta_1(\mathbf{x}_{(1)}) + \ldots + \eta_{(r)}(\mathbf{x}_{(r)}) \tag{4.2}$$

where $\mathbf{x}_{(i)} = \{x_{r_1^{(i)}}, \ldots, x_{r_{n_i}^{(i)}}\}$, and $\{r_1^{(i)}, \ldots, r_{n_i}^{(i)}\} = S_i$. The special case of (4.1) is found when we have $\bigcup_{i=1}^{2} S_i = S$ and $S_1 \bigcap S_2 = \varnothing$.

The use of structural information in order to write $\eta(\mathbf{x})$ as in (4.2) can be thought of as "opening the black box". The strength of structural prior information will no doubt vary from model to model. For some models an expert may be able to determine all the subsets $S_1, \ldots, S_r$, whilst in other models we genuinely have no idea about the form of $\eta(.)$, although some simplification may be possible.

In this chapter we develop a series of models that account for different levels of prior information. We begin the chapter by examining the role that smoothness, and in particular the correlation function plays in functions of many inputs, since this relates to our work later in the chapter. In sections 4.3 and 4.4 we develop models for the case where we have a known decomposition of $\eta(.)$. In section 4.5 we develop additional theory for when our prior information is weaker. We consider decompositions for weak structural information in sections 4.6 and 4.7.

4.2 Parametric Approximations

In the problems that interest us, the dimension, p, of the vector of inputs, $\mathbf{x}$, is large. Computer models often contain many uncertain inputs (in the example we consider in chapter 7 we have $p = 88$). We require robust methodology that can handle problems of this magnitude without an excessive number of design points.

The key feature of the Gaussian Process model that we described in chapter 3, is the assumed correlation structure – the extent to which learning the output, $\eta(\mathbf{x})$, aids our learning about output, $\eta(\mathbf{x}')$. In this chapter we restrict our attention to the correlation function defined in (3.4), which can be written as

$$\mathbf{c}(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{p} c_i(x_i, x_i'), \tag{4.3}$$

where $c_i(x_i, x_i') \leq 1$, and hence $\mathbf{c}(\mathbf{x}, \mathbf{x}') \leq 1$.

The predictive performance of our Gaussian Process model, relative to a standard regression model, depends on 2 related elements:

1. the number of design points;
2. the smoothness of the function.

The rougher the output is (as a function of the inputs), the more design points we will require in order to produce significant improvements upon a standard regression model.

Oakley and O'Hagan(51) managed to tackle a 40 dimensional problem with just 101 design points, whilst *Welch et al.*(64) found that with a well chosen design they could tackle a $30-40$ dimensional problem with as few as 50 runs of the computer model. In both papers, the authors found that most of the output variability was caused by a just a few active inputs. As a result, a well chosen LHD is able to reduce the dimension of the problem to well below p.

Welch et al.(64) fit a model of the form

$$\eta(\mathbf{x}) = \mu + \mathbf{Z}(\mathbf{x}), \tag{4.4}$$

where $\mathbf{Z}(.)$ is a Gaussian Process with zero mean and covariance $\sigma^2\mathbf{c}(.,.)$.

The interpretation of (4.4) is that we have some global point estimate, μ, for any value of the inputs $\mathbf{x}$. The Gaussian Process, $\mathbf{Z}(.)$, corrects this estimate, taking into account the local correlation structure between outputs. If we know the output $\eta(\mathbf{x})$ at $\mathbf{x}$, and this is larger than μ, then for a smooth function the output $\eta(\mathbf{x}')$ at some adjacent input $\mathbf{x}'$, is also likely to be larger than μ.

Welch et al. used the components of (4.3) in order to assess the sensitivity of the model output to the inputs. If the inputs are scaled to have the same range, the parameters of the i^{th} term in the product, (4.3), are a measure of the importance of input i. If the output is active with respect to the i^{th} input, then the correlation, $c_i(x_i, x_i')$, will depend strongly upon $|x_i - x_i'|$, whereas if input i is relatively inactive then $c_i(x_i, x_i')$ will be close to 1 regardless of $|x_i - x_i'|$. A model of the form (4.4) requires most of the model inputs to be inactive, or at least relatively inactive, for a high value of p, unless we have many design points.

Welch et al. used the correlation function

$$\mathbf{c}(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{p} \exp\{-b_i|x - x'|^{a_i}\} \qquad 0 \leq a_i \leq 2, \tag{4.5}$$

estimating the parameters by maximum likelihood.

The active dimensions are identified by large values of b_i (the most active having the largest b_i), whilst relatively inactive and completely dormant inputs have very small and zero values of b_i respectively. *Welch et al.* found that setting all the $a_i = 2$, (which is desirable for a differentiable function), results in little loss in terms of predictive performance.

In chapter 3 we considered a more general mean function, where we replaced

μ by a regression fit

$$\eta(\mathbf{x}) = \mathbf{h}(\mathbf{x})^{\mathbf{T}}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{x}). \tag{4.6}$$

By changing our mean function, the interpretation of $\mathbf{Z}(.)$ changes. This is now a Gaussian Process on the residuals from our parametric approximation to $\eta(.)$. If we know the residual at $\mathbf{x}$, which we denote $e(\mathbf{x})$, to be large and positive, then for a smooth function, the residual $e(\mathbf{x}')$ at some adjacent set of inputs, $\mathbf{x}'$, is also likely to be large and positive.

Our interpretation of the terms in the product (4.3) also changes. The parameters of the i^{th} term in this product are no longer a measure of how sensitive the output is to input i; they are a measure of how smooth the departures are from the mean function in dimension i. A special case is when the regression fit explains all the variability in dimension i, and $c_i(x_i, x_i') = 1$ regardless of $|x_i - x_i'|$.

We illustrate with the example

$$\eta(x_1, x_2, x_3) = 5 + 2x_1 - 0.3x_1^2 + \cos(x_2). \tag{4.7}$$

Taking $\mathbf{h}(\mathbf{x}) = (1, x_1, x_1^2, x_2, x_3)$, we can see that the mean function fits exactly in x_1, and therefore $c_1(x_1, x_1') = 1 \forall |x_1 - x_1'|$. We also find that $c_3(x_3, x_3') = 1 \forall |x_3 - x_3'|$ since a completely dormant input is a special case of the perfectly fitting parametric approximation. Therefore, in (4.7) our correlation function reduces to a function of x_2 only.

In general we could consider any vector of regressor variables, $\mathbf{h}(\mathbf{x})$, including interaction terms. We should incorporate any available expert information when selecting $\mathbf{h}(\mathbf{x})$. Our work in sections 4.6 and 4.7 of this chapter examines a previously untried non parametric form for the mean function.

With a well chosen mean function, and a Latin hypercube design, we are able, in effect, to reduce the dimension of the input vector, $\mathbf{x}$, to well below p. *Welch et al.* considered examples where $p \approx 40$ with only 50 design points, and 5 active dimensions. However, 50 points is vastly inadequate for a 5 dimensional problem if we have large high order interactions, since high dimensional space is only sparsely covered with this small number of points. Therefore, in addition to factor sparsity we also require relative simplicity of $\eta(.)$ in the active dimensions.

The Gaussian Process model is most efficient when high order interactions are negligible compared with main effects and low order interactions. This is because the Latin Hypercube design (or any good design in general) covers marginal distributions and low order space well, but only covers high order space sparsely. In some models, such as the financial models that motivate this research, we expect some interaction terms to be zero. In this case we can identify more efficient correlation structures. In the remainder of this chapter we examine different correlation structures for when $\eta(.)$ can be simplified to a sum of lower order terms.

4.3 Known Additive Decomposition

Suppose that we have a known mutually exclusive and exhaustive partition of $\mathbf{x} = \{\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(r)}\}$ and a corresponding decomposition of $\eta(.)$ into functions of lower dimensional input vectors, that is $S_i \bigcap S_j = \varnothing \, \forall i \neq j$. Then we can write output $\eta(\mathbf{x})$ as

$$\eta(\mathbf{x}) = \eta_1(\mathbf{x}_{(1)}) + \ldots + \eta_r(\mathbf{x}_{(r)}), \qquad (4.8)$$

where the $\eta_j(.)$ are independent functions of sub-vectors $\mathbf{x}_{(j)}$.

In this section we examine two models that depend on very strong structural prior information of this type. We go on to derive posterior distributions for $\eta(.)$ using methodology similar to that of chapter 3.

4.3.1 Observable Functions

In the simplest model we consider, $\eta(.)$ is a fairly transparent function. We are able to make direct observations, $y_{(1)}, y_{(2)}, \ldots, y_{(r)}$, of $\eta_1(.), \eta_2(.), \ldots, \eta_r(.)$ respectively, and $\sum_{j=1}^{r} y_{(j)} = y$ are our observations of $\eta(.)$. Our interest is in finding a cheap approximation to $\eta(.)$ using this additional structural information. This is the strongest possible structural prior information that we might expect.

Since we can make observations of each of these sub functions, we are able to model each of them using the methodology of chapter 3. For function j, our prior estimate of $\mathbf{x}_{(j)}$ is $\mathbf{h}_j(\mathbf{x}_{(j)})^T\boldsymbol{\beta}_j$ and our prior variance is σ_j^2. In general our mean function will differ for each function.

For function j we have hyperparameters $\boldsymbol{\beta}_j, \sigma_j^2$ and $\boldsymbol{\omega}_j$. We adopt a change in notation from chapter 3 here (from $\boldsymbol{\Omega}_j$ to $\boldsymbol{\omega}_j$ since $\boldsymbol{\omega}_j$ is a vector rather than a matrix of parameters). Information about the hyperparameters of $\eta_j(.)$ is likely to be weak. We adopt the non informative prior,

$$p(\boldsymbol{\beta}_j, \sigma_j^2, \boldsymbol{\omega}_j) \propto \sigma_j^{-2}. \tag{4.9}$$

We make n_j observations of $\eta_j(.)$, and follow the prior to posterior analysis of chapter 3. Our posterior distributions are student processes. For a given input $\mathbf{x} = \{\mathbf{x}_{(1)}, \ldots \mathbf{x}_{(r)}\}$, we have

$$\begin{aligned}
\frac{\eta_1(\mathbf{x}_{(1)}) - \mathbf{m}_1^{**}(\mathbf{x}_{(1)})}{\hat{\sigma}_1\sqrt{\frac{n_1-q_1-2}{n_1-q_1}\mathbf{c}_1^{**}(\mathbf{x}_{(1)},\mathbf{x}_{(1)})}}|\mathbf{y}_{(1)},\boldsymbol{\omega}_1 &\sim t_{n_1-q_1}, \\
&\vdots \\
\frac{\eta_r(\mathbf{x}_{(r)}) - \mathbf{m}_r^{**}(\mathbf{x}_{(r)})}{\hat{\sigma}_r\sqrt{\frac{n_r-q_r-2}{n_r-q_r}\mathbf{c}_r^{**}(\mathbf{x}_{(r)},\mathbf{x}_{(r)})}}|\mathbf{y}_{(r)},\boldsymbol{\omega}_r &\sim t_{n_r-q_r},
\end{aligned}$$

where q_j denotes the dimension of $\boldsymbol{\beta}_j$. Terms $\hat{\sigma}_j, \mathbf{m}_j^{**}(\mathbf{x}_{(j)})$ and $\mathbf{c}_j^{**}(\mathbf{x}_{(j)},\mathbf{x}_{(j)})$ are calculated from (3.20), (3.26) and (3.27) respectively from the previous chapter.

We wish to make inference about $\eta(.)$. The posterior distribution for some new value, $\mathbf{x}$, is the sum of r t-distributions, which has no closed form. Since $\eta_1(.), \dots, \eta_r(.)$ are taken to be independent a priori, and since the data are independent, our posteriors $\eta_1(.)|\mathbf{y}_{(1)}, \boldsymbol{\omega}_1, \dots, \eta_r(.)|\mathbf{y}_{(r)}, \boldsymbol{\omega}_r$ are also independent. Therefore, we can easily calculate posterior moments of $\eta(.)$ such as the expectation and variance. However, we frequently wish to provide probability bounds for a new observation, such as a 95% interval, and in the absence of a closed probability distribution this can only be done numerically. Inference in this manner for many values of $\mathbf{x}$ will prove to be time consuming, especially if r is large.

We instead turn our attention to approximate results. For large r (and $n_1, \dots n_r$ not too small), by the central limit theorem we can approximate the output at $\mathbf{x}$ by a normal distribution. The expectation and variance of this approximation would be correct, however the tails would be too light, and for small r this would be a poor approximation. A better approximation can be found using a member of the Pearson family, and we require only the mean, variance, skewness and kurtosis (see *Johnson et al.* (33)) in order to fit a distribution.

Since the skewness is zero and the kurtosis will always be > 3, a Pearson type *VII* distribution is a suitable approximation,

$$p(x) = \frac{\Gamma(d)}{\sqrt(\pi)\Gamma(d - 1/2)} \frac{c^{2d-1}}{(c^2 + (x - \xi)^2)^d}, \tag{4.10}$$

for $c > 0$ and $d > 0$.

The t-distribution is a member of this family, and *Johnson et al.* (33) find that (4.10) can be found as a simple multiplicative transformation of a t-distribution. Our approximation to $\eta(\mathbf{x})$ therefore takes the form of a t-distribution. We show how to calculate the parameters of this distribution below.

For ease of notation we let $v_j = n_j - q_j$ for $j = 1, \dots r$ and

$$Z_j = \frac{\eta_j(\mathbf{x}_{(j)}) - \mathbf{m}_j^{**}(\mathbf{x}_{(j)})}{\hat{\sigma}_j \sqrt{\frac{v_j - 2}{v_j} \mathbf{c}_\mathbf{j}^{**}(\mathbf{x}_{(j)}, \mathbf{x}_{(j)})}} | \mathbf{y}_{(j)}, \boldsymbol{\omega}_j, \tag{4.11}$$

is a standard t-distribution with v_j degrees of freedom.

The posterior expectation and variance of $\eta(\mathbf{x})$ are given by

$$\begin{aligned} E[\eta(\mathbf{x})|\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(r)}, \boldsymbol{\omega}_1, \dots \boldsymbol{\omega}_r] &= \mathbf{m}_1^{**}(\mathbf{x}_{(1)}) + \dots + \mathbf{m}_r^{**}(\mathbf{x}_{(r)}), & (4.12) \\ Var[\eta(\mathbf{x})|\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(r)}, \boldsymbol{\omega}_1, \dots \boldsymbol{\omega}_r] &= \frac{v_1 - 2}{v_1} \hat{\sigma}_1{}^2 \mathbf{c}_\mathbf{1}^{**}(\mathbf{x}_{(1)}, \mathbf{x}_{(1)}) Var(Z_1) \dots \\ &+ \frac{v_r - 2}{v_r} \hat{\sigma}_r{}^2 \mathbf{c}_\mathbf{r}^{**}(\mathbf{x}_{(r)}, \mathbf{x}_{(r)}) Var(Z_r). & (4.13) \end{aligned}$$

We now consider the kurtosis of Z_j, which is given by

$$\beta_2(t_{v_j}) = \frac{\mu_4}{\mu_2^2} = \frac{E[Z_j^4]}{E[Z_j^2]^2}, \tag{4.14}$$

where following *Johnson et al.*(33), we use $\beta_2(t_v)$ to represent the kurtosis. and μ_i denotes the i^{th} cental moment of Z_j. We have the simplification shown in (4.14) since the odd moments of the t-distribution are zero.

For a t-distribution the kurtosis is given by

$$\beta_2(t_{v_j}) = 3\frac{v_j - 2}{v_j - 4}. \tag{4.15}$$

The kurtosis of Z_j invariant to shifting and scaling, therefore $a + bZ_j$ and Z_j have the same kurtosis. Applying this result we find that the kurtosis of $\eta_j(\mathbf{x}_{(j)})$ and Z_j are the same. However, we require the kurtosis of the sum $\eta_1(\mathbf{x}_{(1)}) + \dots + \eta_r(\mathbf{x}_{(r)})$, and therefore have to take the scaling into account.

The kurtosis of this sum is given by

$$\beta_2(t_v) = \frac{E\{[\sum_{j=1}^{r} \hat{\sigma}_j \sqrt{\frac{v_j-2}{v_j}\mathbf{c}_\mathbf{j}^{**}(\mathbf{x}_{(j)}, \mathbf{x}_{(j)})} Z_j]^4\}}{E\{[\sum_{j=1}^{r} \hat{\sigma}_j \sqrt{\frac{v_j-2}{v_j}\mathbf{c}_\mathbf{j}^{**}(\mathbf{x}_{(j)}, \mathbf{x}_{(j)})} Z_j]^2\}^2}. \tag{4.16}$$

The denominator of (4.16) is easily found from (4.13) therefore we just need to evaluate the numerator. In principle we can expand this expression by repeated use of the binomial expansion. Each term is of the form

$$a_{m_1m_2m_3m_4} E\{[\hat{\sigma}_i \sqrt{\frac{v_i-2}{v_i}\mathbf{c}_\mathbf{i}^{**}(\mathbf{x}_{(i)}, \mathbf{x}_{(i)})} Z_i]^{m_1}\} E\{[\hat{\sigma}_j \sqrt{\frac{v_j-2}{v_j}\mathbf{c}_\mathbf{j}^{**}(\mathbf{x}_{(j)}, \mathbf{x}_{(j)})} Z_j]^{m_2}\}$$
$$\times\ E\{[\hat{\sigma}_k \sqrt{\frac{v_k-2}{v_k}\mathbf{c}_\mathbf{k}^{**}(\mathbf{x}_{(k)}, \mathbf{x}_{(k)})} Z_k]^{m_3}\} E\{[\hat{\sigma}_l \sqrt{\frac{v_l-2}{v_l}\mathbf{c}_\mathbf{l}^{**}(\mathbf{x}_{(l)}, \mathbf{x}_{(l)})} Z_l]^{m_4}\}, \tag{4.17}$$

for coefficients $a_{m_1m_2m_3m_4}$ and integers m_1, m_2, m_3, m_4, where $m_i \geq 0$ and $m_1 + m_2 + m_3 + m_4 = 4$. However, by noting that $E[Z_j^m] = 0$ for m odd, it is obvious that almost all these terms will cancel. The numerator simplifies to

$$E\{[\sum_{j=1}^{r} \hat{\sigma}_j \sqrt{\frac{v_j-2}{v_j}\mathbf{c}_\mathbf{j}^{**}(\mathbf{x}_{(j)}, \mathbf{x}_{(j)})} Z_j]^4\} = \sum_{i=1}^{r} [\hat{\sigma}_i \sqrt{\frac{v_i-2}{v_i}\mathbf{c}_\mathbf{i}^{**}(\mathbf{x}_{(i)}, \mathbf{x}_{(i)})}]^4 E\{Z_i^4\}$$
$$+\ \ 6\sum_{i<j} [\hat{\sigma}_i \sqrt{\frac{v_i-2}{v_i}\mathbf{c}_\mathbf{i}^{**}(\mathbf{x}_{(i)}, \mathbf{x}_{(i)})}]^2 [\hat{\sigma}_j \sqrt{\frac{v_j-2}{v_j}\mathbf{c}_\mathbf{j}^{**}(\mathbf{x}_{(j)}, \mathbf{x}_{(j)})}]^2 E\{Z_i^2\}E\{Z_j^2\}, \tag{4.18}$$

where

$$\begin{aligned} E\{Z_j^4\} &= 3\frac{v_j^2}{(v_j-4)(v_j-2)}, \\ E\{Z_i^2\}E\{Z_j^2\} &= \frac{v_i}{v_i-2}\frac{v_j}{v_j-2}. \end{aligned}$$

Thus, by equating (4.15) and (4.16) and rearranging, we find v as

$$v = \frac{4\beta_2(t_v) - 6}{\beta_2(t_v) - 3}. \tag{4.19}$$

We have the variance from (4.13), but in order to ensure that variance of our approximation is correct we have to scale (4.13) by $Var(t_v)$ the variance of our t-distribution (with v degrees of freedom). Hence

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{Var(t_v)}\{\frac{n_1-q_1-2}{n_1-q_1}\hat{\sigma_1}^2\mathbf{c}_1^{**}(\mathbf{x}_{(1)},\mathbf{x}_{(1)})Var(t_{n_1-q_1})+\dots \\ &\qquad \frac{n_r-q_r-2}{n_r-q_r}\hat{\sigma_r}^2\mathbf{c}_r^{**}(\mathbf{x}_{(r)},\mathbf{x}_{(r)})Var(t_{n_r-q_r})\}, \qquad (4.20)\end{aligned}$$

and

$$\frac{\eta(\mathbf{x})-E\{\eta(\mathbf{x})\}}{\hat{\sigma}}|\mathbf{y}_{(1)},\dots\mathbf{y}_{(r)},\boldsymbol{\omega}_1,\dots\boldsymbol{\omega}_r \sim t_v, \qquad (4.21)$$

We demonstrate our approximation by considering the sum of $X = Y_1+Y_2$, where the Y_i are t-distributed with 7 degrees of freedom. From (4.12) the expectation of this sum is zero, and from (4.13) the variance is given by $7/5+7/5 = 14/5$. The kurtosis of X is 4, and we find the t-distribution with this kurtosis from (4.19), solving for $v = 10$ degrees of freedom. Our approximation is therefore

$$\frac{X}{\sqrt{2.8/1.25}} \sim t_{10}.$$

We plot this approximation and the true (numerically evaluated) distribution in *Figure* (4.1). As we can see from the plot, our approximation works well, the two densities showing little separation well into the tails of X.

4.3.2 Unobservable Functions

We once again consider a model of the form (4.8), with a known additive decomposition. However, we now suppose we are only able to directly observe $\eta(\mathbf{x})$. This decomposition requires very precise structural prior information. Once more, we use this additional structural information and find the posterior distribution of

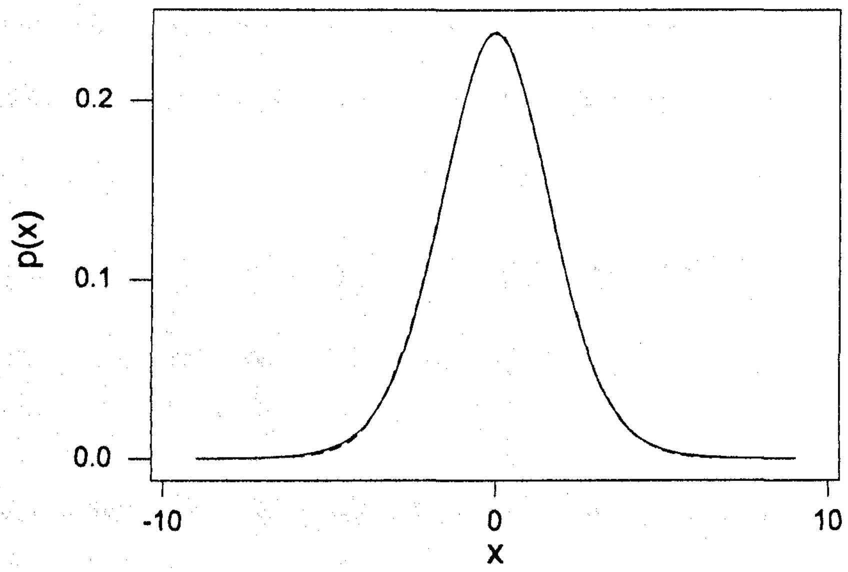

Figure 4.1: Sum of t-distributions: our approx (dash), simulated true density (solid)

$\eta(.)$. For ease of notation, we consider inference for $r = 2$, but the extension to $r > 2$ is straightforward.

Since we know $\eta(.)$ is additive our model can be written as

$$\eta(.) = \eta_1(.) + \eta_2(.), \tag{4.22}$$

$$\eta_1(.) = \eta(., \mathbf{x}_{(2)} = \mathbf{a}_2) + c_2, \tag{4.23}$$

$$\eta_2(.) = \eta(\mathbf{x}_{(1)} = \mathbf{a}_1, .) + c_1, \tag{4.24}$$

for constants $c_1 = -\eta_1(\mathbf{a}_1)$ and $c_2 = -\eta_2(\mathbf{a}_2)$.

We adopt Gaussian Process priors on $\eta(., \mathbf{x}_{(2)} = \mathbf{a}_2)$ and $\eta(\mathbf{x}_{(1)} = \mathbf{a}_1, .)$, and make observations $\mathbf{y}_{(1)} = \{\eta(\mathbf{x}_{(1)1}, \mathbf{x}_{(2)} = \mathbf{a}_2), \ldots, \eta(\mathbf{x}_{(1)n_1}, \mathbf{x}_{(2)} = \mathbf{a}_2)\}$ of $\eta(., \mathbf{x}_{(2)} = \mathbf{a}_2)$ and $\mathbf{y}_{(2)} = \{\eta(\mathbf{x}_{(1)} = \mathbf{a}_1, \mathbf{x}_{(2)1}), \ldots \eta(\mathbf{x}_{(1)} = \mathbf{a}_1, \mathbf{x}_{(2)n_2})\}$ of $\eta(\mathbf{x}_{(1)} = \mathbf{a}_1, .)$.

Following the same methodology as in 4.2.1, we arrive at posterior distributions for $\eta(\mathbf{x}_{(1)}, \mathbf{x}_2 = \mathbf{a}_2)$ and $\eta(\mathbf{x}_{(1)} = \mathbf{a}_1, \mathbf{x}_{(2)})$ of

$$\frac{\eta(\mathbf{x}_{(1)}, \mathbf{x}_{(2)} = \mathbf{a}_2) - \mathbf{m}_1^{**}(\mathbf{x}_{(1)})}{\hat{\sigma}_1\sqrt{\frac{n_1-q_1-2}{n_1-q_1}\mathbf{c}_1^{**}(\mathbf{x}_{(1)}, \mathbf{x}_{(1)})}}|\mathbf{y}_{(1)}, \boldsymbol{\omega}_1 \quad \sim \quad t_{n_1-q_1}, \tag{4.25}$$

$$\frac{\eta(\mathbf{x}_{(1)} = \mathbf{a}_1, \mathbf{x}_{(2)}) - \mathbf{m}_2^{**}(\mathbf{x}_{(2)})}{\hat{\sigma}_2\sqrt{\frac{n_2-q_2-2}{n_2-q_2}\mathbf{c}_2^{**}(\mathbf{x}_{(2)}, \mathbf{x}_{(2)})}}|\mathbf{y}_{(2)}, \boldsymbol{\omega}_2 \quad \sim \quad t_{n_2-q_2}. \tag{4.26}$$

It is possible to approximate the sum of $\eta_1(\mathbf{x}_{(1)}, \mathbf{x}_{(2)} = \mathbf{a}_2|\mathbf{y}_{(1)}, \boldsymbol{\omega}_1)$ and $\eta_2(\mathbf{x}_{(1)} = \mathbf{a}_1, \mathbf{x}_2|\mathbf{y}_{(2)}, \boldsymbol{\omega}_2)$ by a t-distribution similar to (4.21). However, we want a fast approximation to $\eta(\mathbf{x}|\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$, and this is given by

$$\eta(\mathbf{x}|\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \boldsymbol{\omega}_1, \boldsymbol{\omega}_2) = \eta(\mathbf{x}_{(1)}, \mathbf{x}_{(2)} = \mathbf{a}_2|\mathbf{y}_{(1)}, \boldsymbol{\omega}_1) + \eta(\mathbf{x}_{(1)} = \mathbf{a}_1, \mathbf{x}_{(2)}|\mathbf{y}_{(2)}, \boldsymbol{\omega}_2) + c,$$

where the constant $c = c_1 + c_2$ is well defined.

We will in general require one further design point in order to estimate c. Suppose we observe response y^* at design point $\mathbf{x}^* = (\mathbf{x}^*_{(1)}, \mathbf{x}^*_{(2)})$. We can estimate $\eta(\mathbf{x}^*_{(1)}, \mathbf{x}_{(2)} = \mathbf{a}_2|\mathbf{y}_{(1)}, \boldsymbol{\omega}_1)$ by $E\{\eta(\mathbf{x}^*_{(1)}, \mathbf{x}_{(2)} = \mathbf{a}_2|\mathbf{y}_{(1)}, \boldsymbol{\omega}_1)\}$ and $\eta(\mathbf{x}_{(1)} = \mathbf{a}_1, \mathbf{x}^*_{(2)}|\mathbf{y}_{(2)}, \boldsymbol{\omega}_2)$ by $E\{\eta(\mathbf{x}_{(1)} = \mathbf{a}_1, \mathbf{x}^*_{(2)}|\mathbf{y}_{(2)}, \boldsymbol{\omega}_2)\}$. Our point estimate of c is

$$\hat{c} = y^* - E\{\eta(\mathbf{x}^*_{(1)}, \mathbf{x}_{(2)} = \mathbf{a}_2|\mathbf{y}_{(1)}, \boldsymbol{\omega}_1)\} - E\{\eta(\mathbf{x}_{(1)} = \mathbf{a}_1, \mathbf{x}^*_{(2)}|\mathbf{y}_{(2)}, \boldsymbol{\omega}_2)\}. \tag{4.27}$$

However if we take $\mathbf{x}^* = (\mathbf{a}_1, \mathbf{a}_2)$, from (4.22) we have $c = \eta_1(\mathbf{a}_1) + \eta_2(\mathbf{a}_2) = \eta(\mathbf{a}_1, \mathbf{a}_2)$, so it is possible to determine c exactly. In addition, it is possible to utilise this additional point $\mathbf{x}^*$ as data in both of the emulators, thus increasing the efficiency of the design.

Our knowledge about some new output, $\eta(\mathbf{x})$, is approximated by a t-distribution, with expectation $\hat{\mathbf{m}}^{**}(\mathbf{x}) = \mathbf{m}_1^{**}(\mathbf{x}_{(1)}) + \mathbf{m}_2^{**}(\mathbf{x}_{(2)}) + c$, and $\hat{\sigma}^2$, and v are found from (4.20) and (4.19).

We demonstrate the method with the simple function

$$\eta(\mathbf{x}) = x_1 + \sin(x_1) + \cos(x_2), \tag{4.28}$$

where x_1 and x_2 are independent $U(-3,3)$. Taking $a_1 = a_2 = 0$, we observe each function at 7 design points. We show design points and outputs in *Table* (4.1).

$x_{(1)}$	$x_{(2)} = a_2$	$y_{(1)}$	$x_{(1)} = a_1$	$x_{(2)}$	$y_{(2)}$
-3	0	-2.14112	0	- 3	-0.98999
-2	0	-1.90930	0	-2	-0.41615
-1	0	-0.84147	0	-1	0.54030
0	0	1	0	0	1
1	0	2.84147	0	1	0.54030
2	0	3.90930	0	2	-0.41615
3	0	4.14112	0	3	-0.98999

Table 4.1: Design points and outputs

We show the design points $x_{(1)}$ plotted against our 7 outputs $y_{(1)i} = \eta(x_{(1)i}, x_{(2)} = a_2)$ in *Figure* (4.2). For comparison we also show the function $\eta_1(x_{(1)})$.

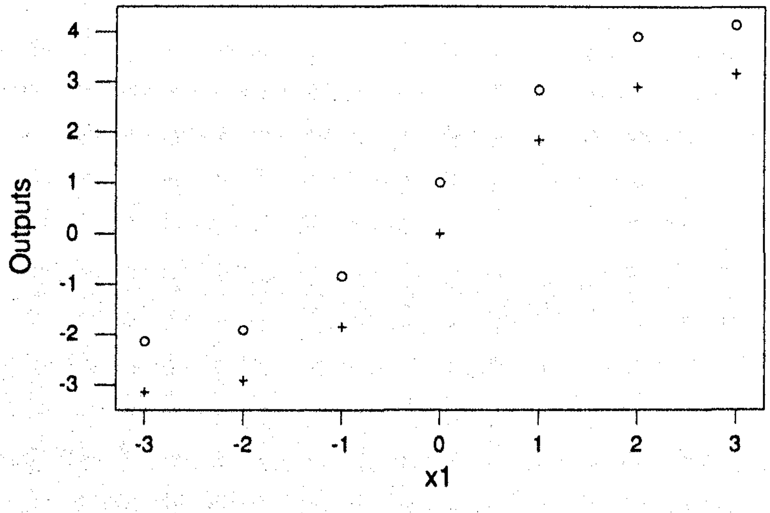

Figure 4.2: Design points and outputs: $y_{(1)i}$ (circles), $\eta_1(x_{(1)})$ (pluses)

We adopt the mean function $\mathbf{h}_1(\mathbf{x}_{(1)}) = (1, x_1)$ and calculate the posterior distribution. In *Figure* (4.3) we show our posterior, $\eta(., x_{(2)} = a_2)|\mathbf{y}_{(1)}, \omega_1$, plotted against $x_{(1)}$. We plot our posterior mean and 99% bounds.

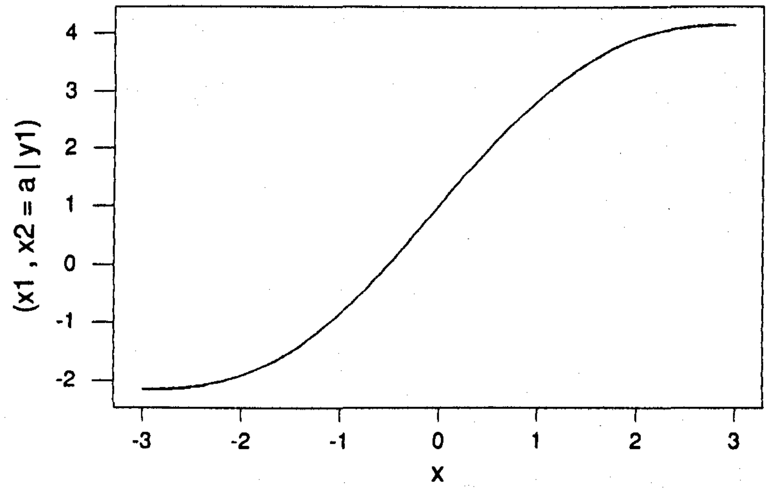

Figure 4.3: Posterior distribution of $\eta(., x_{(2)} = a_2)$

As we see from *Figure* (4.3), our 99% bounds are almost indistinguishable from the posterior mean. We only begin to see separation of these bounds from the posterior mean as $|x| > 3$, which is outside the range of χ.

We find the posterior distribution of $\eta(x_{(1)} = a_1, .)|\mathbf{y}_{(2)}, \omega_2$ similarly. We determine c from $c = \eta(0,0) - \eta_1(0) - \eta_2(0) = 1 - 1 - 1 = -1$.

For comparison we also calculate the posterior distribution of $\eta(.)$ using the methodology we described in chapter 3. We use a 14 point Latin hypercube design. For 100 randomly generated points, generated from $U(-3, 3)$ distributions, we calculate the posterior mean under each model, and compare the results using Root Mean Squared Error, $RMSE = \{100^{-1} \sum_{i=1}^{100} \{\eta(\mathbf{x}_i) - \hat{\eta}(\mathbf{x}_i)\}^2\}^{1/2}$. Our additive model gave $RMSE$ of 0.0866, whilst the Gaussian Process model of chapter 3 gave $RMSE$ of 0.2276.

However, the true value of the additive model is that we can reduce the number of our design points. In this example, the additive model performs better (in terms of $RMSE$) with as few as 9 design points (using design $\{(-3,0),(-1.5,0),(0,0),$ $(1.5,0),(3,0),(0,-3),(0,-1.5),(0,1.5),(0,3)\}$).

4.4 Known Partially Additive Decomposition

We consider a known decomposition of $\eta(\mathbf{x})$ of the form

$$\eta(\mathbf{x}) = \eta_1(\mathbf{x}_{(1)}) + \ldots + \eta_r(\mathbf{x}_{(r)}), \tag{4.29}$$

where the $\eta_j(.)$ are independent functions of sub-vectors $\mathbf{x}_{(j)}$. A partially additive decomposition is more general than the additive decomposition (and includes the additive decomposition as a special case). We still have the condition that $\bigcup_{i=1}^{r} S_i = S$ (providing that all the inputs are active), however we no longer require $S_i \bigcap S_j = \varnothing \forall i \neq j$ – that is input x_k may be present in at least two sub-vectors $\mathbf{x}_{(i)}, \mathbf{x}_{(j)}$. The single requirement we have is that $S_i \not\subseteq S_j \forall i \neq j$.

If we consider the two functions

$$\eta(\mathbf{x}) = \sin(x_1 + x_2 + x_3) + \cos(x_1 + x_2), \tag{4.30}$$

$$\eta(\mathbf{x}) = \sin(x_1 + x_2 + x_3) + \cos(x_1 + x_2 + x_4), \tag{4.31}$$

we see that under this definition (4.31) is partially additive with $S_1 = \{1,2,3\}$ and $S_2 = \{1,2,4\}$, whilst (4.30) is not partially additive.

Under our definition the dimension of each S_i is less than the dimension of S. The decomposition of $\eta(.)$ may be complex; for example r functions with input vectors of dimension $r-1$ and a decomposition with $r >> p$ terms are both consistent with our definition. Our interest lies in decompositions which simplify $\eta(.)$. Therefore, decompositions with $r \leq p$ and where the dimension of each sub-vector is $<< p$ are of interest.

4.4.1 Observable Functions

In this scenario, the computer model provides us with output in such a manner that we are able to make direct observations of the functions, $\eta_1(.), \ldots \eta_r(.)$ respectively. As with the additive case (examined in 4.2.1), we observe $y_{(1)}, \ldots y_{(r)}$, with the property that $\sum_{j=1}^{r} y_j = y$. We want to use this additional structural information in order to improve our cheap approximation to $\eta(.)$.

Since we are able to observe each of these sub functions, we can proceed exactly as we did in 4.2.1. We model each function with a Gaussian process prior, observe data $\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(r)}$ respectively, and after application of Bayes theorem arrive at posterior distributions, $\eta_1(.|\mathbf{y}_{(1)}, \boldsymbol{\omega}_1), \ldots, \eta_r(.|\mathbf{y}_{(r)}, \boldsymbol{\omega}_r)$ respectively.

We want to make inference about $\eta(.)$, and our knowledge about $\eta(.)$ is once more represented by a sum of r t-distributions. We can easily calculate posterior moments of each $\eta_j(.)$ and by independence of the $\eta_j(.)$, we can also calculate posterior moments of $\eta(.)$. Given the expectation, variance and kurtosis, we use a t approximation as in 4.2.1.

4.4.2 Unobservable Functions

For a partially additive decomposition, there is no equivalent model to the one considered in 4.2.2, that is able to exploit a specific design. However, knowledge of the decomposition is useful, and we discuss how to fit a similar model in section 4.6, when we are only able to make observations of $\eta(.)$.

4.5 Structural Uncertainty

In the previous two sections we have exploited known model structure in order to fit additive correlation structures. This correlation structure is more efficient when $\eta(.)$ can be decomposed, as highlighted by our example in 4.2.2. However, if we falsely assume that $\eta(.)$ is decomposable, and our design points do not cover the full input space, we have very little information about interactions between some groups of inputs. We need to ensure that if we have doubts about the structure of $\eta(.)$, our design points should cover the full design space, χ, especially when a single evaluation of the function is computationally expensive. For this reason, we will most likely find that experts are not able to or not willing to decompose $\eta(.)$ with complete certainty.

In this section, we consider the case where we have uncertainty about whether a decomposition of $\eta(.)$ is possible. Our design points are chosen to cover the entire input space, χ. With this structure, we may still fit the Gaussian Process model of chapter 3, but we can also attempt to fit more efficient additive correlation structures. We develop a Gaussian Process model for a decomposition of $\eta(.)$ when we have structural uncertainty. We go on to look at more specific additive and partially additive correlation structures in sections 4.5 and 4.6.

4.5.1 Specification of a Prior Distribution

We suspect that we can decompose $\eta(.)$ into functions of lower dimensional input vectors. However, we have uncertainty about this decomposition. Supposing that we have correctly identified the decomposition, we can write $\eta(.)$ as

$$\eta(.) = \eta_1(.) + \ldots \eta_r(.). \tag{4.32}$$

We are able to make observations of $\eta(.)$, but due to the configuration of the design points, we cannot model $\eta_1(.), \dots \eta_r(.)$ using the hierarchical structure we described in 4.2.2. However, it is possible to model the terms $\eta_1(.), \dots \eta_r(.)$ individually even with a space filling design, although this is more problematic.

We begin by specifying our prior beliefs about $\eta(.)$. The expectation and variance of $\eta(\mathbf{x})$, conditional on regression hyperparameters, $\boldsymbol{\beta}$, variance hyperparameters, $\sigma_1^2, \dots \sigma_r^2$, and smoothness parameters, $\boldsymbol{\omega}_1, \dots \boldsymbol{\omega}_r$, where $\boldsymbol{\omega}_j$ denotes the parameters of the j^{th} correlation function, corresponding to term $\eta_j(.)$ in our decomposition, are given by

$$\begin{aligned} E[\eta(\mathbf{x})|\boldsymbol{\beta}, \sigma_1^2, \dots, \sigma_r^2, \boldsymbol{\omega}_1, \dots \boldsymbol{\omega}_r] &= \mathbf{h}(\mathbf{x})^{\mathbf{T}}\boldsymbol{\beta}, && (4.33) \\ Var[\eta(\mathbf{x})|\boldsymbol{\beta}, \sigma_1^2, \dots, \sigma_r^2, \boldsymbol{\omega}_1, \dots \boldsymbol{\omega}_r] &= \sigma_1^2 + \dots \sigma_r^2. && (4.34) \end{aligned}$$

We assume independence of the terms $\eta_1(.), \dots \eta_r(.)$.

We define the covariance between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ as

$$Cov[\eta(\mathbf{x}) \, , \, \eta(\mathbf{x}')] = \mathbf{d}(\mathbf{x}, \mathbf{x}') = \sigma_1^2 \mathbf{c}_1(\mathbf{x}_{(1)}, \mathbf{x}'_{(1)}) + \dots + \sigma_r^2 \mathbf{c}_\mathbf{r}(\mathbf{x}_{(r)}, \mathbf{x}'_{(r)}), \qquad (4.35)$$

where $\mathbf{x}_{(j)}$ denotes a sub-vector of $\mathbf{x}$, and $\mathbf{c}_\mathbf{j}(.,.)$ for $j = 1, \dots, r$ are correlation functions, with function, $\mathbf{c}_\mathbf{j}(.,.)$, corresponding to term $\eta_j(.)$ in (4.32)

The covariance between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ is $\mathbf{d}(\mathbf{x}, \mathbf{x}')$. This is a weighted sum (weighted by the σ_j^2) of correlation functions. We have no cross products in (4.35) due to the assumed independence of $\eta_1(.), \dots \eta_r(.)$.

We combine our prior beliefs about expectation, variance and covariance, as expressed in (4.33)-(4.35), with distributional assumptions similar to those of chapter 3.

We specify the Gaussian Process prior,

$$\eta(.)\,|\,\boldsymbol{\beta},\, \sigma_1^2,\ldots,\sigma_r^2,\boldsymbol{\omega}_1,\ldots\boldsymbol{\omega}_r \sim GP(\,\mathbf{h}(.)^{\mathbf{T}}\boldsymbol{\beta},\mathbf{d}(.,.)\,). \tag{4.36}$$

The final element of our prior specification requires us to consider our beliefs about the hyperparameters . Previously we considered a conjugate Normal Inverse Gamma prior on $\boldsymbol{\beta}$ and σ^2 however there is no natural multivariate extension for more than one variance parameter. Our beliefs about $\boldsymbol{\beta}$ and the variances, σ_j^2, are likely to be weak in any case. We adopt the non informative prior

$$p(\boldsymbol{\beta},\sigma_1^2,\ldots,\sigma_r^2) \propto \sigma_1^{-2} \times \ldots \times \sigma_r^{-2}, \tag{4.37}$$

on $\boldsymbol{\beta}$ and $\sigma_1^2,\ldots,\sigma_r^2$ and adopt independent improper uniform priors on the elements of parameter vectors $\boldsymbol{\omega}_1,\ldots\boldsymbol{\omega}_r$.

4.5.2 Prior to Posterior Analysis

Suppose that we are able to make n runs of the expensive computer code. We obtain the data vector, $\mathbf{y} = \{\eta(\mathbf{x}_1),\eta(\mathbf{x}_2),\ldots,\eta(\mathbf{x}_n)\}$, at inputs $\{\mathbf{x}_1,\mathbf{x}_2\ldots,\mathbf{x}_n\}$. We wish to update our beliefs about $\eta(.)$ in light of data, $\mathbf{y}$.

The likelihood is written as

$$f(\mathbf{y}\,|\,\boldsymbol{\beta},\sigma_1^2,\ldots,\sigma_r^2,\omega_1,\ldots\omega_r) = \frac{|\mathbf{A}^*|^{-1/2}}{(2\pi)^{n/2}}\exp\{-1/2(\mathbf{y}-\mathbf{H}\boldsymbol{\beta})^{\mathbf{T}}\mathbf{A}^{*-1}(\mathbf{y}-\mathbf{H}\boldsymbol{\beta})\}, \tag{4.38}$$

where

$$\begin{aligned}\mathbf{H} &= \{\mathbf{h}(\mathbf{x}_1)\ldots\mathbf{h}(\mathbf{x}_n)\}^T,\\ \mathbf{A}^* &= \sigma_1^2\mathbf{A}_1+\ldots+\sigma_r^2\mathbf{A}_r,\end{aligned}$$

and

$$\mathbf{A_j} = \begin{pmatrix} \mathbf{c_j}(\mathbf{x}_{(j),1}, \mathbf{x}_{(j),1}) & \mathbf{c_j}(\mathbf{x}_{(j),1}, \mathbf{x}_{(J),2}) & \cdots & \mathbf{c_j}(\mathbf{x}_{(j),1}, \mathbf{x}_{(j),k}) & \cdots & \mathbf{c_j}(\mathbf{x}_{(j),1}, \mathbf{x}_{(j),n}) \\ \mathbf{c_j}(\mathbf{x}_{(j),2}, \mathbf{x}_{(j),1}) & \mathbf{c_j}(\mathbf{x}_{(j),2}, \mathbf{x}_{(j),2}) & & & & \\ \vdots & & \ddots & & & \\ & & \cdots & \mathbf{c_j}(\mathbf{x}_{(j),i}, \mathbf{x}_{(j),k}) & \cdots & \\ \mathbf{c_j}(\mathbf{x}_{(j),n}, \mathbf{x}_{(j),1}) & \cdots & & & & \mathbf{c_j}(\mathbf{x}_{(j),n}, \mathbf{x}_{(j),n}) \end{pmatrix},$$

denotes the $n \times n$ correlation matrix for term $\eta_j(.)$ of the decomposition.

We begin by finding the joint posterior of $\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_r^2$ and $\boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r$, which after application of Bayes theorem gives us

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_r^2, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r \,|\, \mathbf{y}) &\propto f(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_r^2, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r) \\ &\times f(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_r^2, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r). \qquad (4.39) \end{aligned}$$

We can partition the posterior as

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_r^2, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r \,|\, \mathbf{y}) &= f(\boldsymbol{\beta} \,|\, \sigma_1^2, \ldots, \sigma_r^2, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r, \mathbf{y}) \\ &\times f(\sigma_1^2, \ldots, \sigma_r^2, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r \,|\, \mathbf{y}), \qquad (4.40) \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\beta} \,|\, \sigma_1^2, \ldots, \sigma_r^2, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r, \mathbf{y} &\sim N(\hat{\boldsymbol{\beta}}^{(1)}, (\mathbf{H}\mathbf{A}^{*-1}\mathbf{A})^{-1}), \qquad (4.41) \\ f(\sigma_1^2, \ldots, \sigma_r^2, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r \,|\, \mathbf{y}) &\propto \sigma_1^{-2} \ldots \sigma_r^{-2} \times |\mathbf{A}^*|^{-1/2} |\mathbf{H}\mathbf{A}^{*-1}\mathbf{A}|^{-1/2} \\ &\times \exp\{-\frac{1}{2}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}^{(1)})^T \mathbf{A}^{*-1} (\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}^{(1)})\} \qquad (4.42) \end{aligned}$$

and

$$\hat{\boldsymbol{\beta}}^{(1)} = (\mathbf{H}^\mathbf{T}\mathbf{A}^{*-1}\mathbf{H})^{-1}\mathbf{H}^\mathbf{T}\mathbf{A}^{*-1}\mathbf{y}, \qquad (4.43)$$

where we adopt superscript notation to distinguish between the current calculations and those of chapter 3.

We now update our beliefs about $\eta(.)$ in light of the data. Similar to chapter 3, the function, $\eta(.)\,|\,\mathbf{y}, \boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_r^2, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r$, is a Gaussian Process:

$$\eta(.)\,|\,\mathbf{y}, \boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_r^2, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r \sim GP(\mathbf{m}^{(1)*}(.), \mathbf{d}^*(.,.)), \tag{4.44}$$

where

$$\begin{aligned} \mathbf{m}^{(1)*}(\mathbf{x}) &= \mathbf{h}(\mathbf{x})^{\mathbf{T}}\boldsymbol{\beta} + \mathbf{t}^{(1)}(\mathbf{x})^{\mathbf{T}}\mathbf{A}^{*-1}(\mathbf{y} - \mathbf{H}\boldsymbol{\beta}), & (4.45)\\ \mathbf{d}^*(\mathbf{x}, \mathbf{x}') &= \mathbf{d}(\mathbf{x}, \mathbf{x}') - \mathbf{t}^{(1)}(\mathbf{x})^{\mathbf{T}}\mathbf{A}^{*-1}\mathbf{t}^{(1)}(\mathbf{x}'), & (4.46)\\ \mathbf{t}^{(1)}(\mathbf{x}) &= \{\mathbf{d}(\mathbf{x}, \mathbf{x}_1), \ldots, \mathbf{d}(\mathbf{x}, \mathbf{x}_n)\}. & (4.47) \end{aligned}$$

Taking the product of (4.41) and (4.44) and integrating over $\boldsymbol{\beta}$ leaves us with the Gaussian Process

$$\eta(.)\,|\,\mathbf{y}, \sigma_1^2, \ldots, \sigma_r^2, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r \sim GP(\mathbf{m}^{(1)**}(.), \mathbf{d}^{**}(.,.)), \tag{4.48}$$

where

$$\begin{aligned} \mathbf{m}^{(1)**}(\mathbf{x}) &= \mathbf{h}(\mathbf{x})^{\mathbf{T}}\hat{\boldsymbol{\beta}}^{(1)} + \mathbf{t}^{(1)}(\mathbf{x})^{\mathbf{T}}\mathbf{A}^{*-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}^{(1)}), & (4.49)\\ \mathbf{d}^{**}(\mathbf{x}, \mathbf{x}') &= \mathbf{d}^*(\mathbf{x}, \mathbf{x}') + (\mathbf{h}(\mathbf{x})^{\mathbf{T}} - \mathbf{t}^{(1)}(\mathbf{x})^{\mathbf{T}}\mathbf{A}^{*-1}\mathbf{H}) & \\ &\times\ (\mathbf{H}^{\mathbf{T}}\mathbf{A}^{*-1}\mathbf{H})^{-1}(\mathbf{h}(\mathbf{x}')^{\mathbf{T}} - \mathbf{t}^{(1)}(\mathbf{x}')^{\mathbf{T}}\mathbf{A}^{*-1}\mathbf{H})^{T}. & (4.50) \end{aligned}$$

Since we cannot separate the unknown variance parameters from the correlation matrices, it is not possible to remove the conditioning on the variances analytically. We will require numerical methods in order to find $\eta(.)|\mathbf{y}$.

The simplification we adopted in chapter 3 was to estimate the unknown parameters by the posterior mode of their joint distribution. We could apply the same methodology here, estimating $\sigma_1^2, \ldots, \sigma_r^2, \omega_1, \ldots \omega_r$ by the posterior mode of (4.42), and treating these as known. Test problems have shown this approximation is adequate if we just require a point estimate for $\eta(\mathbf{x})$, even though we have substantial uncertainty in both $\mathbf{t}^{(1)}(\mathbf{x})^{\mathbf{T}}$ and $\mathbf{A}^{*-1}$. However, we inevitably underestimate the uncertainty surrounding our point estimate. In chapter 3 we claimed that we only ignored 'second order uncertainty' with our approximation, but our expression (4.42) contains variances, a measure of 'first order uncertainty', so we cannot make the same claim.

For this more difficult problem, we should take all the uncertainties into account. We could sample from the distribution of $\eta(.)|\mathbf{y}$ using MCMC, similar to *Neal*(45) and *Bayarri et al.*(4).

4.5.3 Equal Variances

A special case, and the simplest case we could encounter, is where the variances $\sigma_1^2 = \ldots = \sigma_r^2 = \sigma^2$, that is a we have the same variance for each term in the decomposition (4.32).

Then (4.33)-(4.34) are replaced by

$$E[\eta(\mathbf{x})|\boldsymbol{\beta}, \sigma^2, \omega_1, \ldots \omega_r] = \mathbf{h}(\mathbf{x})^{\mathbf{T}}\boldsymbol{\beta}, \tag{4.51}$$

$$Var[\eta(\mathbf{x})|\boldsymbol{\beta}, \sigma^2, \omega_1, \ldots \omega_r] = r\sigma^2, \tag{4.52}$$

and the covariance is defined as

$$Cov[\eta(\mathbf{x}), \eta(\mathbf{x}')] = \sigma^2\mathbf{c}^{(2)}(\mathbf{x}, \mathbf{x}') = \sigma^2\{\mathbf{c}_1(\mathbf{x}_{(1)}, \mathbf{x}'_{(1)}) + \ldots + \mathbf{c}_\mathbf{r}(\mathbf{x}_{(r)}, \mathbf{x}'_{(r)})\}, \tag{4.53}$$

where the superscript notation is used again to help us distinguish between models.

We combine our prior beliefs about expectation, variance and covariance, as given by (4.51)-(4.53), with distributional assumptions in a Gaussian Process prior as before:

$$\eta(.) \,|\, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r \sim GP(\mathbf{h}(.)^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2 \mathbf{c}^{(2)}(.,.)). \tag{4.54}$$

The final stage of our prior specification is the prior distributions for the model hyperparameters. The simplification of equal variances allows us to place a conjugate prior on $\boldsymbol{\beta}$ and σ^2, although we choose the non-informative prior, $p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$. We still adopt independent, improper uniform priors on the elements of the $\boldsymbol{\omega}_j$.

After observing data, and proceeding as in chapter 3, we arrive at a posterior student process on $\eta(.)|\mathbf{y}, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r$, which for a given $\mathbf{x}$ is written as

$$\frac{\eta(\mathbf{x}) - \mathbf{m}^{(2)**}(\mathbf{x})}{\hat{\sigma}^{(2)}\sqrt{\frac{n-q-2}{n-q}\mathbf{c}^{(2)**}(\mathbf{x},\mathbf{x})}} \,|\, \mathbf{y}, \boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r \sim t_{n-q}, \tag{4.55}$$

where,

$$\begin{aligned}
\mathbf{m}^{(2)**}(\mathbf{x}) &= \mathbf{h}(\mathbf{x})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(2)} + \mathbf{t}^{(2)}(\mathbf{x})^{\mathrm{T}}\mathbf{A}^{**-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}^{(2)}), && (4.56)\\
\mathbf{c}^{(2)**}(\mathbf{x},\mathbf{x}') &= \mathbf{c}^{(2)}(\mathbf{x},\mathbf{x}') - \mathbf{t}^{(2)}(\mathbf{x})^{\mathrm{T}}\mathbf{A}^{**-1}\mathbf{t}^{(2)}(\mathbf{x}') + (\mathbf{h}(\mathbf{x})^{\mathrm{T}} - \mathbf{t}^{(2)}(\mathbf{x})^{\mathrm{T}}\mathbf{A}^{**-1}\mathbf{H})\\
&\quad \times (\mathbf{H}^{\mathrm{T}}\mathbf{A}^{**-1}\mathbf{H})^{-1}(\mathbf{h}(\mathbf{x}')^{\mathrm{T}} - \mathbf{t}^{(2)}(\mathbf{x}')^{\mathrm{T}}\mathbf{A}^{**-1}\mathbf{H})^{T}, && (4.57)\\
\mathbf{t}^{(2)}(\mathbf{x}) &= \{\mathbf{c}^{(2)}(\mathbf{x},\mathbf{x}_1), \ldots, \mathbf{c}^{(2)}(\mathbf{x},\mathbf{x}_n)\}, && (4.58)\\
\mathbf{A}^{**} &= \mathbf{A}_1 + \ldots + \mathbf{A}_{\mathbf{r}}, && (4.59)\\
\hat{\boldsymbol{\beta}}^{(2)} &= (\mathbf{H}^{\mathrm{T}}\mathbf{A}^{**-1}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{H}\mathbf{A}^{**-1}\mathbf{y}, && (4.60)\\
\hat{\sigma}^{(2)2} &= \frac{\mathbf{y}^{\mathrm{T}}(\mathbf{A}^{**-1} - \mathbf{A}^{**-1}\mathbf{H}(\mathbf{H}^{\mathrm{T}}\mathbf{A}^{**-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{A}^{**-1})\mathbf{y}}{n-q-2}. && (4.61)
\end{aligned}$$

We adopt the simplification proposed in chapter 3, and ignore the ‘second order uncertainty', by estimating $\omega_1, \ldots \omega_r$ from their joint posterior mode,

$$f(\boldsymbol{\omega}_\mathbf{1}, \ldots, \boldsymbol{\omega}_\mathbf{r} \,|\, \mathbf{y}) = \hat{\sigma}^{((2)-(n-q))} |\mathbf{A}^{**}|^{-1/2} |\mathbf{H}^\mathbf{T} \mathbf{A}^{**-1} \mathbf{H}|^{-1/2}, \tag{4.62}$$

and treating them as known.

4.5.4 Model Comparison

We have an obvious advantage in the equal variances formulation of 4.4.3, since we are able to avoid numerical methods. We can also consider a simple generalization to variances $a_1\sigma^2, \ldots a_r\sigma^2$ for known weights a_j, with minimal modification to the theory of 4.4.3.

However, we need to know what loss of information we suffer if we apply the methodology of 4.4.3 to any problem. At first sight, it seems rather naive to assume we have equal variances, especially when the dimensions of the $\mathbf{x}_{(j)}$ may differ significantly. However, by examining variances and correlation functions, we can show this assumption is not so unreasonable after all.

The variances of our two models are $\sum_{j=1}^r \sigma_j^2$ and $r\sigma^2$. These describe how far $\eta(.)$ departs from our parametric approximation, $\mathbf{h}(.)^\mathbf{T}\boldsymbol{\beta}$. In this sense there is clearly no advantage in unequal variances.

The correlation functions can be written as

$$\begin{aligned} \frac{1}{\sum_{j=1}^r \sigma_j^2} \mathrm{d}(\mathbf{x}, \mathbf{x}') &= \frac{\sigma_1^2}{\sum_{j=1}^r \sigma_j^2} \mathrm{c}_1(.,.) + \ldots \frac{\sigma_r^2}{\sum_{j=1}^r \sigma_j^2} \mathrm{c}_\mathrm{r}(.,.), & (4.63) \\ \frac{1}{r} \mathrm{c}^{(2)}(\mathbf{x}, \mathbf{x}') &= \frac{1}{r} \{\mathrm{c}_1(.,.) + \ldots \mathrm{c}_\mathrm{r}(.,.)\}, & (4.64) \end{aligned}$$

respectively.

In (4.63) the variances act as weights for the respective correlations functions, giving the illusion of much more flexibility. Of course, as in most statistical models, the addition of parameters improves the performance, but since the correlation functions also contain parameters there is a substantial overlap in what $\mathbf{c}_j(.,.)$ and σ_j^2 are estimating. The variances and smoothing parameters in (4.63) are in some sense competing to model the same source of uncertainty. As a result, the additional variance parameters within (4.63) offer only marginal improvement over (4.64). As the size of our sub vectors increases, and hence the number of parameters in our correlations functions increases, the effect of additional variance parameters is further diminished. One notable exception when the flexibility of (4.63) is desirable is when the contribution from one of our $\mathbf{c}_j(.,.)$ is zero. In this special case, when using (4.64) we have no way of setting $\mathbf{c}_j(.,.)$ equal to zero $\forall \mathbf{x}_{(j)}, \mathbf{x}'_{(j)}$. We discuss this further later on in this chapter.

We compare the two correlation functions, (4.63) and (4.64), using two examples, which are chosen to demonstrate very different behaviors. We take $\eta(.)$ to be a function of x_1 and x_2 in both cases but discuss higher order functions later. In the first of these, the output is a smooth function of both inputs. We use

$$\eta(\mathbf{x}) = x_1 + \sin(x_1) + \cos(x_2), \tag{4.65}$$

which we observe at 14 design points, selected using a Latin hypercube design. We estimate the unknown parameters from the respective posterior modes (4.42) and (4.62). We now consider the correlation as a function of distances $d_1 = x_1 - x'_1$ and $d_2 = x_2 - x'_2$. We show plots of the correlation functions (4.63) and (4.64) in *Figure* (4.4) and *Figure* (4.5) respectively. We plot the difference between the two correlation functions in *Figure* (4.6).

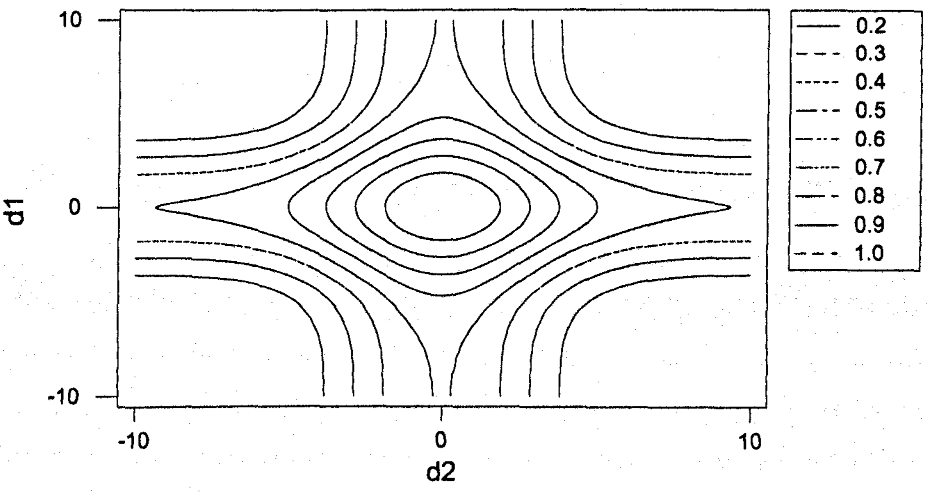

Figure 4.4: Contour plot: unequal variances

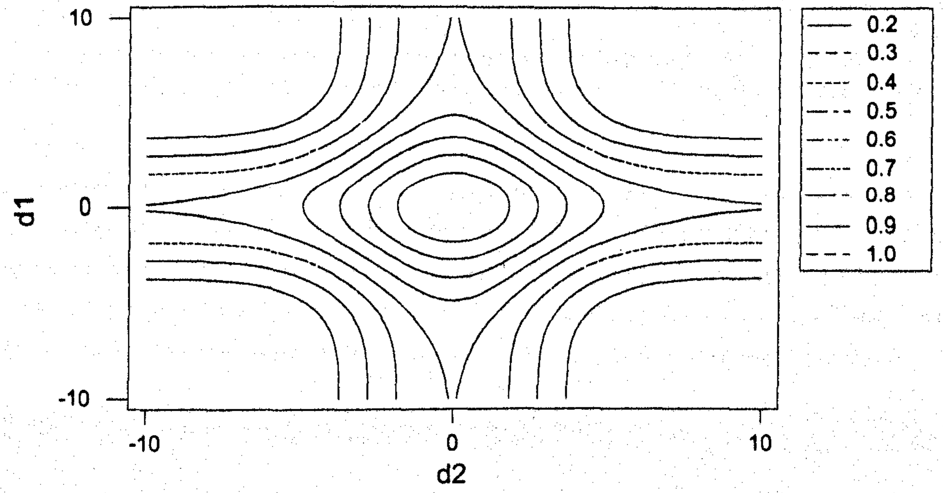

Figure 4.5: Contour plot: equal variances

As we see from *Figure* (4.4) and *Figure* (4.5), the correlation functions have a similar shape as a function of distance. For similar correlation functions we should find the difference is $\approx 0 \forall d_1, d_2$. We see from *Figure* (4.6) that the difference between the two functions is only significantly different from zero in the tails. This is of little concern since the power of our approach is in the large correlations.

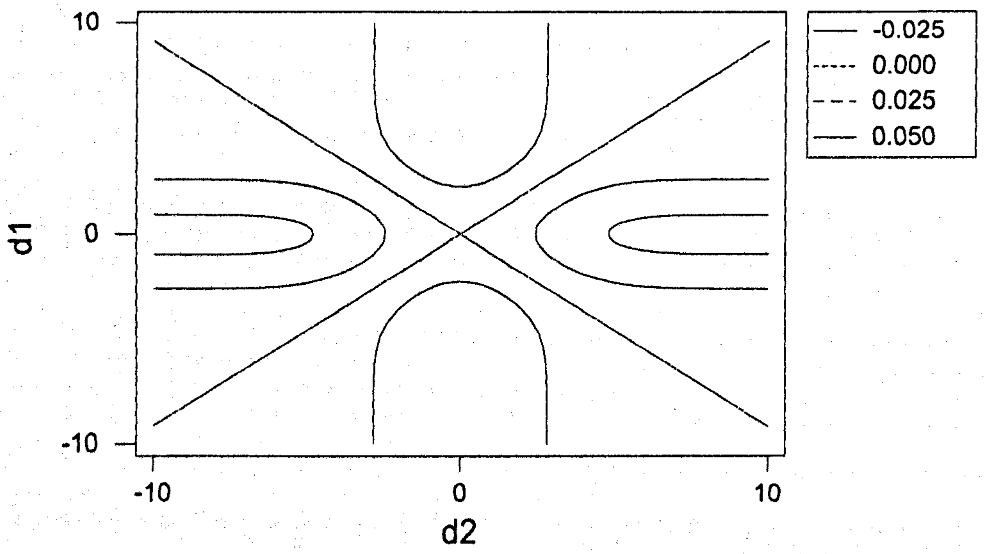

Figure 4.6: Contour plot: difference

In our second example we consider the case when the output is far more sensitive to changes in x_2 than changes in x_1. This represents a case where we might expect multiple variances to offer greater flexibility. We use the example

$$\eta(\mathbf{x}) = x_1 + \sin(x_1) + \cos(3x_2), \tag{4.66}$$

which we observe at the same 14 design points. We show plots of the correlation functions (4.63) and (4.64) in *Figure* (4.7) and *Figure* (4.8) respectively, and plot the difference between the two correlation functions in *Figure* (4.9). We take d_1 over the range -10 to 10 and d_2 over the range -1 to 1.

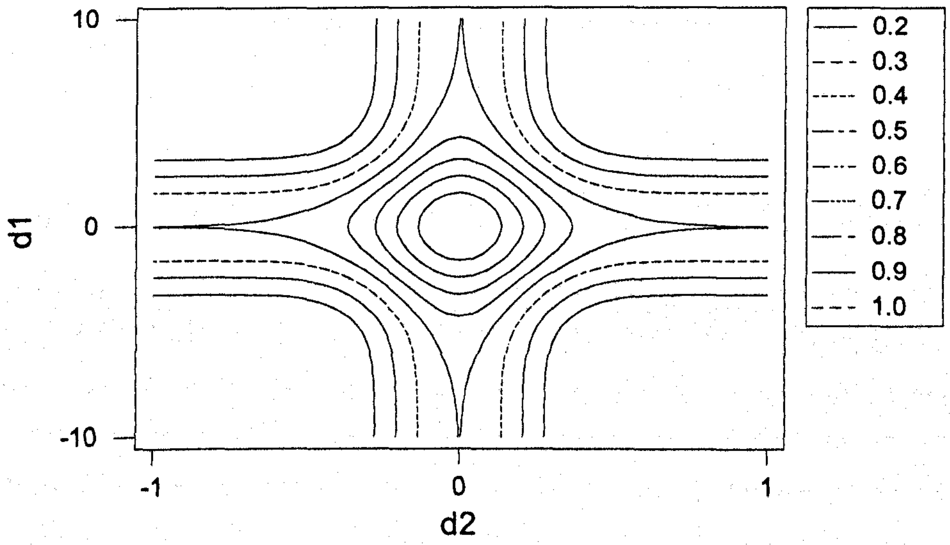

Figure 4.7: Contour plot: unequal variances

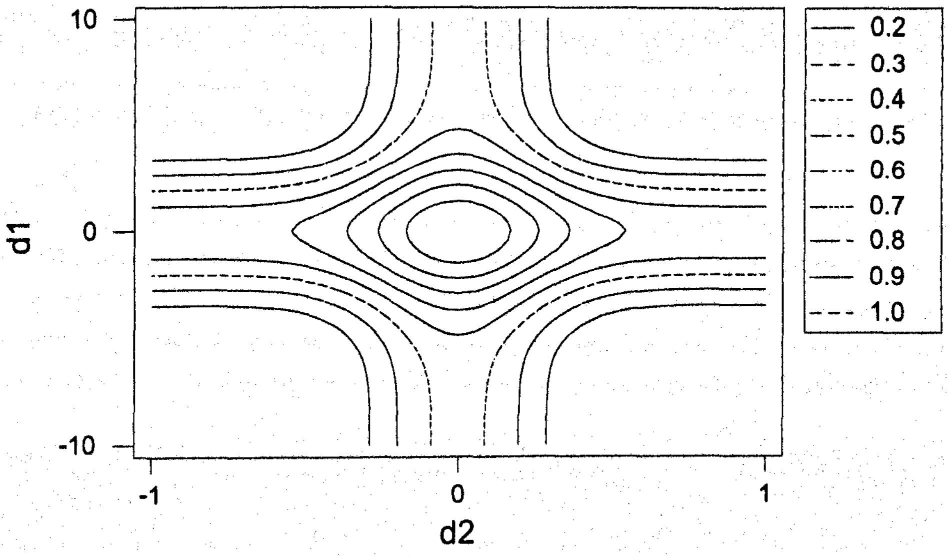

Figure 4.8: Contour plot: equal variances

We note from *Figure* (4.7) and *Figure* (4.8) the difference between the two functions, which is far more apparent than in our previous example. *Figure* (4.9) shows this difference more clearly, and we note the two functions diverge as we move away from $d_1, d_2 = 0$.

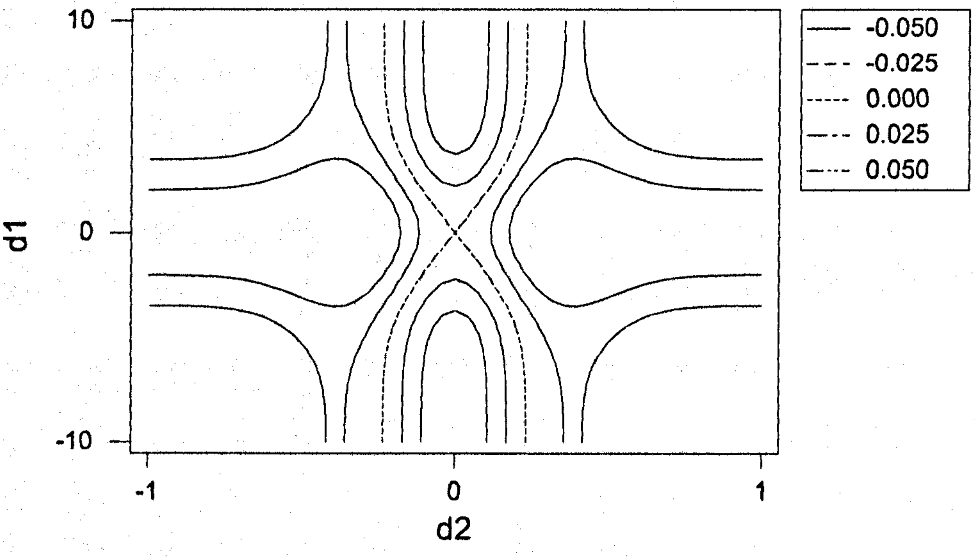

Figure 4.9: Contour plot: difference

In both examples, relatively small changes in the parameters of (4.63) resulted in non trivial changes in the shape of the correlation function. Despite this, we still ran into difficulties in maximizing the posterior, (4.42), which was flat over a large area surrounding the mode. Further investigation revealed that the posterior was very flat over regions where the product $\sigma_i^2 \omega_i$ remains constant, even though changes in w_i and corresponding changes in σ_i^2 resulted in substantial changes to the correlation function in the tails.

The significance of this is apparent when we consider the power series representation of the exponential function. In our examples the correlation, using

(4.63), may be written as

$$\frac{1}{\sum_{i=1}^{2}\sigma_i^2}\sum_{i=1}^{2}\sigma_i^2\{1-\frac{\omega_i d_i^2}{1!}+\frac{\omega_i^2 d_i^4}{2!}-\frac{\omega_i^3 d_i^6}{3!}+\ldots\}, \quad (4.67)$$

where we note that the terms in the expansion (4.67) are functions of d_i^2

Ignoring the constant term, the first function of distance d_i in $\exp\{-\omega_i d_i^2\}$, is $\omega_i d_i^2$ and it is this term that dominates the correlation function for d_i close to zero. We cannot be more precise than 'close to', since this depends on the magnitude of ω_i – that is how quickly the exponential decays toward zero. The large correlations, where the power of the Gaussian Process model lies are dominated by this term. It is only as we move away from $d_i = 0$ to smaller correlations that the higher order terms in the power series expansion of $\exp\{-\omega_i d_i^2\}$ begin to have a larger influence on the correlation. However, provided that we have enough design points in order to model the large correlations well, the smaller correlations have a relatively small effect on the performance of the Gaussian Process model. Resultantly, the data are unable to easily distinguish between correlation functions that model the smaller correlations differently, hence the difficulty in maximizing (4.42). When using (4.64), it is possible to select the parameters such that (4.63) and (4.64) are identical up to the first order term. The higher order terms may of course differ substantially.

For higher order problems it is difficult to visualize the correlation as a function of distance, so we cannot easily verify if (4.63) and (4.64) produce similar correlations for a larger sum of 1 dimensional correlation functions. However, our numerical work has identified flat posterior distributions – an indicator of over-parametrization, when using (4.63). When each correlation function is of dimension $k > 1$ and can be written as the product $\prod_{i=1}^{k}\exp\{-w_i d_i^2\}$, by expanding each term as in (4.67), we note the higher order terms in the expansion have even less importance than in the $k = 1$ case.

In the following two sections we develop methodology for searching for additive and partially additive decompositions. For this work we assume equal variances.

4.6 Unknown Additive Decomposition

In this section we consider a similar situation to that of section 4.2.2. We suspect that we have mutually exclusive and exhaustive partition of $\mathbf{x} = \{\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(r)}\}$, and a corresponding decomposition of the output $\eta(\mathbf{x})$;

$$\eta(\mathbf{x}) = \eta_1(\mathbf{x}_{(1)}) + \ldots + \eta_r(\mathbf{x}_{(r)}). \qquad (4.68)$$

We may suspect that we know all of, some of, or none of the subsets S_i. In light of our uncertainty, we have chosen design points in order to cover χ, as described in 4.5.

Suppose we want to test if $\eta(\mathbf{x})$ can be decomposed as in (4.68). Presuming that the function is additive, and we have correctly identified the decomposition, we define the prior expectation, variance and covariance as in section 4.5.3, and specify the Gaussian Process prior (4.54). We update our beliefs in light of the data, $\mathbf{y}$, arriving at a posterior student process as described in the previous section. However, if the partition is erroneous, it is possible to model $\eta(.)$ using the Gaussian Process prior (3.5) from the previous chapter, updating our beliefs as described in section 3.2.2.

Our problem is to determine which of the two correlation structures (and as a result which model) we should use. If we have found the correct decomposition of $\eta(.)$, then the additive structure will predict better, whilst the standard correlation structure will perform far better if our decomposition of $\eta(.)$ is erroneous.

Choosing the most appropriate correlation structure, given data, $\mathbf{y}$, requires a novel approach. Standard methods of model comparison are of little use since both of our models will interpolate the data exactly. We could conclusively verify a proposed decomposition by making more observations however this may be impractical, especially for a computationally expensive function. We consider two approaches here, that use just the observations, $\mathbf{y}$.

4.6.1 Cross Validation

Cross validation, where we leave each design point, $\mathbf{x}_i$, out in turn, and predict $\eta(\mathbf{x}_i)$ using the remaining designs point, is a useful tool for detecting if the additive model is inadequate. When the additive model is inadequate the cross validation prediction errors, $\eta(\mathbf{x}_i)-\hat{\eta}(\mathbf{x}_i)$, where $\hat{\eta}(\mathbf{x}_i)$ denotes our prediction at $\mathbf{x}_i$ using the remaining $n-1$ design points, will in general exhibit some structure. However, this test may not be able to distinguish between the cases where we have a very small interaction between inputs, and additivity.

In the cases when cross validation can identify the decomposition (4.68) is incorrect, it does not indicate how the choice of subsets S_i for $i=1,\dots,r$ is incorrect. We will not know which terms in the decomposition are incorrect, and which inputs we should have included/omitted from these terms.

4.6.2 Regression Based Model Comparison

The approach to model comparison that we propose is based upon the perfectly fitting prior mean function that we described in 4.2. To briefly recap, we considered the model

$$\eta(\mathbf{x}) = \mathbf{h}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{x}), \tag{4.69}$$

where $\mathbf{Z}(.)$ is a Gaussian Process with zero mean and covariance $\sigma^2\mathbf{c}(.,.)$. Using the product form (4.3) to express the correlation, we noted that the parameters of the i^{th} term in the product were a measure of how smooth departures from the mean function (which takes a parametric form) were in dimension i. In particular, if the mean function fits the data perfectly in dimension i, then $c(x_i, x_i') = 1 \forall x_i, x_i'$.

Using the correlation function

$$\mathbf{c}(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{p} \exp\{-\omega_i(x_i - x_i')^2\}, \tag{4.70}$$

this special case is identified by $\omega_i = 0$.

Suppose that we believe $\eta(\mathbf{x})$ can be written as

$$\eta(\mathbf{x}) = \eta_{S_A}(\mathbf{x}_{S_A}) + \eta_{S-S_A}(\mathbf{x}_{S-S_A}), \tag{4.71}$$

that is a function of the inputs contained in subset S_A, and a function of all the remaining inputs (denoted by $S - S_A$).

Our approach to model comparison involves replacing $\mathbf{h}(\mathbf{x})^{\mathbf{T}}\boldsymbol{\beta}$ in (4.69) by a term specifically introduced to model $\eta_{S_A}(\mathbf{x}_{S_A})$.

We write $\eta(\mathbf{x})$ as

$$\eta(\mathbf{x}) = \mu(\mathbf{x}) + \mathbf{Z}(\mathbf{x}), \tag{4.72}$$

where $\mu(.)$ is our prior mean function, and $\mathbf{Z}(.)$ is a Gaussian Process with zero mean and covariance, $\sigma^2\mathbf{c}_S(.,.)$. The correlation function, $\mathbf{c}_S(.,.)$, is a function of all inputs.

Previously we have only considered parametric forms for our prior mean function however we now adopt a non-parametric form. We model $\mu(.)$ with a Gaussian

Process prior. The prior expectation of $\mu(.)$ is $\mathbf{h}(.)^{\mathrm{T}}\boldsymbol{\beta}$, and the prior covariance is $\sigma^2 \mathbf{c}_{S_A}(.,.)$, where $\mathbf{c}_{S_A}(.,.)$ is a function of just the dimensions in the set S_A.

Using properties of normal distributions, and assuming independence of the correlations, the sum (4.72) is a Gaussian Process with expectation $\mathbf{h}(.)^{\mathrm{T}}\boldsymbol{\beta}$ and covariance

$$\sigma^2 \mathbf{c}^{(2)}(.,.) = \sigma^2 \{\mathbf{c}_{S_A}(.,.) + \mathbf{c}_S(.,.)\}, \tag{4.73}$$

which we can see is a special case of (4.53).

Our first correlation function depends on sub-vector $\mathbf{x}_{S_A}$, with parameter vector $\boldsymbol{\omega}_{S_A}$. Our second correlation function depends on the full vector of inputs, $\mathbf{x}$, with p dimensional parameter vector $\boldsymbol{\omega}_S$. Taking S_A to contain $1, 2, \ldots d$, where $d < p$, and using our exponential correlation function, we write the covariance as

$$\sigma^2 \mathbf{c}^{(2)}(\mathbf{x}, \mathbf{x}') = \sigma^2 \{\prod_{i=1}^{d} \exp[\omega_{S_A i}(x_i - x_i')^2] + \prod_{i=1}^{p} \exp[\omega_{S i}(x_i - x_i')^2]\}. \tag{4.74}$$

We observe data, $\mathbf{y}$, and follow the prior to posterior analysis of 4.5.3, arriving at a posterior student process. Our posterior is conditional on data $\mathbf{y}$ and parameter vectors $\boldsymbol{\omega}_{S_A}$ and $\boldsymbol{\omega}_S$. We find the joint posterior of $\boldsymbol{\omega}_{S_A}$ and $\boldsymbol{\omega}_S$ as

$$f(\boldsymbol{\omega}_{S_A}, \boldsymbol{\omega}_S \,|\, \mathbf{y}) = \hat{\sigma}^{((2)-(n-q))} |\mathbf{A}^{**}|^{-1/2} |\mathbf{H}^{\mathrm{T}} \mathbf{A}^{**-1} \mathbf{H}|^{-1/2}, \tag{4.75}$$

and estimate $\boldsymbol{\omega}_{S_A}$ and $\boldsymbol{\omega}_S$ from the posterior mode.

Any mixture of correlation functions of the form (4.73) will always interpolate the data exactly however the posterior (4.75) will reflect that some structures fit the data, $\mathbf{y}$, better than others.

If the parametric component of the model, $\mathbf{h}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta}$, fits input x_i exactly we have a very similar interpretation to the case we discussed in detail in section 4.2.

We find the corresponding elements of $\hat{\omega}_{S_A}$ (which only exists if $i \leq d$) and $\hat{\omega}_S$ are zero. The special structure of (4.74) means we can use a similar result in order to identify additive groups. If the first d elements of $\hat{\omega}_S$ are zero and $\hat{\omega}_{S_A} \neq \mathbf{0}$, then (4.74) reduces to

$$\sigma^2 \mathbf{c}^{(2)}(\mathbf{x}, \mathbf{x}') = \sigma^2 \{\prod_{i=1}^{d} \exp[\omega_{S_{A_i}}(x_i - x_i')^2] + \prod_{i=d+1}^{p} \exp[\omega_{Si}(x_i - x_i')^2]\}, \qquad (4.76)$$

and the covariance mirrors the additive form of (4.71). In (4.76) we have $S_A = \{1, \ldots, d\}$ and $S_B = S - S_A = \{d+1, \ldots, p\}$ and clearly $S_1 \bigcap S_2 = \varnothing$. Thus, the elements of $\hat{\omega}_{S_A}$ and $\hat{\omega}_S$ can be used to identify the decomposition.

Suppose the first $d' < d$ elements of $\mathbf{x}$ form an additive group. Once more the correct additive structure is nested within (4.74). If we have the first d' elements of $\omega_S = 0$ and the latter $d - d'$ elements of $\omega_{S_A} = 0$, then (4.74) reduces to

$$\sigma^2 \mathbf{c}^{(2)}(\mathbf{x}, \mathbf{x}') = \sigma^2 \{\prod_{i=1}^{d'} \exp[\omega_{S_A i}(x_i - x_i')^2] + \prod_{i=d'+1}^{p} \exp[\omega_{Si}(x_i - x_i')^2]\}, \qquad (4.77)$$

and the covariance mirrors the additive structure of $\eta(.)$. We now have $S_A = \{1, \ldots, d'\}$ and $S_B = S - S_A = \{d' + 1, \ldots, p\}$. This latter result forms the basis of an efficient search algorithm, which we describe presently, for finding any additive decomposition of $\eta(.)$.

Finally, we consider the case when neither S_A or a subset of S_A contains an additive group. Since the Gaussian Process $\eta_{S_A}(.)$ cannot adequately model the dimensions S_A, the optimal covariance structure will clearly be independent of $\mathbf{c}_{S_A}(.,.)$. The more general covariance structure we discussed in section 4.5, $\mathbf{d}(.,.) = \sigma^2_{S_A} \mathbf{c}_{S_A}(.,.) + \sigma^2_S \mathbf{c}_S(.,.)$, would clearly be advantageous here since by estimating $\sigma^2_{S_A} = 0$ the covariance is independent of $\mathbf{c}_{S_A}(.,.)$. Using our equal variances formulation, we find the covariance is independent of $\mathbf{c}_{S_A}(.,.)$ if it can be written as

$$\sigma^2 \mathbf{c}^{(2)}(.,.) = \sigma^2\{1 + \mathbf{c}_S(.,.)\}, \tag{4.78}$$

which occurs when $\boldsymbol{\omega}_{S_A} = \mathbf{0}$.

By examining (4.78), we note the variance is $2\sigma^2$ and the correlation is bounded by 0.5 and 1. Clearly this constraint limits our flexibility in modelling the correlation. However, we have already argued that the power of the Gaussian Process approach is in modelling the large correlations well, which (4.78) is able to do. Since the correlation is bounded below by 0.5, this will have some effect on predictive capability, and as a result this correlation function will be inferior to the more flexible separate variances formulation.

However, our interest is in model selection and for the purpose of model selection we have found this formulation to be sufficient. Using (4.74) we detect we have not found an additive group when $\hat{\boldsymbol{\omega}}_{S_A} = \mathbf{0}$.

4.6.3 Searching for Additive Groups

In principle we could search for all possible additive groups by repeatedly applying the methodology 4.5.2. We would just need to consider all possible subsets of S. However, this would be a very time consuming and inefficient procedure, which would not be possible for large p.

We can implement a far more efficient procedure by using one of our results from 4.6.2. We found that if d' elements of subset S_A form an additive group, then if the $d - d'$ elements of $\boldsymbol{\omega}_{S_A}$, corresponding to the remaining $d - d'$ inputs, are zero, then $\mathbf{c}_{S_A}(.,.)$ depends only on the additive group. We increase our chances of finding an additive group by making S_A as large as possible. This forms the basis of an efficient procedure for finding all additive groups. For p model inputs, we can determine the decomposition of $\eta(.)$ into functions of lower dimensional input vectors, by examining a maximum of p additive correlation structures.

We implement the following algorithm:

1. Let $\Psi = S$
2. Repeat steps 3-5 until Ψ is the empty set.
3. Let S_A contain all elements of Ψ except the first.
4. We fit the Gaussian Process model (4.72), and estimate the smoothing parameters $\boldsymbol{\omega}_{S_A}$ and $\boldsymbol{\omega}_S$ from the posterior mode (4.75).
   (a) The non zero elements of $\hat{\boldsymbol{\omega}}_{S_A}$ (and corresponding zero elements in $\hat{\boldsymbol{\omega}}_S$) indicate one additive group.
   (b) The non zero elements of $\hat{\boldsymbol{\omega}}_S$ (and corresponding zero elements in $\hat{\boldsymbol{\omega}}_{S_A}$) indicate a second additive group. At iteration i these identify subset S_i.
5. Remove the elements of S_i from the set Ψ. Return to step 3 to find the remaining additive groups.

The algorithm works very efficiently for a small number of groups, r. The motivation for the algorithm is that the non zero elements of $\hat{\boldsymbol{\omega}}_S$ should indicate the inputs that interact with the input corresponding to the first element of Ψ. The non zero elements of $\hat{\boldsymbol{\omega}}_{S_A}$ should indicate the inputs that do not interact with the input corresponding to the first element of Ψ. However, when $r > 2$, and at iteration i of the algorithm we have more than one possible decomposition, the algorithm identifies the decomposition that models $\mathbf{y}$ the best, which may not necessarily correspond to our above interpretation. Our numerical work (which we discuss in section 4.6.5) has shown that when we have $r > 2$ groups and consequently more than one decomposition of the form (4.71), the data tend to select a model such that the dimensions of S_A and $S_B = S - S_A$ are similar.

In practice this result means that the sets determined by $\hat{\omega}_{S_A}$ in the early iterations of our algorithm may themselves contain subsets. We find all additive groups by repeatedly applying our algorithm. In step 1 of the algorithm we let $\Psi = S_i$ and the remaining steps are unchanged. We usually only need to do this for the first few large subsets that the algorithm identifies.

One modification to the algorithm proposed above, that improves efficiency is to write the covariance at iteration k of our algorithm as

$$\sigma^2 \mathbf{c}^{(2)}(.,.) = \sigma^2\{\sum_{i=1}^{k}[\mathbf{c}_{S_i}(.,.)] + \mathbf{c}_{S_A}(.,.) + \mathbf{c}_{S_B}(.,.)\}, \tag{4.79}$$

where the first term models the k subsets, $S_1, \ldots S_k$, of the decomposition that we have we have found. Subset S_B contains all inputs not in $S_1, \ldots S_k$. The subset S_A contains all elements of S_B except the first. The form (4.79) contains the same number of unknown parameters as (4.73). However, with this modification, at termination of the algorithm not only do we know the decomposition of $\eta(.)$, but we also have estimates of all the parameters.

4.6.4 Example

We demonstrate the algorithm with the 12 input example

$$\begin{aligned}\eta(\mathbf{x}) &= \eta_1(x_1, x_3, x_6, x_{11}) + \eta_2(x_2, x_7, x_8, x_9) + \eta_3(x_4, x_5, x_{10}, x_{12}) \quad (4.80)\\ &= (x_1^2 + x_3^2 + x_{11}^2 + x_6^2)^{1/2} + (x_2^2 + x_7^2 + x_8^2 + x_9^2)^{1/2}\\ &\quad + (x_4^2 + x_5^2 + x_{10}^2 + x_{12}^2)^{1/2},\end{aligned}$$

which we observe at 100 design points, selected according to a Latin hypercube design. The inputs were independent $U(0,1)$ distributed. The number of terms in the decomposition, and the inputs within these terms were unknown before applying our algorithm.

We begin by specifying the covariance between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ as

$$\sigma^2 \mathbf{c}^{(2)}(\mathbf{x},\mathbf{x}') = \sigma^2\{\mathbf{c}_S(\mathbf{x}_S,\mathbf{x}'_S) + \mathbf{c}_{S_A}(\mathbf{x}_{S_A},\mathbf{x}'_{S_A})\},$$

where $S = S_B = \{1,\dots 12\}$ and $S_A = \{2,\dots 12\}$ and we have parameter vectors $\boldsymbol{\omega}_{S_B} = (\omega_{S_B,1},\dots,\omega_{S_B,12})$ and $\boldsymbol{\omega}_{S_A} = (-,\omega_{S_A,2},\dots,\omega_{S_A,12})$. We begin labelling the elements of $\boldsymbol{\omega}_{S_A}$ from an index of 2 so we can easily identify which input each parameter corresponds to. The $-$ in place of $\omega_{S_A,1}$ signifies we have no parameter corresponding to input x_1 in this correlation function.

We estimate $\boldsymbol{\omega}_{S_B}$ and $\boldsymbol{\omega}_{S_A}$ from their posterior mode as

$$\begin{aligned}\hat{\boldsymbol{\omega}}_{S_B} &= (\hat{\omega}_{S_B,1},0,\hat{\omega}_{S_B,3},\hat{\omega}_{S_B,4},\hat{\omega}_{S_B,5},\hat{\omega}_{S_B,6},0,0,0,\hat{\omega}_{S_B,10},\hat{\omega}_{S_B,11},\hat{\omega}_{S_B,12}),\\ \hat{\boldsymbol{\omega}}_{S_A} &= (-,\hat{\omega}_{S_A,2},0,0,0,0,\hat{\omega}_{S_A,7},\hat{\omega}_{S_A,8},\hat{\omega}_{S_A,9},0,0,0),\end{aligned}$$

where $\hat{\omega}_{S_B,k}$ indicates the k^{th} element of $\boldsymbol{\omega}_{S_B}$ is non zero. The elements of $\hat{\boldsymbol{\omega}}_{S_A}$ and $\hat{\boldsymbol{\omega}}_{S_B}$ that we have indicated as being zero, in most instances were exactly zero. For some parameters the maximization was not at exactly zero. In the algorithm we set a threshold of 0.005, and took any parameter < 0.005 to be zero.

At iteration 1 we have identified the decomposition

$$\eta(\mathbf{x}) = \eta_1(x_1,x_3,x_4,x_5,x_6,x_7,x_{10},x_{11},x_{12}) + \eta_2(x_2,x_7,x_8,x_9). \tag{4.81}$$

Following our algorithm we attempt to further decompose $\eta_2(.)$. Our covariance is written as

$$\sigma^2 \mathbf{c}^{(2)}(\mathbf{x},\mathbf{x}') = \sigma^2\{\mathbf{c}_{S_1}(\mathbf{x}_{S_1},\mathbf{x}'_{S_1}) + \mathbf{c}_{S_A}(\mathbf{x}_{S_A},\mathbf{x}'_{S_A}) + \mathbf{c}_{S_B}(\mathbf{x}_{S_B},\mathbf{x}'_{S_B})\},$$

where $S_1 = \{1,3,4,5,6,7,10,11,12\}$, $S_B = \{2,7,8,9\}$ and $S_A = \{7,8,9\}$. We estimate parameter vectors $\boldsymbol{\omega}_{S_1}$, $\boldsymbol{\omega}_{S_A}$ and $\boldsymbol{\omega}_{S_B}$ from their posterior mode as

$$\begin{aligned}
\hat{\boldsymbol{\omega}}_{S_1} &= (\hat{\omega}_{S_1,1}, -, \hat{\omega}_{S_1,3}, \hat{\omega}_{S_1,4}, \hat{\omega}_{S_1,5}, \hat{\omega}_{S_1,6}, -, -, -, \hat{\omega}_{S_1,10}, \hat{\omega}_{S_1,11}, \hat{\omega}_{S_1,12}),\\
\hat{\boldsymbol{\omega}}_{S_B} &= (-, \hat{\omega}_{S_B,2}, -, -, -, -, \hat{\omega}_{S_B,7}, \hat{\omega}_{S_B,8}, \hat{\omega}_{S_B,9}, -, -, -),\\
\hat{\boldsymbol{\omega}}_{S_A} &= (-, -, -, -, -, -, 0, 0, 0, -, -, -).
\end{aligned}$$

A comparison of $\hat{\boldsymbol{\omega}}_{S_A}$ and $\hat{\boldsymbol{\omega}}_{S_B}$ indicates that no further decomposition of $\eta_2(.)$ is possible. We re-run our algorithm and search for a further decomposition of $\eta_1(.)$. Our covariance is again written as (4.82), but $S_1 = \{2,7,8,9\}$, $S_B = \{1,3,4,5,6,7,10,11,12\}$ and $S_A = \{3,4,5,6,7,10,11,12\}$.

We estimate $\boldsymbol{\omega}_{S_1}$, $\boldsymbol{\omega}_{S_B}$ and $\boldsymbol{\omega}_{S_A}$ from their posterior mode as

$$\begin{aligned}
\hat{\boldsymbol{\omega}}_{S_1} &= (-, \hat{\omega}_{S_1,2}, -, -, -, -, \hat{\omega}_{S_1,7}, \hat{\omega}_{S_1,8}, \hat{\omega}_{S_1,9}, -, -, -),\\
\hat{\boldsymbol{\omega}}_{S_B} &= (\hat{\omega}_{S_B,1}, -, \hat{\omega}_{S_B,3}, 0, 0, \hat{\omega}_{S_B,6}, -, -, -, 0, \hat{\omega}_{S_B,11}, 0),\\
\hat{\boldsymbol{\omega}}_{S_A} &= (-, -, 0, \hat{\omega}_{S_3,4}, \hat{\omega}_{S_A,5}, 0, -, -, -, \hat{\omega}_{S_A,10}, 0, \hat{\omega}_{S_A,12}).
\end{aligned}$$

Thus, we identify the model

$$\eta(\mathbf{x}) = \eta_1(x_2, x_7, x_8, x_9) + \eta_2(x_1, x_3, x_6, x_{11}) + \eta_3(x_4, x_5, x_{10}, x_{12}). \tag{4.82}$$

We attempt to fit 2 further models in order to ensure we have the simplest possible decomposition however we cannot further simplify the function. Thus we have $S_1 = \{2,7,8,9\}$, $S_2 = \{1,3,6,11\}$ and $S_3 = \{4,5,10,12\}$. We found the correct decomposition by examining a total of 5 different correlation structures. In each case we found a strong posterior mode – the data clearly indicated the best correlation structure for each model comparison.

To complete the example we compared the predictive performance of the additive correlation structure with the multiplicative structure of chapter 3. We used $\mathbf{h}(\mathbf{x}) = (1, \mathbf{x})$ in each model and fitted the models using the same 100 design points. We show prediction errors for a further 100 points in *Figure* (4.10).

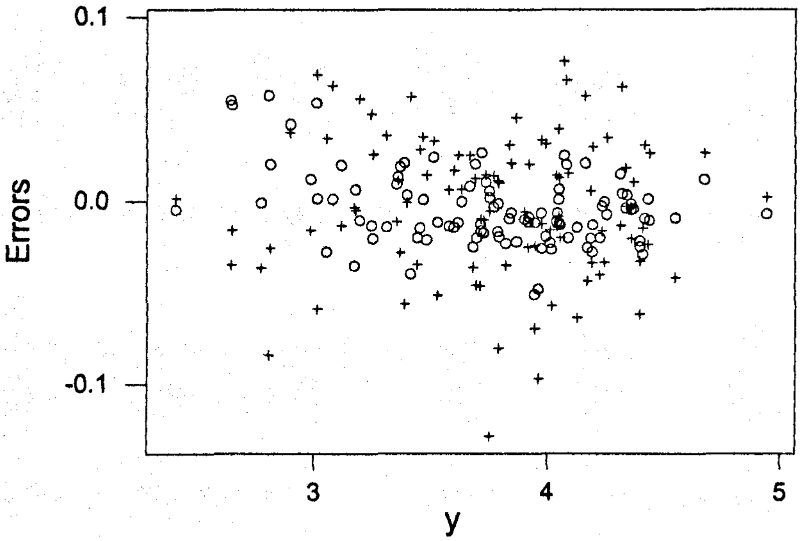

Figure 4.10: Prediction errors: additive model (circles), multiplicative model (crosses)

The mean error was approximately zero using both models and in general the prediction errors were very small under both correlation structures; an artefact of the small variance of Y. However, as can be seen graphically in *Figure* (4.10), the prediction error variance was reduced by a half when using the additive correlation structure.

4.6.5 Discussion

The procedure appears to be very efficient at finding additive groups and we have found that given enough data the additive groups are always identified. In order to attempt to establish properties of the search routine and how many design points are required for different dimensional inputs vectors, and different numbers and sizes of the subsets S_i, we have undertaken a significant amount of empirical work.

In Table (4.2) we show a subset of the problems that we have studied, with the number of inputs in the problem and the minimum number of design points, selected using Latin Hypercube Designs, that were required to consistently find the decomposition.

Inputs	Function	Design Points
2	$x_1 + \sin(x_1) + \cos(x_2)$	10
2	$x_1 + \sin(x_1) + \cos(3x_2)$	12
3	$\sin(x_1) + \cos(x_2 + x_3)$	15
4	$x_1 + \sin(x_1) + \cos(x_2)$	10
4	$x_1 + \sin(x_1) + \cos(x_2) + 0.05\sin(x_1 + x_2)$	12
4	$x_1 + \sin(x_1) + \cos(3x_2)$	12
4	$\sin(x_1 + x_2) + \cos(x_3) + \exp(x_4)$	22
6	$(x_1^2 + x_2^2)^{1/2} + (x_3^2 + x_4^2)^{1/2} + (x_5^2 + x_6^2)^{1/2}$	25
6	$(x_1^2 + x_2^2 + x_3^2)^{1/2} + (x_4^2 + x_5^2 + x_6^2)^{1/2}$	30
6	$(x_1^2 + x_2^2 + x_3^2 + x_4^2)^{1/2} + (x_5^2 + x_6^2)^{1/2}$	28
8	$(x_1^2 + x_2^2 + x_3^2 + x_4^2)^{1/2} + (x_5^2 + x_5^2 + x_7^2 + x_8^2)^{1/2}$	40

Table 4.2: Example problems

Our work has found that the number of points is closely related to the number of active dimensions; in Table (4.2) we show two examples of a function with 2 completely dormant inputs for the $p = 4$ case, and we require the same number of design points as the corresponding $p = 2$ problem. We have also found that for a given p the number of design points that we require to identify a decomposition is related to the numbers and sizes of the subsets S_i; our limited numerical work has found that when most of these subsets are small, we tend to require fewer design points than when these subsets are large. In conclusion our numerical work has not been able to quantify how many design points we require for given p; this clearly depends on both model complexity and the number of active dimensions, which may be less than p.

We have also attempted to assess how robust our method is to a small interaction. Can we detect whether a function is completely additive, or merely almost additive, with a very small interaction? We cannot claim to have studied this case exhaustively however our limited numerical work in this area has indicated that our method is robust to small interaction terms. We considered the example (4.80) with the additional term $c(\sum_{i=1}^{12} x_i^2)^{1/2}$, taking $c \rightarrow 0$. Even with $c = 0.01$ we detected the model was no longer additive. Our method appears to work on the principle of 'accepting' the additive decomposition if the data suggests this is significantly better fitting than the standard model.

4.7 Unknown Partially Additive Decomposition

We now consider a similar situation to that of section 4.4. We have a partially additive decomposition of output $\eta(\mathbf{x})$:

$$\eta(\mathbf{x}) = \eta_1(\mathbf{x}_{(1)}) + \ldots + \eta_r(\mathbf{x}_{(r)}), \tag{4.83}$$

where the $\eta_j(.)$ are functions of lower dimensional input vectors $\mathbf{x}_{(j)}$. We may know all of, some of, or none of the subsets $S_1, \ldots S_r$. In this case we do not have the condition that $S_i \bigcap S_j = \varnothing \, \forall i \neq j$.

4.7.1 Known Decomposition

In section 4.4.2 we noted that for a model of the form (4.83), even if the partially additive decomposition of $\eta(.)$ was known, we could not select our design points in order to exploit this structure. As a result, we noted that for a partially additive model there is no equivalent to our work of section 4.3.2. However, by a straightforward application of the methodology of section 4.5, we can use this structural prior information to fit a partially additive correlation structure.

4.7.2 Regression Based Model Comparison

In practice we are unlikely to know the full partially additive decomposition of $\eta(.)$ with certainty. Suppose that we want to test if $\eta(\mathbf{x})$ can be decomposed as in (4.83). We can compare the partially additive model with the Gaussian Process model of chapter 3, similar to our method of the previous section. We again adopt a method based upon the perfectly fitting mean function.

Suppose we believe that output, $\eta(\mathbf{x})$, contains a function of inputs $\mathbf{x}_{S_A}$. For ease of notation we once more assume that $S_A = \{1, \ldots d\}$. The output can be decomposed as

$$\eta(\mathbf{x}) = \eta_{S_A}(\mathbf{x}_{S_A}) + \eta_S(\mathbf{x}_S). \tag{4.84}$$

We can think of $\eta_{S_A}(.)$ as our non-parametric mean function $\mu(.)$, which in turn has mean function $\mathbf{h}(.)^{\mathrm{T}}\boldsymbol{\beta}$ and correlation function $\mathbf{c}_{S_A}(.,.)$, which is a function of the first d inputs in $\mathbf{x}$. The second term $\eta_S(.)$ is a zero mean Gaussian Process with correlation function $\mathbf{c}_S(.,.)$, which is a function of all p inputs. Therefore, (4.84) is a Gaussian Process with mean function $\mathbf{h}(.)^{\mathrm{T}}\boldsymbol{\beta}$ and covariance

$$\sigma^2 \mathbf{c}^{(2)}(\mathbf{x}, \mathbf{x}') = \sigma^2\{\mathbf{c}_{S_A}(\mathbf{x}_{S_A}, \mathbf{x}'_{S_A}) + \mathbf{c}_S(\mathbf{x}_S, \mathbf{x}'_S)\}. \tag{4.85}$$

Using our exponential correlation function, we write the covariance as

$$\sigma^2 \mathbf{c}^{(2)}(\mathbf{x},\mathbf{x}') = \sigma^2\{\prod_{i=1}^{d} \exp[\omega_{S_A i}(x_i - x_i')^2] + \prod_{i=1}^{p} \exp[\omega_{Si}(x_i - x_i')^2]\}, \tag{4.86}$$

and estimate the parameters $\boldsymbol{\omega}_S, \boldsymbol{\omega}_{S_A}$ from the posterior mode of

$$f(\boldsymbol{\omega}_S, \boldsymbol{\omega}_{S_A} \,|\, \mathbf{y}) = \hat{\sigma}^{((2)-(n-q))}|\mathbf{A}^{**}|^{-1/2}|\mathbf{H}^{\mathrm{T}}\mathbf{A}^{**-1}\mathbf{H}|^{-1/2}. \tag{4.87}$$

A partially additive decomposition of $\eta(.)$ is more complex than the additive decomposition, since each input may be present in at least two partially additive functions. As a result we have more special structures nested within (4.86). We can use $\hat{\boldsymbol{\omega}}_S, \hat{\boldsymbol{\omega}}_{S_A}$ to detect the correct structure.

1. $\hat{\boldsymbol{\omega}}_{S_A} = \mathbf{0}$. This tells us that the subset S_A does not contain a partially additive group. If we also find the i^{th} element of $\hat{\boldsymbol{\omega}}_S$ is equal to zero, this indicates the parametric fit, $\mathbf{h}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta}$, explains the variation with respect to input x_i perfectly – this follows from the discussion in section 4.6.

2. $\hat{\boldsymbol{\omega}}_S > \mathbf{0}$ and at least one element of $\hat{\boldsymbol{\omega}}_{S_A}$ is greater than zero. This tells us that we have found a partially additive group. However, since all elements of $\hat{\boldsymbol{\omega}}_S$ are greater than zero, S_A does not contain any unique elements.

3. $\hat{\boldsymbol{\omega}}_{S_A} > \mathbf{0}$ and the first d elements of $\hat{\boldsymbol{\omega}}_S$ are zero. This tells us that we have identified $S_A = \{1, \ldots, d\}$ and $S_B = S - S_A = \{d+1, \ldots, p\}$ such that $S_A \bigcap S_B = \varnothing$ – that is an additive decomposition.

4. $\hat{\boldsymbol{\omega}}_{S_A} > \mathbf{0}$ and d' (but not all) of the first d elements of $\hat{\boldsymbol{\omega}}_S$ are zero. We have explained all the variability in d' dimensions, but $\eta(.)$ contains other functions of the remaining $d - d'$ inputs. That is S_A contains some but not all unique elements.

Given enough data the posterior mode of (4.87) will be able to distinguish between these four structures.

4.7.3 Searching for Partially Additive Groups

Finding a single partially additive group is not difficult in principle. By adopting the same procedure as proposed in 4.6.3 and taking S_A to be as large a set as possible, we maximize our chances of finding a partially additive group. However, in order to determine all the partially additive groups we need a more structured approach than we proposed in 4.6.3.

Additive groups of inputs are a special case of a partially additive decomposition of $\eta(.)$. Our numerical work suggests that additive groups are always detected first. We begin by finding all the additive groups of inputs using the methodology of section 4.6.2. We can search within each of these additive groups for a partially additive decomposition.

We find all r^* subsets of S, such that $S_i \bigcap S_j = \varnothing \, \forall \, i \neq j$ using the algorithm of section 4.5.2. The output may then be written as

$$\eta(\mathbf{x}) = \eta_{S_1}(\mathbf{x}_{S_1}) + \ldots \eta_{S^*_r}(\mathbf{x}_{S_{r^*}}). \tag{4.88}$$

We then search for subsets of each S_i for $j = 1, \ldots r^*$. We search these additive groups one at a time so that we may limit the dimension of our numerical maximization.

We propose the following algorithm for decomposing $\eta_{S_j}(\mathbf{x}_{S_j})$

1. Let Ψ denote the set of inputs within $\mathbf{x}_{S_j}$.

2. Set $n = 1$.

3. Set $m = 1$.

4. Repeat steps 5 to 7 until termination

5. Let $\mathbf{x}_{S_j^n}$ contain all elements of Ψ but the m^{th}. We fit our Gaussian Process model with covariance

$$\sigma^2 \mathbf{c}^{(2)}(\mathbf{x},\mathbf{x}') = \sigma^2 \sum_{k=1}^{r^*} \mathbf{c}_{S_k}(\mathbf{x}_{S_k},\mathbf{x}'_{S_k}) + \sigma^2 \sum_{i=1}^{n} \mathbf{c}_{S_j^i}(\mathbf{x}_{S_j^i},\mathbf{x}'_{S_j^i}), \tag{4.89}$$

   and estimate parameters $\boldsymbol{\omega}_{S_1},\ldots,\boldsymbol{\omega}_{S_r^*}$ and $\boldsymbol{\omega}_{S_j^i},\ldots,\boldsymbol{\omega}_{S_j^n}$

6. We examine the parameters of the correlation functions $\mathbf{c}_{S_j}(\mathbf{x}_{S_j},\mathbf{x}'_{S_j})$ and $\mathbf{c}_{S_j^n}(\mathbf{x}_{S_j^n},\mathbf{x}'_{S_j^n})$:

   (a) If $\hat{\boldsymbol{\omega}}_{S_j^n} = \mathbf{0}$, set $m = m + 1$.

   (b) If $\hat{\boldsymbol{\omega}}_{S_j^n} \neq \mathbf{0}$, we have found a partially additive group. Set $n = n + 1$. If any elements of $\boldsymbol{\omega}_{S_j}$ are zero, set $m = 1$ and remove the corresponding inputs from the set Ψ.

7. Terminate when $m + 1$ exceeds the dimension of the set Ψ.

Given enough data, $\mathbf{y}$, the algorithm will find all partially additive groups. However, the amount of data we require will depend on the order of the decomposition, and how many inputs each sub function contains. The decomposition may contain many more parameters than we had in our Gaussian Process model of chapter 3, so it may not be feasible to find the full partially additive decomposition of $\eta(.)$.

4.7.4 Example

We demonstrate the algorithm with the 12 input example

$$\begin{aligned}\eta(\mathbf{x}) &= \eta_1(x_1+x_4+x_7)+\eta_2(x_7+x_{10}+x_{11})+\eta_3(x_1+x_{11}+x_{12}) \\ &+ \eta_4(x_5+x_8+x_9)+\eta_5(x_3+x_6+x_9)+\eta_6(x_2+x_6+x_8) \qquad (4.90) \\ &= (x_1^2+x_4^2+x_7^2)^{1/2}+(x_7^2+x_{10}^2+x_{11}^2)^{1/2}+(x_1^2+x_{11}^2+x_{12}^2)^{1/2} \\ &+ (x_5^2+x_8^2+x_9^2)^{1/2}+(x_3^2+x_6^2+x_9^2)^{1/2}+(x_2^2+x_6^2+x_8^2)^{1/2},\end{aligned}$$

which we observe at the same 100 design point as in the previous example. Again, the problem is designed to be challenging, containing 2 additive groups of 6 inputs, which can each be further decomposed into 3 partially additive groups.

We begin by specifying the covariance between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ as

$$\sigma^2\mathbf{c}^{(2)}(\mathbf{x},\mathbf{x}') = \sigma^2\{\mathbf{c}_{S_A}(\mathbf{x}_{S_A},\mathbf{x}'_{S_A})+\mathbf{c}_{S_B}(\mathbf{x}_{S_B},\mathbf{x}'_{S_B})\},$$

where $S_A = \{2,\dots,12\}$ and $S = S_B = \{1,\dots 12\}$. We have parameter vectors $\boldsymbol{\omega}_{S_A} = (-,\omega_{S_A,2},\dots,\omega_{S_A,12})$ and $\boldsymbol{\omega}_{S_B} = (\omega_{S_B,1},\dots,\omega_{S_B,12})$. Note this first stage is identical to the procedure we used for a purely additive decomposition.

We estimate $\boldsymbol{\omega}_{S_A}$ and $\boldsymbol{\omega}_{S_B}$ from their posterior mode as

$$\begin{aligned}\hat{\boldsymbol{\omega}}_{S_A} &= (-,\hat{\omega}_{S_A,2},\hat{\omega}_{S_A,3},0,\hat{\omega}_{S_A,5},\hat{\omega}_{S_A,6},0,\hat{\omega}_{S_A,8},\hat{\omega}_{S_A,9},0,0,0). \\ \hat{\boldsymbol{\omega}}_{S_B} &= (\hat{\omega}_{S_B,1},0,0,\hat{\omega}_{S_B,4},0,0,\hat{\omega}_{S_B,7},0,0,\hat{\omega}_{S_B,10},\hat{\omega}_{S_B,11},\hat{\omega}_{S_B,12}),\end{aligned}$$

Thus, we find that $\eta(\mathbf{x})$ can be written as

$$\eta(\mathbf{x}) = \eta_1(x_1,x_4,x_7,x_{10},x_{11},x_{12})+\eta_2(x_2,x_3,x_5,x_6,x_8,x_9). \qquad (4.91)$$

We find no further decomposition into additive groups is possible. We have identified the two subsets of S as $S_1 = \{1,4,7,10,11,12\}$ and $S_2 = \{2,3,5,6,8,9\}$. We now begin our algorithm to find all partially additive groups.

Following our algorithm, we specify the covariance between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ as

$$\sigma^2 \mathbf{c}^{(2)}(\mathbf{x},\mathbf{x}') = \sigma^2\{\mathbf{c}_{S_1}(\mathbf{x}_{S_1},\mathbf{x}'_{S_1}) + \mathbf{c}_{S_2}(\mathbf{x}_{S_2},\mathbf{x}'_{S_2}) + \mathbf{c}_{S_1^1}(\mathbf{x}_{S_1^1},\mathbf{x}'_{S_1^1})\},$$

where $S_1 = \{1,4,7,10,11,12\}$ and $S_2 = \{2,3,5,6,8,9\}$ and $S_1^1 = \{4,7,10,11,12\}$.

We estimate $\boldsymbol{\omega}_{S_1}$, $\boldsymbol{\omega}_{S_1^1}$ $\boldsymbol{\omega}_{S_2}$ from their posterior mode as

$$\begin{aligned}
\hat{\boldsymbol{\omega}}_{S_1} &= (\hat{\omega}_{S_1,1}, -, -, \hat{\omega}_{S_1,4}, -, -, \hat{\omega}_{S_1,7}, -, -, 0, \hat{\omega}_{S_1,11}, \hat{\omega}_{S_1,12}),\\
\hat{\boldsymbol{\omega}}_{S_1^1} &= (-, -, -, 0, -, -, \hat{\omega}_{S_1^1,7}, -, -, \hat{\omega}_{S_1^1,10}, \hat{\omega}_{S_1^1,11}, 0),\\
\hat{\boldsymbol{\omega}}_{S_2} &= (-, \hat{\omega}_{S_2,2}, \hat{\omega}_{S_2,3}, -, \hat{\omega}_{S_2,5}, \hat{\omega}_{S_2,6}, -, \hat{\omega}_{S_2,8}, \hat{\omega}_{S_2,9}, -, -, -).
\end{aligned}$$

We find our first partially additive group, with $S_1^1 = \{7,10,11\}$. We examine $\hat{\boldsymbol{\omega}}_{S_1}$ and note that $\omega_{\hat{S_1},10}$ is zero, and therefore input x_{10} is unique to S_1^1.

Thus, $\eta(\mathbf{x})$ can be written as

$$\eta(\mathbf{x}) = \eta_1(x_1,x_4,x_7,x_{11},x_{12}) + \eta_2(x_7,x_{10},x_{11}) + \eta_3(x_2,x_3,x_5,x_6,x_8,x_9). \quad (4.92)$$

We specify the covariance between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ as

$$\sigma^2 \mathbf{c}^{(2)}(\mathbf{x},\mathbf{x}') = \sigma^2\{\mathbf{c}_{S_1}(\mathbf{x}_{S_1},\mathbf{x}'_{S_1}) + \mathbf{c}_{S_2}(\mathbf{x}_{S_2},\mathbf{x}'_{S_2}) + \sum_{i=1}^{2} \mathbf{c}_{S_1^i}(\mathbf{x}_{S_1^i},\mathbf{x}'_{S_1^i})\},$$

where $S_1 = \{1,4,7,11,12\}$, $S_2 = \{2,3,5,6,8,9\}$, $S_1^1 = \{7,10,11\}$ $S_1^2 = \{4,7,11,12\}$.

For the above covariance structure we found $\hat{\boldsymbol{\omega}}_{S_1^2} = \mathbf{0}$, which indicates we have not found a partially additive group on this iteration of the algorithm. Following our algorithm we again specify a covariance of the form (4.93), but with $S_1^2 = \{1, 7, 11, 12\}$ whilst the remaining subsets are unchanged.

We estimate $\boldsymbol{\omega}_{S_1}$, $\boldsymbol{\omega}_{S_1^1}$, $\boldsymbol{\omega}_{S_1^2}$ and $\boldsymbol{\omega}_{S_2}$ from their posterior mode as

$$\begin{aligned}
\hat{\boldsymbol{\omega}}_{S_1} &= (\hat{\omega}_{S_1,1}, -, -, \hat{\omega}_{S_1,4}, -, -, \hat{\omega}_{S_1,7}, -, -, -, 0, 0), \\
\hat{\boldsymbol{\omega}}_{S_1^1} &= (-, -, -, -, -, -, \hat{\omega}_{S_1^1,7}, -, -, \hat{\omega}_{S_1^1,10}, \hat{\omega}_{S_1^1,11}, 0), \\
\hat{\boldsymbol{\omega}}_{S_1^2} &= (\hat{\omega}_{S_1^2,1}, -, -, -, -, -, 0, -, -, -, \hat{\omega}_{S_1^2,11}, \hat{\omega}_{S_1^2,12}), \\
\hat{\boldsymbol{\omega}}_{S_2} &= (-, \hat{\omega}_{S_2,2}, \hat{\omega}_{S_2,3}, -, \hat{\omega}_{S_2,5}, \hat{\omega}_{S_2,6}, -, \hat{\omega}_{S_2,8}, \hat{\omega}_{S_2,9}, -, -, -).
\end{aligned}$$

We find our second partially additive group with $S_{1^2} = \{1, 11, 12\}$. We examine $\hat{\boldsymbol{\omega}}_{S_1}$ and note that $\omega_{\hat{S_1},11}$ and $\omega_{\hat{S_1},12}$ are zero. Our two partially additive groups model all the variability from inputs x_{10}, x_{11}, x_{12}.

Thus, $\eta(\mathbf{x})$ can be written as

$$\begin{aligned}
\eta(\mathbf{x}) &= \eta_1(x_1, x_4, x_7) + \eta_2(x_7, x_{10}, x_{11}) + \eta_3(x_1, x_{11}, x_{12}) \\
&\quad + \eta_4(x_2, x_3, x_5, x_6, x_8, x_9). \qquad (4.93)
\end{aligned}$$

We fit 3 more models, in an attempt to simplify $\eta_1(x_1, x_4, x_7)$, however these correctly indicated no further simplification was possible, hence $\eta_1(x_1, x_4, x_7)$ is our final partially additive group.

We now try to decompose the second additive group. We specify the covariance between $\eta(\mathbf{x})$ and $\eta(\mathbf{x}')$ as

$$\sigma^2 \mathbf{c}^{(2)}(\mathbf{x}, \mathbf{x}') = \sigma^2\{\mathbf{c}_{S_1}(\mathbf{x}_{S_1}, \mathbf{x}'_{S_1}) + \mathbf{c}_{S_2}(\mathbf{x}_{S_2}, \mathbf{x}'_{S_2}) + \mathbf{c}_{S_2^1}(\mathbf{x}_{S_2^1}, \mathbf{x}'_{S_2^1})\},$$

where $S_1 = \{1, 4, 7, 10, 11, 12\}$, $S_2 = \{2, 3, 5, 6, 8, 9\}$ and $S_2^1 = \{3, 5, 6, 8, 9\}$.

We estimate $\boldsymbol{\omega}_{S_1}$, $\boldsymbol{\omega}_{S_2}$ and $\boldsymbol{\omega}_{S_2^1}$ from their posterior mode as

$$\begin{aligned}
\hat{\boldsymbol{\omega}}_{S_1} &= (\hat{\omega}_{S_1,1}, -, -, \hat{\omega}_{S_1,4}, -, -, \hat{\omega}_{S_1,7}, -, -, \hat{\omega}_{S_1,10}, \hat{\omega}_{S_1,11}, \hat{\omega}_{S_1,12}),\\
\hat{\boldsymbol{\omega}}_{S_2} &= (-, \hat{\omega}_{S_2,2}, 0, -, \hat{\omega}_{S_2,5}, \hat{\omega}_{S_2,6}, -, \hat{\omega}_{S_2,8}, \hat{\omega}_{S_2,9}, -, -, -),\\
\hat{\boldsymbol{\omega}}_{S_2^1} &= (-, -, \hat{\omega}_{S_2^1,3}, -, 0, \hat{\omega}_{S_2^1,6}, -, 0, \hat{\omega}_{S_2^1,9}, -, -, -).
\end{aligned}$$

We find the first partially additive group with $S_2^1 = \{3, 6, 9\}$. We note that dimension 3 is unique to S_2^1. Thus

$$\begin{aligned}
\eta(\mathbf{x}) &= \eta_1(x_7, x_{10}, x_{11}) + \eta_2(x_1, x_{11}, x_{12}) + \eta_3(x_1, x_4, x_7)\\
&+ \eta_4(x_2, x_5, x_6 x_8, x_9) + \eta_5(x_3, x_6, x_9).
\end{aligned} \tag{4.94}$$

We now attempt to find the remaining partially additive groups. Our covariance now takes the form

$$\sigma^2 \mathbf{c}^{(2)}(\mathbf{x}, \mathbf{x}') = \sigma^2\{\mathbf{c}_{S_1}(\mathbf{x}_{S_1}, \mathbf{x}'_{S_1}) + \mathbf{c}_{S_2}(\mathbf{x}_{S_2}, \mathbf{x}'_{S_2}) + \sum_{i=1}^{2} \mathbf{c}_{S_2^i}(\mathbf{x}_{S_2^i}, \mathbf{x}'_{S_2^i})\},$$

where $S_1 = \{1, 4, 7, 10, 11, 12\}$, $S_2 = \{2, 5, 6, 8, 9\}$, $S_2^1 = \{3, 6, 9\}$ and $S_2^2 = \{5, 6, 8, 9\}$.

We estimate the parameters via their joint posterior mode and find

$$\begin{aligned}
\hat{\boldsymbol{\omega}}_{S_1} &= (\hat{\omega}_{S_1,1}, -, -, \hat{\omega}_{S_1,4}, -, -, \hat{\omega}_{S_1,7}, -, -, \hat{\omega}_{S_1,10}, \hat{\omega}_{S_1,11}, \hat{\omega}_{S_1,12}),\\
\hat{\boldsymbol{\omega}}_{S_2} &= (-, \hat{\omega}_{S_2,2}, -, -, \hat{\omega}_{S_2,5}, \hat{\omega}_{S_2,6}, -, \hat{\omega}_{S_2,8}, 0, -, -, -),\\
\hat{\boldsymbol{\omega}}_{S_2^1} &= (-, -, \omega_{\hat{S}_2^1,3}, -, -, \omega_{\hat{S}_2^1,6}, -, -, \omega_{\hat{S}_2^1,9}, -, -, -),\\
\hat{\boldsymbol{\omega}}_{S_2^2} &= (-, -, -, -, \omega_{\hat{S}_2^2,5}, \omega_{\hat{S}_2^2,6}, -, \omega_{\hat{S}_2^2,8}, \omega_{\hat{S}_2^2,9}, -, -, -).
\end{aligned}$$

This final covariance structure produces only partial success. We expected to find $\omega_{\hat{S}_2^2,6} = 0$ and $\omega_{\hat{S}_2,5} = 0$, which would identify the final two partially additive groups. The estimates of these two parameters were close to zero, but the posterior was quite flat in the area around the mode. The 100 design points were insufficient in order to identify this final group. However, when we used a different 100 point LHS we were able to correctly identify all the partially additive groups.

To complete the example we compared the predictive performance of the partially additive correlation structure with the multiplicative structure of chapter 3. Once more we use $\mathbf{h}(\mathbf{x}) = (1, \mathbf{x})$ in each model and fit the models using the same 100 design points. We show prediction errors using these two models for a further 100 points in *Figure* (4.11).

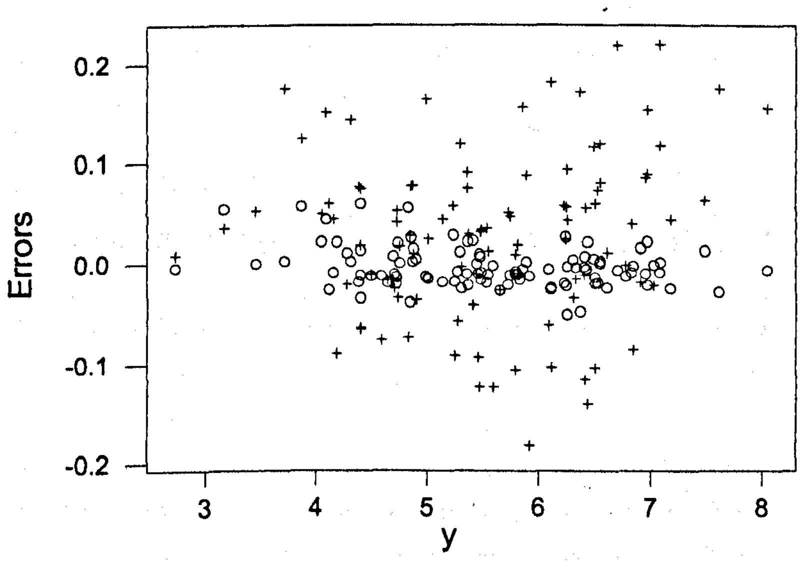

Figure 4.11: Prediction errors: partially additive model (circles), multiplicative model (crosses)

We note from *Figure* (4.11) that the variability about a mean error of zero is far less when using our partially additive correlation structure. The partially additive correlation structure offers a significant improvement over the multiplicative structure.

4.7.5 Discussion

The above example highlights that a partially additive decomposition is complex. Our method can work well, but performance obviously depends on how complex the decomposition of $\eta(.)$ is. Given enough data we can find all partially additive groups, but the number of design points we require to identify a partially additive decomposition may well outnumber the number of design points required to fit the Gaussian Process model reviewed in chapter 3.

This example was a difficult test with all 12 inputs active. In such a function 100 design points is not many to determine whether an order 5 interaction can be decomposed. We can in principle handle many more inputs provided we have factor sparsity. In problems with factor sparsity, like those considered in *Welch et al.*(64) our algorithm is very effective.

In summary:

1. given enough data we can find all partially additive groups;

2. the algorithm works best when we have factor sparsity;

3. we may not be able to fully decompose $\eta(.)$;

4. any decomposition we find aids modelling.

4.8 Conclusions

In this chapter we have considered a series of models for use when we have structural prior information about $\eta(.)$. We defined the two distinct cases of additive and partially additive decompositions of $\eta(.)$. We considered inference when we had known decompositions of $\eta(.)$, and showed that predictive performance was better when using the additive structure as opposed to the multiplicative structure in the model we reviewed in chapter 3. Moreover we achieve better performance with fewer design points.

We then considered inference for $\eta(.)$ in light of uncertainty about a decomposition. We developed methodology for searching for additive and partially additive groups. We used the parameters of the correlation functions to identify model structure. We used a specific form of correlation function; however any correlation function that can be written as a product of one dimensional, one parameter correlation functions can be used. We found additive groups are easily identified – we require fewer design points than we would normally model $\eta(.)$ with, using the methodology of chapter 3. We found partially additive models are more difficult to fully decompose, and we may not be able to fully decompose $\eta(.)$. We showed by way of two examples that the additional computational burden of searching for additive and partially additive structures is justified, especially for a computationally expensive function since predictive performance is improved.

A final conclusion relates to the wider applicability of the methodology developed in this chapter. The motivation for the methodology of sections 4.6-4.7 was to search for models with an unknown simpler structure. However, the application that we discuss in chapter 7 showed that the method may be more widely applicable in the context of model validation. Often the creation of a computer model is an iterative procedure and the earlier versions of the model frequently contain omissions and programming errors; our method could be used to efficiently audit such models, assessing whether the model contained an erroneous (and unintended) simplified structure that did not reflect important aspects of the process being modelled or whether the model contained interactions that were not expected. Equally the method could be useful in showing that specific structure expected in the model is actually present.

Chapter 5

Uncertainty and Sensitivity Analysis for Decomposable Functions

5.1 Introduction

In this chapter we consider uncertainty and sensitivity analysis for the case when we can decompose the function $\eta(.)$ as

$$\eta(.) = \eta_1(.) + \ldots + \eta_r(.). \tag{5.1}$$

We use the methodology developed by *Haylock and O'Hagan*(23) and *Oakley and O'Hagan*(49),(51), which was reviewed in chapter 3. Whilst the methodology we use for uncertainty and sensitivity analysis is not new, we do need modifications to the calculations given in sections 3.3 and 3.4 as a result of the changes to the Gaussian Process model that we made in chapter 4. We describe these modifications in detail in this chapter.

We have two broad categories of decomposition to consider.

1. *Decomposition known a priori.*

   The first of these classes is where we have data available, or sufficient knowledge about the form of the decomposition, such that we have been able to use a one at a time design in order to observe each of the sub functions, $\eta_j(.)$, directly. We found posterior distributions for $\eta(.)$ under these decompositions in sections 4.3 and 4.4.

2. *Decomposition unknown a priori.*

   In our second class, the decomposition was unknown a priori. Resultantly we use a space filling design. Using the methodology described in the previous chapter, the decomposition can be identified, and we act as if the decomposition is known with certainty. We do not have observations of the sub functions, $\eta_j(.)$, however we have been able to model $\eta(.)$, with an additive covariance structure. We found the posterior distribution for $\eta(.)$ in 4.5 of the previous chapter, and examined additive and partially additive decompositions in detail in 4.6 and 4.7.

In this chapter we consider measures of uncertainty and sensitivity for these two classes of decomposition, that make use of the structural information (5.1). An example demonstrating the methodology developed in this chapter is given in chapter 7.

5.2 Inference for Known Decompositions

We can decompose the output, $\eta(\mathbf{x})$ as

$$\eta(\mathbf{x}) = \eta_1(\mathbf{x}_{(1)}) + \ldots + \eta_r(\mathbf{x}_{(r)}). \tag{5.2}$$

As in the previous chapter we let S represent the set of integers $i = 1, \ldots, p$ and $S_1, \ldots, S_r$ are subsets of S. Function $\eta_j(\mathbf{x}_{(j)})$ is a function of the sub vector $\mathbf{x}_{(j)}$, whose elements are identified by the elements of S_j. For additive models we have a mutually exclusive and exhaustive partition of $\mathbf{x}$, and resultantly $S_i \bigcap S_j = \varnothing \forall i \neq j$. For partially additive models each x_i may be present in more than one $\mathbf{x}_{(j)}$ and resultantly this condition need not hold. As discussed in the previous chapter, we take the functions $\eta_j(\mathbf{x}_{(j)})$ for $j = 1 \ldots r$, to be independent.

In this section we consider inference for the observable functions case (discussed in section 4.3.1), where we are able to make observations $y_{(j)}$ of each function, with the property that $\sum_{j=1}^{r} y_{(j)} = y$. We don't explicitly consider inference for the unobservable functions model (discussed in section 4.3.2), since inference for this model is trivial given the theory that we develop. However, we indicate how inference for this model differs from our analysis at the end of this section.

5.2.1 Uncertainty

We consider measures of uncertainty about Y, where $Y = \eta(\mathbf{X})$. We are interested in the same summaries as discussed in chapter 3, $E[Y|\eta(.)]$, $Var[Y|\eta(.)]$ and the distribution function of $Y|\eta(.)$.

Expectation

We first consider the expectation of Y, conditional on $\eta(.)$, which can be written

$$E[Y|\eta(.)] = E[Y_{(1)}|\eta_1(.)] + \ldots + E[Y_{(r)}|\eta_r(.)]. \tag{5.3}$$

In chapter 3 we found the posterior distribution of $E[Y|\eta(.)]$ for the case where we did not utilize structural prior information. We can find the posterior distributions of $E[Y_{(1)}|\eta_1(.)], \ldots E[Y_{(r)}|\eta_r(.)]$ similarly. Letting $K_{1,(j)}$ denote $E[Y_{(j)}|\eta_j(.)]$, we have

$$K_{1,(j)} = \int_{\mathcal{X}} \eta_j(\mathbf{x}_{(j)})\, dG(\mathbf{x}), \tag{5.4}$$

and the posterior distribution of $K_{1,(j)}$ is given by

$$\frac{K_{1,(j)} - \hat{k}_{1,(j)}}{\hat{\sigma}_j\sqrt{\frac{n_j-q_j-2}{n_j-q_j}W_{(j)}}} \mid \boldsymbol{\omega}_j, \mathbf{y}_{(j)} \sim t_{n_j-q_j}. \tag{5.5}$$

In arriving at (5.5) we followed the same steps as in section 3.3 so we need not repeat them here. The expectation and variance of $K_{1,(j)}$ involve integrals $\mathbf{R}_{(j)}, \mathbf{T}_{(j)}$ and $U_{(j)}$. Again, these integrals are almost identical to equations (3.51)-(3.53) of chapter 3 so we do not list them again. By properties of expectation, the integral (5.4) and hence $\mathbf{R}_{(j)}, \mathbf{T}_{(j)}$ and $U_{(j)}$ can be reduced to integrals with respect to just the sub vector $\mathbf{x}_{(j)}$.

We are interested in $K_1 = E[Y|\eta(.)]$ and this is given by the sum

$$K_1 = \sum_{j=1}^{r} K_{1,(j)}. \tag{5.6}$$

It is simple to calculate the expectation and variance of (5.6). Letting $v_j = n_j - q_j$ and

$$Z_j = \frac{K_{1,(j)} - \hat{k}_{1,(j)}}{\hat{\sigma}_j\sqrt{\frac{v_j-2}{v_j}W_{(j)}}} \mid \boldsymbol{\omega}_j, \mathbf{y}_{(j)}, \tag{5.7}$$

for $j = 1, \ldots, r$, we have

$$\begin{aligned} E[K_1|\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r, \mathbf{y}_{(1)}, \ldots \mathbf{y}_{(r)}] &= \hat{k}_{1,(1)} + \ldots + \hat{k}_{1,(r)}, & (5.8)\\ Var[K_1|\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r, \mathbf{y}_{(1)}, \ldots \mathbf{y}_{(r)}] &= \{\sigma_1^2\frac{v_1-2}{v_1}W_{(1)}Var(Z_1) + \ldots \\ &\quad + \sigma_r^2\frac{v_r-2}{v_r}W_{(r)}Var(Z_r)\}. & (5.9) \end{aligned}$$

We showed how to approximate a weighted sum of t-distributions in 4.3.1. If we wish to calculate summaries of K_1 such as $P(K_1 < c)$, for constant c, then a similar approximation will be useful here, especially for large r. We equate the first 4 moments of (5.6) to those of a t-distribution with v degrees of freedom. We calculate v by equating the kurtosis of (5.6), denoted $\beta_2(t_v)$ to that of a t-distribution with v degrees of freedom. The calculation of $\beta_2(t_v)$ was described in detail in 4.3.1 and is not repeated here.

We find v using

$$v = \frac{4\beta_2(t_v) - 6}{\beta_2(t_v) - 3}, \tag{5.10}$$

which follows from chapter 4.

Our approximation to (5.6) is therefore

$$\frac{K_1 - E[K_1]}{\hat{\sigma_W}} |\boldsymbol{\omega}_1, \dots \boldsymbol{\omega}_r, \mathbf{y}_{(1)}, \dots \mathbf{y}_{(r)} \sim t_v, \tag{5.11}$$

where

$$\hat{\sigma_W} = \frac{Var[K_1]}{Var[t_v]}, \tag{5.12}$$

and $Var[t_v]$ is the variance of a t-distribution with v degrees of freedom. It is trivial to calculate summaries from (5.11).

Variance

We now consider the variance of Y conditional on $\eta(.)$, which can be written as

$$Var[Y|\eta(.)] = Var[Y_{(1)}|\eta_1(.)] + \dots + Var[Y_{(r)}|\eta_r(.)]. \tag{5.13}$$

We can consider inference for each of the variances in (5.13) individually. Following the same process as in chapter 3, we find the posterior expectation (with respect to the posterior distribution of $\eta_j(.)$) of the variance of $Y_{(j)}$ as

$$\begin{aligned} E^*[Var\{Y_{(j)}\}|\boldsymbol{\omega}_j,\mathbf{y}_{(j)}] &= E^*[K_{2,(j)} - K^2_{1,(j)}|\boldsymbol{\omega}_j,\mathbf{y}_{(j)}] \\ &= E^*[K_{2,(j)}|\boldsymbol{\omega}_j,\mathbf{y}_{(j)}] - (Var^*[K_{1,(j)}|\boldsymbol{\omega}_j,\mathbf{y}_{(j)}] + E^*[K_{1,(j)}|\boldsymbol{\omega}_j,\mathbf{y}_{(j)}]^2). \end{aligned}$$

where

$$K_{2,(j)}|\boldsymbol{\omega}_j,\mathbf{y}_{(j)} = E^*\left[\int_{\mathcal{X}} \eta_j^2(\mathbf{x}_{(j)})|\boldsymbol{\omega}_j,\mathbf{y}_{(j)}\, dG(\mathbf{x})\right]. \tag{5.14}$$

Again, we find that $K_{2,(j)}|\boldsymbol{\omega}_j,\mathbf{y}_{(j)}$ reduces to an integral with respect to the sub vector $\mathbf{x}_{(j)}$, and hence $E^*[Var\{Y_{(j)}\}|\boldsymbol{\omega}_j,\mathbf{y}_{(j)}]$ depends on just $\mathbf{x}_{(j)}$.

After making inference about each term in (5.13), we calculate the expected value of the variance of Y from the sum

$$E^*[Var\{Y\}|\boldsymbol{\omega}_1,\ldots\boldsymbol{\omega}_r,\mathbf{y}_{(1)},\ldots\mathbf{y}_{(r)}] = \sum_{j=1}^{r} E^*[Var\{Y_{(j)}\}|\boldsymbol{\omega}_j,\mathbf{y}_{(j)}]. \tag{5.15}$$

Distribution Function

The final summary we require is the distribution function of $Y|\eta(.)$. The posterior distribution of the distribution function requires us to calculate the integral

$$F_{Y|\eta(.)}(s) = \int_{\mathcal{X}} I\{\eta(\mathbf{x}) \leq s\}\, dG(\mathbf{x}). \tag{5.16}$$

However, as we found in chapter 3, the posterior distribution of $F_Y(s)$ is intractable, so we calculate moments of $F_{Y|\eta(.)}(s)$ instead. For the previous two summaries we have been able to make inference about each term of the decomposition independently. For the distribution function, the form of the integral, (5.16), means this is not possible. Therefore, in order to make inference about (5.16) we use an approximation. In section 4.3.1 we found that the posterior distribution of $\eta(\mathbf{x})$ could be well approximated by a t-distribution. We use the same approximation here,

$$\frac{\eta(\mathbf{x}) - E[\eta(\mathbf{x})]}{\hat{\sigma_x}}|\boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r, \mathbf{y}_{(1)}, \ldots \mathbf{y}_{(r)} \sim t_v, \tag{5.17}$$

where, $E[\eta(\mathbf{x})]$, v and $\hat{\sigma}_x$ were given in (4.12), (4.19) and (4.20).

The first two posterior moments of $F_{Y|\eta(.)}(s)$, are almost identical to those from chapter 3:

$$\begin{aligned}
& E^*\{F_{Y|\eta(.)}(s)|\boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r, \mathbf{y}_{(1)}, \ldots \mathbf{y}_{(r)}\} \\
= & \int_{\mathcal{X}} E^*[I\{\eta(\mathbf{x}) \leq s\}|\boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r, \mathbf{y}_{(1)}, \ldots \mathbf{y}_{(r)}]\, dG(\mathbf{x}) \\
= & \int_{\mathcal{X}} P[\{\frac{\eta(\mathbf{x}) - E[\eta(\mathbf{x})]}{\hat{\sigma}_x} \leq \frac{s - E[\eta(\mathbf{x})]}{\hat{\sigma}_x}\}|\boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r, \mathbf{y}_{(1)}, \ldots \mathbf{y}_{(r)}]\, dG(\mathbf{x}) \\
= & \int_{\mathcal{X}} \int_{-\infty}^{\frac{s-E[\eta(\mathbf{X})]}{\hat{\sigma}_X}} f_{T_v}\, dt\, dG(\mathbf{x}),
\end{aligned} \tag{5.18}$$

where f_{T_v} is the density of a t-distribution with v degrees of freedom.

Using the result

$$P\{\eta(\mathbf{z}) \leq s_2\}P\{\eta(\mathbf{x}) \leq s_1|\eta(\mathbf{z}) \leq s_2\} = \int_{-\infty}^{s_2} P\{\eta(\mathbf{x}) \leq s_1|\eta(\mathbf{z}) = k\}f_{\eta(\mathbf{z})}(k)\, dk, \tag{5.19}$$

where $f_{\eta(\mathbf{z})}(k)$ is the density function of $\eta(\mathbf{z})$, we arrive at

$$\begin{aligned}
& E^*\{F_Y(s_1)F_Y(s_2)|\boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_r, \mathbf{y}_{(1)}, \ldots \mathbf{y}_{(r)}\} \\
= & \int_{\mathcal{X}} \int_{\mathcal{X}} \int_{-\infty}^{s_2} \int_{-\infty}^{\frac{s_1-E[\eta(\mathbf{x})]}{\hat{\sigma}_X}} f_{T_{v+1}}\, f_{\eta(\mathbf{z})}(k) dt\, dk\, dG(\mathbf{x})\, dG(\mathbf{z}).
\end{aligned} \tag{5.20}$$

We omit some of the intermediate steps in arriving at (5.20), but the calculation follows the method shown in more detail in chapter 3.

We demonstrate our approximation using the test function

$$\begin{aligned}\eta(x_1,x_2) &= \eta_1(x_{(1)})+\eta_2(x_{(2)})\\ &= x_1+\sin(x_1)+\cos(x_2),\end{aligned} \tag{5.21}$$

where $x_1, x_2 \sim N(0,1)$. We make 7 observations of $\eta_1(.)$ and $\eta_2(.)$ and obtain the outputs $\mathbf{y}_{(1)}$ and $\mathbf{y}_{(2)}$ respectively, at design points $D_1 = \{-3,-2,-1,0,1,2,3\}$ and $D_2 = \{-3,-2,-1,0,1,2,3\}$. In *Figure* (5.1) we plot the distribution function, and percentiles using (5.18). We note the error in using this approximation is very small.

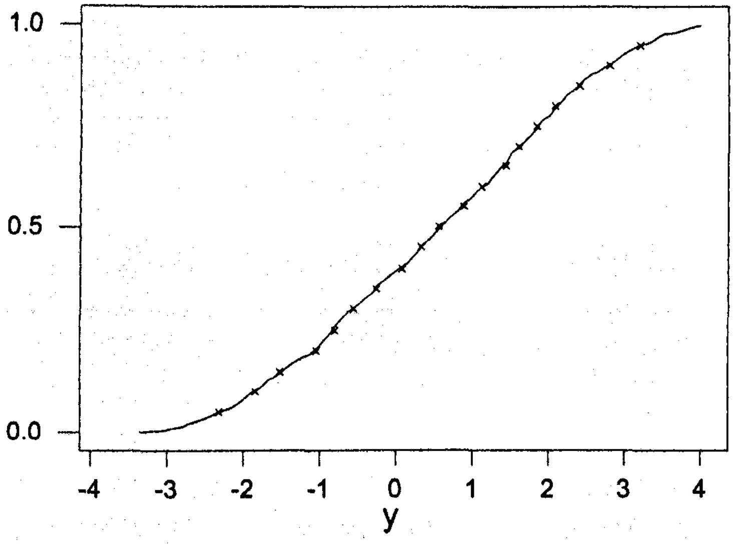

Figure 5.1: Distribution function (line) and percentiles (crosses)

In section 3.3, we briefly mentioned the simulation method developed by *Oakley and O'Hagan*(49) that could be used to evaluate the first two moments of the distribution function. The method can be easily generalized for a model of this form, and is efficient.

5.2.2 Sensitivity

We wish to learn the importance of some sub-vector $\mathbf{X_r}$, which we assess by a decomposition of $\eta(.)$ into main effects and interactions, and a decomposition of the variance.

We first consider conditional expectations of the form, $K_{1,r} = E(Y|\mathbf{X_r} = \mathbf{x_r}, \eta(.))$, where $\mathbf{X_r}$ is a sub vector of $\mathbf{X}$. The sub vector $\mathbf{X_r}$ is quite distinct from the groups $\mathbf{X}_{(1)}, \ldots \mathbf{X}_{(r)}$.

The expectation, $K_{1,r}$, can be decomposed as

$$K_{1,r} = E[(Y_1|\mathbf{X_r} = \mathbf{x_r}), \eta_1(.)] + \ldots + E[(Y_r|\mathbf{X_r} = \mathbf{x_r}), \eta_r(.)], \tag{5.22}$$

and the j^{th} term of (5.22) can be expressed as the integral

$$K_{1,r,(j)} = E(Y_j|\mathbf{X_r} = \mathbf{x_r}, \eta_j(.)) = \int_{\chi_{-r}} \eta_j(\mathbf{x}_{(j)})\, dG_{\mathbf{x_{-r}}|\mathbf{x_r}}(\mathbf{x_{-r}}|\mathbf{x_r}), \tag{5.23}$$

where χ_{-r} denotes the space of possible values for $\mathbf{x_{-r}}$, and $G_{\mathbf{x_{-r}}|\mathbf{x_r}}$ denotes the conditional distribution of $\mathbf{X_{-r}}$ given $\mathbf{X_r}$.

Following from the methodology of chapter 3, quantities $K_{1,r,(j)}$ have t-distributions,

$$\frac{K_{1,r,(j)} - \hat{k}_{1,r,(j)}}{\hat{\sigma}_j \sqrt{\frac{n_j - q_j - 2}{n_j - q_j} W_{r,r,(j)}}} \mid \boldsymbol{\omega}_j, \mathbf{y}_{(j)} \sim t_{n_j - q_j}. \tag{5.24}$$

The expectation and variance of (5.24) require us to evaluate integrals $\mathbf{R}_{\mathbf{r},(j)}$, $\mathbf{T}_{\mathbf{r},(j)}$ and $U_{\mathbf{r},\mathbf{r},(j)}(\mathbf{x_r}, \mathbf{x}'_\mathbf{s})$ which are almost identical to (3.69)(3.70) and (3.72) from chapter 3, and $W_{r,r,(j)}$ is given by a similar calculation to (3.71).

Conditional expectation $K_{1,r}$ is a sum of t-distributions, and we approximate this in the same manner that we approximated the expectation in 5.1.1.

Main Effects and Interactions

We now consider the decomposition of $\eta(.)$ into main effects and interactions.

$$y = \eta(\mathbf{x}) = E\left(Y\right) + \sum_{i=1}^{p} z_i(x_i) + \sum_{1 \leq i \leq j} z_{ij}(x_i, x_j) + \ldots + z_{1,\ldots,p}(\mathbf{x}), \tag{5.25}$$

where

$$\begin{aligned} z_i(x_i) &= E(Y|x_i) - E(Y), & (5.26) \\ z_{i,j}(x_i, x_j) &= E(Y|x_i, x_j) - z_i(x_i) - z_j(x_j) - E(Y), & (5.27) \end{aligned}$$

with higher order interactions following similarly. Note that we don't explicitly show the conditioning on $\eta(.)$ for ease of notation, but all the expressions in (5.25) are conditional on $\eta(.)$.

For a decomposable model we have already shown how to calculate the posterior distribution of expectations $E(Y|\mathbf{X_r} = \mathbf{x_r})$. Main effects and interactions are simply functions of these expectations. Since we have shown that expectations may be written in the form $E(Y|\mathbf{X_r} = \mathbf{x_r}) = \sum_{j=1}^{r} E(Y_{(j)}|\mathbf{X_r} = \mathbf{x_r})$, main effects and interactions can be written in a similar form for example $z_i(x_i)$ may be written as

$$z_i(x_i) = \sum_{j=1}^{r} \{E(Y_{(j)}|x_i) - E(Y_{(j)})\}. \tag{5.28}$$

Each of the r terms in (5.28) has a t-distribution, and $z_i(x_i)$ may be approximated by a sum of t-distributions, as seen earlier. Interaction terms can also be expressed using sums.

From the definition of the conditional expectation, $E(Y_{(j)}|\mathbf{X_r} = \mathbf{x_r})$, we can see this calculation depends on the full vector of inputs. Therefore, even if inputs x_i and x_j are non interacting (do not appear in the same term of our decomposition), it does not follow that the corresponding interaction effect is zero. We demonstrate using the simple example

$$Y = \eta(X) = X_1 + X_2, \tag{5.29}$$

where X_1 and X_2 are normally distributed with respective expectations of μ_1 and μ_2, respective variances σ_1^2 and σ_2^2, and ρ is the correlation between X_1 and X_2.

We find the expectation and conditional expectations as

$$\begin{aligned} E[Y] &= \mu_1 + \mu_2, \\ E[Y|X_1] &= X_1 + \mu_2 - \frac{\sigma_2}{\sigma_1}\rho(X_1 - \mu_1), \\ E[Y|X_2] &= X_2 + \mu_1 - \frac{\sigma_1}{\sigma_2}\rho(X_2 - \mu_2), \end{aligned}$$

and hence we have main effects

$$z_1(x_1) = X_1 - \mu_1 - \frac{\sigma_2}{\sigma_1}\rho(X_1 - \mu_1),$$
$$z_2(x_2) = X_2 - \mu_2 - \frac{\sigma_1}{\sigma_2}\rho(X_2 - \mu_2),$$

and interaction effect

$$z_{1,2}(x_1, x_2) = \frac{\sigma_2}{\sigma_1}\rho(X_1 - \mu_1) + \frac{\sigma_1}{\sigma_2}\rho(X_2 - \mu_2).$$

The interaction term is zero when $\rho = 0$. Therefore we see additivity combined with the additional property of independence are sufficient conditions for the interaction effect to be zero $\forall x_1, x_2$. For a more general model, with $p > 2$ inputs the additional assumption of independence of x_i and x_j is not sufficient to ensure $z_{i,j}(x_i, x_j) = 0 \, \forall x_i, x_j$. We also need to consider the distribution of the other inputs. However, under certain conditions we can guarantee that if inputs are non interacting, the corresponding interaction is exactly zero. We now discuss some conditions under which we can show the main effect $z_i(x_i)$ and first order interaction depend only upon subset S_k, and conditions under which the first order interaction $z_{i,j}(x_i, x_j) = 0 \forall x_i, x_j$.

Suppose we have an additive partition of the inputs, and the joint distribution of the inputs, $G(\mathbf{x})$, can be partitioned into independent components $G(\mathbf{x}_{(1)}), \ldots, G(\mathbf{x}_{(r)})$.

Beginning with the main effect, if input x_i is in the k^{th} group, that is $i \epsilon S_k$, then

$$\begin{aligned} z_i(x_i) &= E(Y|x_i) - E(Y), \\ &= E(Y_{(1)}|x_i) + \ldots + E(Y_{(r)}|x_i) - \{E(Y_{(1)}) + \ldots + E(Y_{(r)})\}, \\ &= E(Y_{(k)}|x_i) + \sum_{n \neq k} E(Y_{(n)}) - \{E(Y_{(1)}) + \ldots + E(Y_{(r)})\}, \\ &= E(Y_{(k)}|x_i) - E(Y_{(k)}), \end{aligned} \tag{5.30}$$

and $z_i(x_i)|\eta(.)$, depends only upon subset S_k.

Now consider the first order interaction between inputs x_i and x_j.

$$\begin{aligned} z_{i,j}(x_i, x_j) &= E(Y|x_i, x_j) - z_i(x_i) - z_j(x_j) - E(Y), \\ &= E(Y|x_i, x_j) - \{E(Y|x_i) - E(Y)\} - \{E(Y|x_j) - E(Y)\} - E(Y). \end{aligned}$$

If $\{i, j\} \epsilon S_k$, then $z_{i,j}(x_i, x_j)$ reduces to

$$\begin{aligned} z_{i,j}(x_i, x_j) &= E(Y_{(k)}|x_i, x_j) + \sum_{n \neq k} E(Y_{(n)}) - \{E(Y_{(k)}|x_i) - E(Y_{(k)})\} \\ &\qquad - \{E(Y_{(k)}|x_j) - E(Y_{(k)})\} - \sum_{n=1}^{r} E(Y_{(n)}), \\ &= E(Y_{(k)}|x_i, x_j) - E(Y_{(k)}|x_i) - E(Y_{(k)}|x_j) + E(Y_{(k)}). \end{aligned} \tag{5.31}$$

Our uncertainty once again depends only upon the subset S_k.

If $i \epsilon S_k$ and $j \epsilon S_l$, then we have

$$\begin{aligned} z_{i,j}(x_i, x_j) &= \sum_{n=1}^{r} E(Y_{(n)}|x_i, x_j) - \sum_{n=1}^{r}\{E(Y_{(n)}|x_i) - E(Y_{(n)})\} \\ &\quad - \sum_{n=1}^{r}\{E(Y_{(n)}|x_j) - E(Y_{(n)})\} - \sum_{n=1}^{r} E(Y_{(n)}), \\ &= E(Y_{(k)}|x_i) + E(Y_{(l)}|x_j) + \sum_{n \neq k,l} E(Y_{(n)}) - \{E(Y_{(k)}|x_i) - E(Y_{(k)})\} \\ &\quad - \{E(Y_{(l)}|x_j) - E(Y_{(l)})\} - \sum_{n=1}^{r} E(Y_{(n)}) = 0 \end{aligned} \tag{5.32}$$

That is, our interaction is exactly zero, with no uncertainty. We can extend this result for higher order interaction terms. It is straightforward to show that under our assumptions about the form of $G(\mathbf{x})$, any interaction effect is zero unless all inputs are contained within the same additive group.

For a partially additive decomposition we will in general require greater independencies of the inputs in order to simplify the main effects and interactions. If $G(\mathbf{x})$ can be partitioned into $G(x_i)$ and $G(\mathbf{x}_{-\mathbf{i}})$ with X_i, $\mathbf{X}_{-\mathbf{i}}$ independent, then $z_i(x_i)$ will depend only upon the subsets of S containing i. If $G(\mathbf{x})$ can be partitioned into $G(x_i)$, $G(x_j)$ and $G(\mathbf{x}_{-\mathbf{ij}})$, then $z_{ij}(x_i, x_j) = 0$ if $\nexists k$ such that $\{i, j\} \subset S_k$. This latter result has an obvious extension to higher order interactions.

Variances

Finally, we consider the decomposition of variance. By independence, the variance of the output conditional on sub vector $\mathbf{X}_\mathbf{r}$, can be written as

$$Var\{E(Y\,|\mathbf{X}_\mathbf{r})\} = Var\{E(Y_{(1)}\,|\mathbf{X}_\mathbf{r})\} + \ldots + Var\{E(Y_{(r)}\,|\mathbf{X}_\mathbf{r})\}, \tag{5.33}$$

where

$$Var\{E(Y_{(j)}\,|\mathbf{X_r})\} = E\{E(Y_{(j)}\,|\mathbf{X_r})^2\} - E(Y_{(j)})^2. \qquad (5.34)$$

As we discussed in chapter 3, the posterior distribution of $E\{E(Y_{(j)}\,|\mathbf{X_r})^2\}$ is intractable however we can calculate posterior moments. By a straightforward adaptation of the methodology of section 3.4 we can calculate the posterior expectation, with respect to the posterior distribution of $\eta_j(.)$, of $E\{E(Y_{(j)}\,|\mathbf{X_r})^2\}$.

The posterior expectation of $Var\{E(Y_{(j)}\,|\mathbf{X_r})\}$ is given by

$$\begin{aligned} E^*\{Var\{E(Y_{(j)}\,|\mathbf{X_r})\}|\omega_j, \mathbf{y}_{(j)}\} &= E^*\{E\{E(Y_{(j)}\,|\mathbf{X_r})^2\}|\omega_j, \mathbf{y}_{(j)}\} \\ &\quad -(Var^*[K_{1,(j)}|\omega_j, \mathbf{y}_{(j)}] + E^*[K_{1,(j)}|\omega_j, \mathbf{y}_{(j)}]^2), \qquad (5.35) \end{aligned}$$

where $E^*\{E\{E(Y_{(j)}\,|\mathbf{X_r})^2\}|\omega_j, \mathbf{y}_{(j)}\}$ requires a similar calculation to (3.77).

The variance of the output, Y, conditional on sub vector $\mathbf{X_r}$ is found from

$$\begin{aligned} &E^*\{Var\{E(Y\,|\mathbf{X_r})\}|\omega_1, \ldots \omega_r, \mathbf{y}_{(1)}, \ldots \mathbf{y}_{(r)}\} \\ &\quad = \sum_{j=1}^{r} E^*\{Var\{E(Y_{(j)}\,|\mathbf{X_r})\}|\omega_j, \mathbf{y}_{(j)}\}. \end{aligned}$$

We are able to calculate sensitivity indices by dividing these partial variances by $E^*\{Var(Y)|\omega_1, \ldots \omega_r, \mathbf{y}_{(1)}, \ldots \mathbf{y}_{(r)}\}$.

For additive models, if $G(\mathbf{x})$ can be partitioned into independent components $G(\mathbf{x}_{(1)}), \ldots, G(\mathbf{x}_{(r)})$, then $E^*\{Var\{E(Y_{(j)}\,|\mathbf{X_r})\}|\omega_j, \mathbf{y}_{(j)}\} = 0$ unless $\mathbf{X}_{(j)}$ and $\mathbf{X_r}$ contain at least one common element. We will require greater independencies in order to simplify partially additive models. If we have complete independence of the inputs, the calculations for variance based sensitivity indices can be vastly simplified for both additive and partially additive models.

5.2.3 Inference for Unobservable Functions

We briefly consider inference for the unobservable functions model that we examined in section 4.3.2. This differed from the simple additive model in that our posterior expectation had the additional constant term, c. As a result, for this model $E^*\left[K_{1,(j)}\,|\boldsymbol{\omega}_1,\ldots\boldsymbol{\omega}_r,\mathbf{y}_{(1)},\ldots\mathbf{y}_{(r)}\right]$ and $E^*\{F_Y(s)|\boldsymbol{\omega}_1,\ldots\boldsymbol{\omega}_r,\mathbf{y}_{(1)},\ldots\mathbf{y}_{(r)}\}$ will be inflated by c. The variance, main effects and interactions and partial variances for this model are identical to the values given earlier on in this section.

5.3 Inference for Unknown Decompositions

We once more consider a decomposition of $\eta(\mathbf{x})$ into

$$\eta(\mathbf{x}) = \eta_1(\mathbf{x}_{(1)}) + \ldots + \eta_r(\mathbf{x}_{(r)}). \tag{5.36}$$

In this second case, the subsets S_j were unknown a priori. In section 4.5 we used a space filling design and modelled $\eta(.)$ with an additive covariance structure. In sections 4.6 and 4.7 we used the data, $\mathbf{y}$ to identify the subsets S_j.

In deriving measures of uncertainty and sensitivity for this class of decomposition, we assume that the decomposition of $\eta(.)$ that we found in chapter 4 is correct although certainly for the partially additive case further decomposition of $\eta(.)$ may be possible.

5.3.1 Uncertainty

We begin by calculating measures of uncertainty – the expectation, variance and distribution function of $Y|\eta(.)$.

Expectation

We let $K_1^{(2)}$ denote $E[Y|\eta(.)]$. Following the methodology of section 3.3, we calculate the posterior distribution of $K_1^{(2)}$ as

$$\frac{K_1^{(2)} - \hat{k}^{(2)}}{\hat{\sigma}^{(2)}\sqrt{\frac{n-q-2}{n-q}W^{(2)}}} \mid \mathbf{y}, \omega_1, \ldots, \omega_r \sim t_{n-q}, \tag{5.37}$$

where

$$\begin{aligned} E^*\left[K_1^{(2)}|\sigma^2, \omega_1, \ldots, \omega_r, \mathbf{y}\right] &= \hat{k}^{(2)} = \int_{\mathcal{X}} \mathbf{m}^{(2)**}(\mathbf{x})\, dG(\mathbf{x}) \\ &= \mathbf{R}\hat{\boldsymbol{\beta}}^{(2)} + \mathbf{T}^{(2)}\mathbf{A}^{**-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}^{(2)}), \qquad (5.38) \\ Var^*\left[K_1^{(2)}|\sigma^2, \omega_1, \ldots, \omega_r, \mathbf{y}\right] &= \sigma^{2(2)}W^{(2)} = \sigma^{2(2)} \int_{\mathcal{X}}\int_{\mathcal{X}} \mathbf{c}^{(2)**}(\mathbf{x}, \mathbf{x}')\, dG(\mathbf{x})\, dG(\mathbf{x}'), \\ &= \sigma^{2(2)}\{U^{(2)} - \mathbf{T}^{(2)}\mathbf{A}^{**-1}\mathbf{T}^{(2)T} + (\mathbf{R} - \mathbf{T}^{(2)}\mathbf{A}^{**-1}\mathbf{H}) \\ &\quad \times (\mathbf{H}^{\mathbf{T}}\mathbf{A}^{**-1}\mathbf{H})^{-1}(\mathbf{R} - \mathbf{T}^{(2)}\mathbf{A}^{**-1}\mathbf{H})^T\}. \qquad (5.39) \end{aligned}$$

Removing the dependency on σ^2 results in σ^2 being replaced by $\hat{\sigma}^{(2)}\frac{n-q-2}{n-q}$ in the variance. The quantities $\mathbf{A}^{**}, \hat{\boldsymbol{\beta}}^{(2)}, \mathbf{m}^{(2)**}(\mathbf{x})$ and $\mathbf{c}^{(2)**}(\mathbf{x}, \mathbf{x}')$ are defined in equations (4.56)-(4.60) of chapter 4. The latter 2 terms, $\mathbf{m}^{(2)**}(\mathbf{x})$ and $\hat{\sigma}^{2(2)}\mathbf{c}^{(2)**}(\mathbf{x}, \mathbf{x}')$ are the standard posterior mean and covariance functions in the Gaussian Process model, but derived from the alternative prior correlation function (4.53).

We defined the integral $\mathbf{R}$ in equation (3.51) of chapter 3. However, $\mathbf{T}^{(2)}$ and $U^{(2)}$ differ from $\mathbf{T}$ and U (defined in equations (3.52)-(3.62) since they are functions of $\mathbf{c}^{(2)**}(.,.)$. These require us to evaluate the integrals

$$\begin{aligned} \mathbf{T}^{(2)} &= \int_{\mathcal{X}} \mathbf{t}^{(2)}(\mathbf{x})^T dG(\mathbf{x}), \qquad (5.40) \\ U^{(2)} &= \int_{\mathcal{X}}\int_{\mathcal{X}} \mathbf{c}^{(2)}(\mathbf{x}, \mathbf{x}')\, dG(\mathbf{x})\, dG(\mathbf{x}'). \qquad (5.41) \end{aligned}$$

Both $\mathbf{T}^{(2)}$ and $U^{(2)}$ may be decomposed into a sum of r terms. From (4.58) we have

$$\mathbf{t}^{(2)}(\mathbf{x}) = \{\mathbf{c}^{(2)}(\mathbf{x}, \mathbf{x}_1), \ldots, \mathbf{c}^{(2)}(\mathbf{x}, \mathbf{x}_n)\}, \tag{5.42}$$

and from (4.53)

$$Cov[\eta(\mathbf{x}) \, , \, \eta(\mathbf{x}')] = \sigma^2 \mathbf{c}^{(2)}(\mathbf{x}, \mathbf{x}') = \sigma^2 \{\mathbf{c}_1(\mathbf{x}_{(1)}, \mathbf{x}'_{(1)}) + \ldots + \mathbf{c}_r(\mathbf{x}_{(r)}, \mathbf{x}'_{(r)})\}. \tag{5.43}$$

Therefore, we can write

$$\mathbf{T}^{(2)} = \mathbf{T}^{(2)}_{(1)} + \ldots \mathbf{T}^{(2)}_{(r)}, \tag{5.44}$$

where the m^{th} element of $\mathbf{T}^{(2)}_{(j)}$ is given by

$$\mathbf{T}^{(2)}_{(j)}[m] = \int_{\mathcal{X}} \mathbf{c}_\mathrm{j}(\mathbf{x}_{(j)} \mathbf{x}_{(j),m}) \, dG(\mathbf{x}), \tag{5.45}$$

and $\mathbf{x}_{(j),m}$ is the the sub-vector $\mathbf{x}_{(j)}$ from the m^{th} design point.

The scalar $U^{(2)}$ is a sum of m terms, $U^{(2)} = U^{(2)}_{(1)} + \ldots + U^{(2)}_{(r)}$. The m^{th} term of $U^{(2)}$ requires us to evaluate the integral

$$U^{(2)}_m = \int_{\mathcal{X}} \int_{\mathcal{X}} \mathbf{c}_\mathrm{m}(\mathbf{x}_{(m)}, \mathbf{x}'_{(m)}) \; dG(\mathbf{x}) \; dG(\mathbf{x}'). \tag{5.46}$$

Variance

For the variance of Y conditional on $\eta(.)$ we require the posterior distribution of $K_2^{(2)} = \int_{\mathcal{X}} \eta(\mathbf{x})^2 dG(\mathbf{x})$. As we found in chapter 3, this form is intractable. Therefore, we just calculate the first posterior moment of $Var[Y|\eta(.)]$, as done previously (in both sections 3.3 and 5.1)

We have

$$\begin{aligned} & E^*[K_2^{(2)}|\omega_1,\ldots,\omega_r,\mathbf{y}] \\ = \;& tr(\hat{\boldsymbol{\beta}}^{(2)T}\mathbf{Q}\hat{\boldsymbol{\beta}}^{(2)}) + tr((\mathbf{y}-\mathbf{H}\hat{\boldsymbol{\beta}}^{(2)})^T\mathbf{A}^{**-1}\mathbf{P}^{(2)}\mathbf{A}^{**-1}(\mathbf{y}-\mathbf{H}\hat{\boldsymbol{\beta}}^{(2)}) \\ + \;& 2tr(\hat{\boldsymbol{\beta}}^{(2)}\mathbf{T}^{(2)T}\mathbf{R}\mathbf{A}^{**-1}(\mathbf{y}-\mathbf{H}\hat{\boldsymbol{\beta}}^{(2)}) + \hat{\sigma}^{2(2)}[1 - tr(\mathbf{A}^{**-1}\mathbf{P}^{(2)}) \\ + \;& tr(\mathbf{H}^T\mathbf{A}^{**-1}\mathbf{H})^{-1}\mathbf{Q} - 2tr((\mathbf{H}^T\mathbf{A}^{**-1}\mathbf{H})^{-1}\mathbf{S}^{(2)}\mathbf{A}^{**-1}\mathbf{H}) \\ + \;& tr((\mathbf{H}^T\mathbf{A}^{**-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{A}^{**-1}\mathbf{P}^{(2)}\mathbf{A}^{**-1}\mathbf{H})]. \end{aligned} \tag{5.47}$$

where E^* denotes the expectation with respect to the posterior distribution of $\eta(.)$. Expressions $\mathbf{P}^{(2)}$ and $\mathbf{S}^{(2)}$ require us to evaluate the integrals

$$\mathbf{P}^{(2)} = \int_{\mathcal{X}} \mathbf{t}^{(2)}(\mathbf{x})\mathbf{t}^{(2)}(\mathbf{x})^{\mathbf{T}}\, dG(\mathbf{x}), \tag{5.48}$$

$$\mathbf{S}^{(2)} = \int_{\mathcal{X}} \mathbf{t}^{(2)}(\mathbf{x})\mathbf{h}(\mathbf{x})^{\mathbf{T}}\, dG(\mathbf{x}). \tag{5.49}$$

The above expressions differ from $\mathbf{P}$ and $\mathbf{S}$ (equations (3.58) and (3.60)) since they are functions of $\mathbf{c}^{(2)**}(.,.)$. Resultantly, both of these quantities can be expanded. The $[i,j]^{th}$ element of $\mathbf{P}^{(2)}$ can be written as

$$\mathbf{P}^{(2)}[i,j] = \mathbf{c}^{(2)**}(\mathbf{x},\mathbf{x}_i)\,\mathbf{c}^{(2)**}(\mathbf{x},\mathbf{x}_j)\,dG(\mathbf{x}), \tag{5.50}$$

and from (5.43), this can be expanded into a sum of r^2 terms. Similarly each element of $\mathbf{S}^{(2)}$ can be expanded into a sum of r terms.

We find the posterior expectation of $Var\{Y|\eta(.)\}$ as

$$\begin{aligned} E^*[Var\{Y\}|\omega_1,\ldots\omega_r,\mathbf{y}] &= E^*[K_2^{(2)}|\omega_1,\ldots\omega_r,\mathbf{y}] \\ &- (Var^*[K_1^{(2)}|\omega_1,\ldots\omega_r,\mathbf{y}] + E^*[K_1^{(2)}|\omega_1,\ldots\omega_r,\mathbf{y}]^2). \end{aligned}$$

Distribution Function

The first 2 moments of $F_{Y|\eta(.)}(s)|\omega_1, \ldots, \omega_r, \mathbf{y}$ can be calculated from equations (3.73) and (3.75). We simply need to substitute $\mathbf{m}^{(2)**}(\mathbf{X})$ and $\mathbf{c}^{(2)**}(\mathbf{X}, \mathbf{X})$ for $\mathbf{m}^{**}(\mathbf{X})$ and $\mathbf{c}^{**}(\mathbf{X}, \mathbf{X})$ in equations (3.63) and (3.65). The moments have to be evaluated numerically, however in this case we are able to calculate these efficiently by utilizing the simulation method of *Oakley and O'Hagan*(49).

5.3.2 Sensitivity

We first consider inference for conditional expectations, $E(Y|\mathbf{X_r} = \mathbf{x_r}, \eta(.))$, before considering the decomposition of $\eta(\mathbf{x})$ into main effects and interactions, and a decomposition of the variance.

The posterior distribution of $K^{(2)}_{1,r} = E(Y|\mathbf{X_r} = \mathbf{x_r}, \eta(.))$, where we use the subscript $\mathbf{r}$ to denote the expectation is conditional on $\mathbf{X_r} = \mathbf{x_r}$, is a t-distribution,

$$\frac{K^{(2)}_{1,r} - \hat{k}^{(2)}_{1,r}}{\hat{\sigma}^{(2)}\sqrt{\frac{n-q-2}{n-q} W^{(2)}_{r,r}}} \,|\, \omega_1, \ldots, \omega_r, \mathbf{y} \sim t_{n-q}. \tag{5.51}$$

where

$$\hat{k}^{(2)}_{1,r} = E^*\{K^{(2)}_{1,r}|\sigma^2, \omega_1, \ldots, \omega_r, \mathbf{y}\} = \mathbf{R_r}(\mathbf{x_r})\hat{\boldsymbol{\beta}}^{(2)} + \mathbf{T}^{(2)}_{\mathbf{r}}(\mathbf{x_r})\mathbf{A}^{**-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}^{(2)}), \tag{5.52}$$

and

$$\begin{aligned} & Cov^*\{K^{(2)}_{1,r}\, K^{(2)}_{1,s}|\sigma^2, \omega_1, \ldots, \omega_r, \mathbf{y}\} \\ = \;& \sigma^{2(2)} \int_{\chi-r} \int_{\chi-s} \mathbf{c}^{(2)**}(\mathbf{x}, \mathbf{x}')\; dG_{-r\,|\,r}(\mathbf{x_{-r}}\,|\,\mathbf{x_r})\, dG_{-s\,|\,s}(\mathbf{x}'_{\mathbf{-s}}\,|\,\mathbf{x}'_{\mathbf{s}}) \\ = \;& \sigma^{2(2)} W^{(2)}_{r,s} = \sigma^{2(2)}\{U^{(2)}_{r;s}(\mathbf{x_r}, \mathbf{x}'_{\mathbf{s}}) - \mathbf{T}^{(2)}{}_{\mathbf{r}}(\mathbf{x_r})A^{**-1}\mathbf{T}^{(2)}{}_{\mathbf{s}}(\mathbf{x_s})^{\mathbf{T}} + (\mathbf{R_r}(\mathbf{x_r}) \\ - \;& \mathbf{T}^{(2)}{}_{\mathbf{r}}(\mathbf{x_r})\mathbf{A}^{**-1}\mathbf{H})(\mathbf{H^T}\mathbf{A}^{**-1}\mathbf{H})^{-1}(\mathbf{R_s}(\mathbf{x_s}) - \mathbf{T}^{(2)}{}_{\mathbf{s}}(\mathbf{x_s})\mathbf{A}^{**-1}\mathbf{H})^{T}\}. \end{aligned} \tag{5.53}$$

Quantities $\mathbf{T}_{\mathbf{r}}^{(2)}(\mathbf{x}_{\mathbf{r}})$ and $U_{r;s}^{(2)}(\mathbf{x}_{\mathbf{r}}, \mathbf{x}_{\mathbf{s}}^{'})$ require us to evaluate

$$\begin{aligned} \mathbf{T}_{\mathbf{r}}^{(2)}(\mathbf{x}_{\mathbf{r}}) &= \int_{\mathcal{X}_{-r}} \mathbf{t}^{(2)}(\mathbf{x})^{T} dG_{-r\,|\,r}(\mathbf{x}_{-\mathbf{r}}\,|\,\mathbf{x}_{\mathbf{r}}), \qquad (5.54) \\ U_{r;s}^{(2)}(\mathbf{x}_{\mathbf{r}}, \mathbf{x}_{\mathbf{s}}^{'}) &= \int_{\mathcal{X}_{-r}} \int_{\mathcal{X}_{-s}} \mathbf{c}^{(2)}(\mathbf{x}, \mathbf{x}^{'})\; dG_{-r\,|\,r}(\mathbf{x}_{-\mathbf{r}}\,|\,\mathbf{x}_{\mathbf{r}})\, dG_{-s\,|\,s}(\mathbf{x}_{-\mathbf{s}}^{'}\,|\,\mathbf{x}_{\mathbf{s}}^{'})(5.55) \end{aligned}$$

whilst $\mathbf{R}_{\mathbf{r}}(\mathbf{x}_{\mathbf{r}})$ was given in (3.69). Both $\mathbf{T}_{\mathbf{r}}^{(2)}(\mathbf{x}_{\mathbf{r}})$ and $U_{r;s}^{(2)}(\mathbf{x}_{\mathbf{r}}, \mathbf{x}_{\mathbf{s}}^{'})$ can be expanded in the same manner that $\mathbf{T}^{(2)}(\mathbf{x})$ and $U^{(2)}$ could be expanded. We remove the dependency on σ^2, resulting in σ^2 being replaced by $\hat{\sigma}^{(2)}\frac{n-q-2}{n-q}$ in the covariance.

Main Effects and Interactions

We now consider inference for the main effect and first order interaction

$$\begin{aligned} z_i(x_i) &= E(Y|x_i) - E(Y), \qquad (5.56) \\ z_{i,j}(x_i, x_j) &= E(Y|x_i, x_j) - z_i(x_i) - z_j(x_j) - E(Y), \qquad (5.57) \end{aligned}$$

as we did in 5.1.2. Higher order interactions require similar calculations to these, and are discussed later.

Since $K_{1,r}^{(2)}|\sigma^2\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r, \mathbf{y}$ is normally distributed for any $\mathbf{r}$, it is simple to note that conditional on σ^2 the main effects and interactions are linear functions of correlated normal distributions.

The posterior expectations of (5.56) and (5.57) are given by

$$\begin{aligned} & E^{*}\{z_i(x_i)|\sigma^2, \boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r, \mathbf{y}\} = \{\mathbf{R}_i(x_i) - \mathbf{R}\}\hat{\boldsymbol{\beta}}^{(2)} \\ & \qquad + \{\mathbf{T}_i^{(2)}(x_i) - \mathbf{T}^{(2)}\}\mathbf{A}^{**-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}^{(2)}), \qquad (5.58) \\ & E^{*}\{z_{ij}(x_i, x_j)|\sigma^2, \boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r, \mathbf{y}\} = \{\mathbf{R}_{ij}(x_i, x_j) - \mathbf{R}_i(x_i) - \mathbf{R}_j(x_j) + \mathbf{R}\}\hat{\boldsymbol{\beta}}^{(2)} \\ & \qquad + \{\mathbf{T}_{ij}^{(2)}(x_i, x_j) - \mathbf{T}_i^{(2)}(x_i) - \mathbf{T}_j^{(2)}(x_j) + \mathbf{T}^{(2)}\}\mathbf{A}^{**-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}^{(2)}), \quad (5.59) \end{aligned}$$

whilst the posterior variances of these terms are calculated using (5.53) as

$$\begin{aligned}
& Var^*\{z_i(x_i)|\sigma^2,\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_r,\mathbf{y}\} \\
= \; & \sigma^{2(2)}\{U^{(2)} + U^{(2)}_{i,i} - 2U^{(2)}_{0,i} - [\mathbf{T}^{(2)}_i(x_i) - \mathbf{T}^{(2)}]\mathbf{A}^{**-1}[\mathbf{T}^{(2)}_i(x_i) - \mathbf{T}^{(2)}]^T \\
& - ([\mathbf{R}_i(x_i) - \mathbf{R}] - [\mathbf{T}^{(2)}_i(x_i) - \mathbf{T}^{(2)}]\mathbf{A}^{**-1}\mathbf{H})(\mathbf{H}^\mathbf{T}\mathbf{A}^{**-1}\mathbf{H})^{-1} \\
& \quad \times([\mathbf{R}_i(x_i) - \mathbf{R}] - [\mathbf{T}^{(2)}_i(x_i) - \mathbf{T}^{(2)}]\mathbf{A}^{**-1}\mathbf{H})^T\}, \qquad (5.60)
\end{aligned}$$

and

$$\begin{aligned}
& Var^*\{z_{ij}(x_i, x_j)|\sigma^2,\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_r,\mathbf{y}\} = \\
& \sigma^{2(2)}\{U^{(2)*} - [\mathbf{T}^{(2)}_{ij}(x_i,x_j) - \mathbf{T}^{(2)}_i(x_i) - \mathbf{T}^{(2)}_j(x_j) + \mathbf{T}^{(2)}]\mathbf{A}^{**-1}[\mathbf{T}^{(2)}_{ij}(x_i,x_j) - \mathbf{T}^{(2)}_i(x_i) \\
& -\mathbf{T}^{(2)}_j(x_j) + \mathbf{T}^{(2)}]^T - ([\mathbf{R}_{ij}(x_i,x_j) - \mathbf{R}_i(x_i) - \mathbf{R}_j(x_j) + \mathbf{R}] - [\mathbf{T}^{(2)}_{ij}(x_i,x_j) \\
& -\mathbf{T}^{(2)}_i(x_i) - \mathbf{T}^{(2)}_j(x_j) + \mathbf{T}^{(2)}]\mathbf{A}^{**-1}\mathbf{H})(\mathbf{H}^\mathbf{T}\mathbf{A}^{**-1}\mathbf{H})^{-1}([\mathbf{R}_{ij}(x_i,x_j) - \mathbf{R}_i(x_i) \\
& -\mathbf{R}_j(x_j) + \mathbf{R}] - [\mathbf{T}^{(2)}_{ij}(x_i,x_j) - \mathbf{T}^{(2)}_i(x_i) - \mathbf{T}^{(2)}_j(x_j) + \mathbf{T}^{(2)}]\mathbf{A}^{**-1}\mathbf{H})^T\}, \qquad (5.61)
\end{aligned}$$

where $U^{(2)*}$ is expanded as

$$\begin{aligned}
U^{(2)*} \; = \; & U^{(2)} + U^{(2)}_{i,i} + U^{(2)}_{j,j} + U^{(2)}_{ij,ij} - 2U^{(2)}_{ij,i} - 2U^{(2)}_{ij,j} \\
& + 2U^{(2)}_{0,ij} + 2U^{(2)}_{i,j} - 2U^{(2)}_{i,0} - 2U^{(2)}_{j,0}. \qquad (5.62)
\end{aligned}$$

All the terms (5.62) are calculated using (5.55), so for example $U^{(2)}_{ij,j}$ is found by letting $\mathbf{x_r} = \{x_i, x_j\}$ and $\mathbf{x_s} = \{x_j\}$ in (5.55). The subscript 0 in the terms in (5.62) denotes the null set, and we therefore have $U^{(2)} = U^{(2)}_{0,0}$.

After removing the dependency on σ^2 the main effects and interactions have t-distributions. The variances (5.60) and (5.61) are the same except with σ^2 replaced by $\hat{\sigma}^{(2)}\frac{n-q-2}{n-q}$.

From (5.54)-(5.55) it is obvious that the integrals required in order to calculate main effects and interactions depend upon the full vector of inputs $\mathbf{x}$. However, under the same independence assumptions that we had in 5.1.2 we can show analogous results to those of the *a priori* known decomposition case. That is, for an additive partition of the inputs, and where the joint distribution of the inputs, $G(\mathbf{x})$, can be partitioned into independent components $G(\mathbf{x}_{(1)}), \ldots, G(\mathbf{x}_{(r)})$, we can show that if $\{i, j\} \epsilon S_k$, then $z_i(x_i)$ and $z_{ij}(x_i,\, x_j)$ are functions of just $\eta_k(.)$. For higher order terms we can show that any interaction effect is zero unless all inputs are contained within the same additive group.

A partially additive decomposition we will in general require almost total independence of the inputs in order to simplify the main effects and interactions. In general if we have $k \leq p$ independent inputs then the corresponding interaction effect is zero, unless all k inputs appear in the same partially additive group.

Variances

Finally we consider partial variances

$$Var\{E(Y\,|\mathbf{X_r})\} = E\{E(Y\,|\mathbf{X_r})^2\} - E(Y)^2. \tag{5.63}$$

We wish to calculate the first posterior moment of $Var\{E(Y\,|\mathbf{X_r})\}$, and for this we require the first posterior moment of $E\{E(Y\,|\mathbf{X_r})^2\}$. By applying the methodology from section 3.4, we find the posterior expectation with respect to the posterior distribution of $\eta(.)$ can be written as

$$\begin{aligned}
&E^*[E\{E(Y\,|\mathbf{X_r})^2\}|\omega_1,\ldots,\omega_r,\mathbf{y}] \\
&= tr((\mathbf{y}-\mathbf{H}\hat{\beta}^{(2)})^T\mathbf{A}^{**-1}\mathbf{P_r^{(2)}}\mathbf{A}^{**-1}(\mathbf{y}-\mathbf{H}\hat{\beta}^{(2)})) + 2tr(\hat{\beta}^{(2)}\mathbf{S_r^{(2)}}\mathbf{A}^{**-1}(\mathbf{y}-\mathbf{H}\hat{\beta}^{(2)})) \\
&+ tr(\hat{\beta}^{(2)T}\mathbf{Q_r^{(2)}}\hat{\beta}^{(2)}) + \sigma^{\hat{(2)}2}[U_r^{(2)} - tr(\mathbf{A}^{**-1}\mathbf{P_r^{(2)}}) + tr((\mathbf{H}^T\mathbf{A}^{**-1}\mathbf{H})^{-1} \\
&\qquad \times (\mathbf{Q_r^{(2)}} - \mathbf{S_r^{(2)}}\mathbf{A}^{**-1}\mathbf{H} - \mathbf{H}^T\mathbf{A}^{**-1}\mathbf{S_r^{(2)T}} + \mathbf{H}^T\mathbf{A}^{**-1}\mathbf{P_r^{(2)}}\mathbf{A}^{**-1}\mathbf{H}))],
\end{aligned} \tag{5.64}$$

where

$$\begin{aligned}
U_{\mathbf{r}}^{(2)} &= \int_{\chi_r}\int_{\chi_{-r}}\int_{\chi_{-r}} \mathbf{c}^{(2)}(\mathbf{x},\mathbf{x}^*)\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}'_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{\mathbf{r}}(\mathbf{x}_{\mathbf{r}}),\\
\mathbf{P}_{\mathbf{r}}^{(2)} &= \int_{\chi_r}\int_{\chi_{-r}}\int_{\chi_{-r}} \mathbf{t}^{(2)}(\mathbf{x})\mathbf{t}^{(2)}(\mathbf{x}^*)^T\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}'_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{\mathbf{r}}(\mathbf{x}_{\mathbf{r}}),\\
\mathbf{Q}_{\mathbf{r}}^{(2)} &= \int_{\chi_r}\int_{\chi_{-r}}\int_{\chi_{-r}} \mathbf{h}(\mathbf{x})\mathbf{h}(\mathbf{x}^*)^{\mathbf{T}}\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}'_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{\mathbf{r}}(\mathbf{x}_{\mathbf{r}}),\\
\mathbf{S}_{\mathbf{r}}^{(2)} &= \int_{\chi_r}\int_{\chi_{-r}}\int_{\chi_{-r}} \mathbf{h}(\mathbf{x})\mathbf{t}^{(2)}(\mathbf{x}^*)^{\mathbf{T}}\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{-\mathbf{r}|\mathbf{r}}(\mathbf{x}'_{-\mathbf{r}}\,|\mathbf{x}_{\mathbf{r}})\, dG_{\mathbf{r}}(\mathbf{x}_{\mathbf{r}}).
\end{aligned}$$

We use $\mathbf{x}^*$ to denote the vector $\mathbf{x}^* = (\mathbf{x}_{\mathbf{r}}, \mathbf{x}'_{-\mathbf{r}})$, whilst $\mathbf{x} = (\mathbf{x}_{\mathbf{r}}, \mathbf{x}_{-\mathbf{r}})$, and $G_{\mathbf{r}}(.)$ denotes the marginal distribution of $\mathbf{X}_{\mathbf{r}}$. The terms $U_{\mathbf{r}}^{(2)}, \mathbf{P}_{\mathbf{r}}^{(2)}$ and $\mathbf{S}_{\mathbf{r}}^{(2)}$ can all be expanded as sums.

Thus, the variance of the output conditional on sub vector $\mathbf{X}_{\mathbf{r}}$ can be found from

$$\begin{aligned}
E^*\{Var\{E(Y\,|\mathbf{X}_{\mathbf{r}})\}|\omega_1,\ldots,\omega_r,\mathbf{y}\} &= E^*\{E\{E(Y\,|\mathbf{X}_{\mathbf{r}})^2\}|\omega_1,\ldots,\omega_r,\mathbf{y}\}\\
&- (Var^*[K_1^{(2)}|\omega_1,\ldots,\omega_r,\mathbf{y}] + E^*[K_1^{(2)}|\omega_1,\ldots,\omega_r,\mathbf{y}]^2). \qquad (5.65)
\end{aligned}$$

We can now calculate sensitivity indices by dividing the partial variances by $E^*\{Var(Y)|\omega_1,\ldots,\omega_r,\mathbf{y}\}$.

Following on from our work on main effects and interactions, we can see that with independencies between inputs these calculations will simplify. In particular, we may be able to infer that some interaction variances are exactly zero.

5.4 Conclusions

In this chapter we have calculated measures of uncertainty and sensitivity for decomposable functions. Our measures were the same as those considered in chapter 3, although we used structural information in this chapter in order to obtain more accurate results, and with less uncertainty about posterior expectations. We showed that in general, even when we have a known decomposition, measures of uncertainty and sensitivity depend on the full vector of inputs. We examined certain independence assumptions, that when combined with a decomposition simplify inference.

Chapter 6

Elicitation of Expert Opinion in Autoregressive Models

6.1 Introduction

Elicitation is the process of formulating a person's subjective knowledge/beliefs about uncertain quantities into a probability distribution. In our work we take the person to be an expert in some field, although not necessarily an expert in probability. The elicited probability distribution should be an accurate representation of the expert's beliefs.

The literature on elicitation is vast and ever growing, and covers both psychological and mathematical considerations. *Garthwaite, Kadane and O'Hagan*(20) recently provided an excellent review of the literature. Whilst we have a great many considerations when designing questions, training experts, and conducting an elicitation, these are beyond the scope of this thesis.

We constrain our attention to the key issue of what summaries to elicit in order to estimate a (joint) distribution, and how to estimate the parameters of this distribution given these assessments. There has been recent progress in a non-parametric approach to distribution fitting, with *Oakley and O'Hagan*(50) modelling the expert's distribution as an unknown function. Within this framework, the expert's distribution is not constrained to be a member of a specified parametric family. In our work, we adopt a simpler approach, considering the case where we have an (assumed) known parametric form for the (joint) distribution, but with unknown parameters.

In section 6.2 we discuss the summaries that we might elicit from an expert, when we have a chosen parametric family of distributions. A priori we have unknown parameters, but we are able to use the expert's assessments to estimate these, such that the chosen distribution from our parametric family closely approximates the expert's stated beliefs. We consider continuous symmetric families. In the remainder of the chapter we concentrate on the more difficult task of eliciting a joint distribution for the parameters of an autoregressive model. We discuss theory in 6.3-6.5, and provide an example in 6.6.

6.2 Expert Judgements

One of the most common summaries to elicit is a measure of central tendency, such as mean, median or mode. The literature (see *Garthwaite, Kadane and O'Hagan*(20) and the references therein) shows that in general experts are able to provide reasonable estimates of these quantities for a symmetric (or at least approximately symmetric) distribution.

In our work we use the median as our measure of central tendency. Suppose we are eliciting our expert's beliefs about an observable quantity X. Then we ask our expert:

1. can you provide a value (median) such that X is equally likely to be less than or greater than this point.

We choose the median because of its simple definition, which a non statistician might reasonably be expected to understand. Moreover, we are able to define quantiles very similarly, and therefore we have consistency in our questions.

For a chosen two parameter family of distributions (with unknown parameters), one more piece of information, conveying information about scale, is enough to estimate the parameters. For a symmetric distribution (especially a normal or t-distribution), the variance is an obvious summary to elicit. However, experts have been shown to be poor at assessing variances directly.

The most common approach to eliciting a variance, is to ask an expert for quantiles or a credible interval, and we are able to infer the expert's variance from these. *Garthwaite and Dickey*(18) found that experts were most comfortable when asked for equal odds judgements, so the interquartile range is a natural quantity to elicit.

We might ask the additional questions:

2a suppose you were told X is below your assessed median. Can you provide a new value (lower quartile) such that X is equally likely to be less than or greater than this value?

2b Suppose you were told X is above your assessed median. Can you provide a new value (upper quartile) such that X is equally likely to be less than or greater than this value?

These questions are an example of the *variable interval method.* The expert provides the points that correspond to specified percentiles of his distribution. We can simplify this task by using a method of bisection, as seen above. An alternative method is the *fixed interval method.* The range of plausible value for X is divided into intervals. The expert then supplies the probability that X lies within each interval. This latter approach is considered to be more difficult since the expert is no longer making equal odds judgements – we also need to know the plausible range for X prior to questioning.

When we are eliciting the parameters of a scaled t-distribution, we face a more difficult task. We have an infinite number of scaled and shifted t-distributions that correspond to the same median and interquartile range. We need additional information about the tails of the distribution, X.

Kadane et al.(35) considered estimating the degrees of freedom parameter, n, by eliciting the expert's median, 75^{th} and 93.75^{th} percentile. This latter percentile is unusual but arises due to the method of repeated bisection proposed by the authors. They formed the tail ratio,

$$\frac{y_{.9375} - y_{.50}}{y_{.75} - y_{.50}}, \tag{6.1}$$

which they compared with a similar ratio based on tabulated values of the t-distribution in order to select n.

In *Figure* (6.1) we plot the ratio, (6.1), for different values of the degrees of freedom parameter n. We see there is clear separation for small values of n, and we can distinguish between the different t-distributions. However, for $n > 5$ it becomes difficult to distinguish between distributions.

In theory we can distinguish between different t-distributions better by eliciting the $\alpha > 0.9375$ percentile in the extreme tails of X. As we take $\alpha \to 1$, we can distinguish between distributions, using a ratio similar to (6.1), even for large values of n. However, we encounter a contradiction between the theory and what

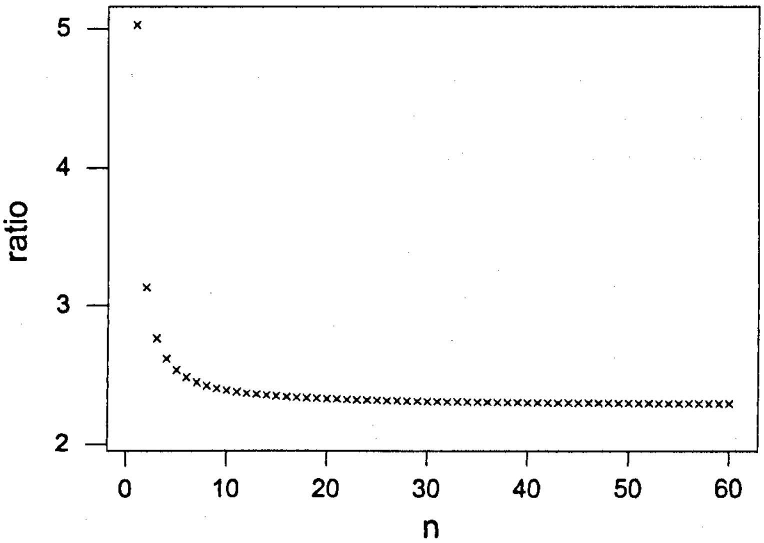

Figure 6.1: Tail ratios of t-distributions

we can achieve in practice.

The empirical work of *Alpert and Raiffa*(2) showed that participants were poor at assessing tail probabilities. The authors undertook fairly large studies in the 1980's, and found that participants were *overconfident* with tail probability assessments. They asked their participants to provide their $\alpha/2$ and $1 - \alpha/2$ percentiles, using various values of α ($\alpha \rightarrow 1$), for quantities like:

'The United States total egg consumption last year.'

For a calibrated participant, if we ask for a credible interval of size α, we expect the credible interval to contain the true quantity approximately $\alpha\%$ of the time. *Alpert and Raiffa* found the credible interval contained the true value far less than $\alpha\%$ of the time.

When referring to *Alpert and Raiffa's* work we deliberately use the word 'participants' rather than 'experts', since subjects had no relevant expertise about what they were being questioned upon. *Alpert and Raiffa* show that the layman is poor at expressing his uncertainty about unlikely scenarios using probability.

Other empirical work in the psychological literature draws similar conclusions to *Alpert and Raiffa*, however other work does not focus on tail probabilities.

It seems sensible to assume that an expert will perform better than a layman. However, the small amount of empirical work in the literature (see *Garthwaite, Kadane and O'Hagan*(20) for a review), shows us that despite some clever procedures such as the method of bisection, even experts tend to be *overconfident*. As far as we are aware, no large empirical studies involving expert knowledge have focussed on assessing tail probabilities.

We later consider two approaches that are designed to avoid the assessment of tail probabilities, when eliciting the parameters of a scaled t-distribution.

6.3 A Bayesian Autoregressive Model

Autoregressive models are a special class of the normal linear model, where the response at time t, denoted by y_t, is a linear function of responses at times $t-1, t-2, \ldots, t-p$, denoted by $y_{t-1}, y_{t-1}, \ldots, y_{t-p}$. Elicitation for the parameters of a linear model has been well studied (see for example *Kadane et al.*(35), *Garthwaite and Dickey* (19)(18)). *Kadane at al.*(34) and *Wolfson*(65) have studied the more complex case of a unit root model for an $AR(1)$ process. They considered a model of the form

$$y_t = \rho y_{t-1} + \mathbf{x}^T \boldsymbol{\beta} + \epsilon_t, \tag{6.2}$$

for some general vector of covariates, $\mathbf{x}$. However, they were unable to exploit the features of a purely autoregressive model.

In this chapter, we consider a stationary autoregressive model of the form

$$y_t = \mu + \sum_{i=1}^{p} \beta_i (y_{t-i} - \mu) + \epsilon_t. \tag{6.3}$$

The literature (see for example *Chatfield*(11)) tells us that a classical analysis of a model of the form (6.3) requires between 50 and 100 points in order to estimate the order of the process and the unknown parameters. We consider the case where we have a small amount of data, $\mathbf{y_d} = \{y_1, \ldots, y_m\}$, but far less than 50 points. However, we do have the subjective information of an expert to supplement this small amount of data. In our application we have just 8 data points.

The standard Bayesian approach requires us to formulate our beliefs about the model parameters using expert knowledge, before observing data, and updating our beliefs via Bayes theorem. However, in our application the data had already been observed, so this was clearly not possible. The prior distribution is theoretically independent of the data, so it can be elicited before or after the data are observed. In practice though, it is difficult for an expert to disregard data once he has been exposed to it. A conventional Bayesian approach is therefore likely to use the data twice. Hence, we elicit the posterior (based on subjective beliefs and the data, $\mathbf{y_d} = \{y_1, \ldots, y_m\}$) directly.

6.3.1 Model and Notation

Our Bayesian autoregressive model takes the form

$$\begin{aligned} Y_{t+1} \,|\, Y_t, \ldots, Y_1, \boldsymbol{\beta}, \sigma^2, &\sim N(\mu + (\mathbf{Y}_{t-(p-1):t} - \mathbf{1}\mu)^T \boldsymbol{\beta}, \sigma^2), && (6.4) \\ \boldsymbol{\beta} \,|\, Y_t, \ldots, Y_1, \sigma^2 &\sim N(\mathbf{b}, (\sigma^2/w)\mathbf{U}), && (6.5) \\ \sigma^2 | Y_t, \ldots, Y_1 &\sim wn/\chi^2_n, && (6.6) \end{aligned}$$

where Y_{t+1} is our (unknown) quantity at time $t+1$, and $\mathbf{Y}_{t-(p-1):t}$ represents the $p \times 1$ vector of past observations, $\mathbf{Y}_{t-(p-1):t} = (Y_t, Y_{t-1}, \ldots, Y_{t-(p-1)})$. That is, the predictive distribution of Y_{t+1}, at some future time $t+1$, conditional on past observations is normally distributed as (6.4).

We let $\boldsymbol{\beta}$ denote our $p \times 1$ vector of autoregressive coefficients, σ^2 denote the variance, and $\mathbf{1}$ denote the $p \times 1$ vector of ones. For simplicity, the order of the autoregressive component, p, is assumed to be known. Methods are available in a linear models context (see *Garthwaite and Dickey* (19)) for when we wish to explicitly model the uncertainty in p.

Our biggest assumption is that the process mean, μ is known. However, (6.4) already implies strong beliefs that the series is stationary. If we are prepared to assume stationarity, it is not unreasonable to assume we might have strong beliefs about μ. In the application that motivates this research, and which is discussed at the end of this chapter, μ represents a treasury target inflation rate. The Bank of England has a range of powers in order to control the level of inflation, so our expert assured us that μ could be treated as known and assumed to be constant over time.

The elicitation task is to quantify our expert's opinion about the unknown parameters of the autoregressive model, (6.4), in the form of a joint probability distribution for $\boldsymbol{\beta}$ and σ^2. In order to make the problem tractable a conjugate prior (with as yet unspecified hyperparameters) is chosen. Marginally, σ^2 is distributed as wn times the reciprocal of a chi squared random variable with n degrees of freedom. Conditional on σ^2, $\boldsymbol{\beta}$ has a multivariate normal distribution, with mean $\mathbf{b}$ and variance/covariance matrix $(\sigma^2/w)\mathbf{U}$.

We need to question the expert in order to obtain his beliefs about our model hyperparameters. *Kadane et al.*(35) state there have been previous (unpublished) attempts at obtaining this information directly. However, even if we simplify the model by taking the variance, σ^2, to be known, and the elements of $\boldsymbol{\beta}$ are taken to be independent *a priori*, an expert may well struggle to provide estimates of $\mathbf{b}$

and the (diagonal) elements of $\mathbf{U}$. The more general case as given in (6.4)-(6.6) is much more difficult to tackle directly.

The more common approach in elicitation, as used by *Kadane et al.*(35), *Garthwaite and Dickey* (19),(18) and *Oakley* (48) in a linear models context, involves indirect questioning. We ask the expert about observable quantities, which they might reasonably be expected to offer opinions about. We ask the expert to provide responses such as a median, mode or quantiles for given data. We translate these responses back into statements about the hyperparameters.

For a hierarchial model of the form (6.4)-(6.6), it seems natural to partition the elicitation process. We propose a two phased process to estimating the model hyperparameters. As in *Garthwaite and Dickey* (19),(18), we structure the elicitation process as:

1. *phase 1* – judgements about the variance hyperparameters n and w;

2. *phase 2* – judgements about the expectation hyperparameters $\mathbf{b}$ and $\mathbf{U}$.

The methodology we propose to estimate n and w is straightforward, and similar to *Garthwaite and Dickey* (19),(18). We estimate these parameters, and treat them as known when we elicit the expectation hyperparameters in the second phase of our process. We develop this methodology in section 6.4.

The problem of estimating expectation hyperparameters $\mathbf{b}$ and $\mathbf{U}$ is more difficult. In particular the off diagonal elements of $(1/w)\mathbf{U}$ pose problems. In our model these represent the strength of the correlation between the autoregressive parameters, $\boldsymbol{\beta}$. We consider methodology for estimating expectation hyperparameters $\mathbf{b}$ and $\mathbf{U}$ in section 6.5.

6.4 Variance hyperparameters

In the first phase of our elicitation process, we develop a method for eliciting variance hyperparameters n and w independently of $\mathbf{b}$ and $\mathbf{U}$. Our method is similar in spirit to the work of *Garthwaite and Dickey*(18), in the context of a linear model. They achieved independence by questioning their expert about the difference between two responses observed at the same design point. We develop a method based upon the very strong prior information about μ.

We consider the predictive distribution of Y_{t+1}. We wish to question our expert about $Y_{t+1}|\mathbf{Y}_{t-(p-1):t}$, and from his assessments deduce the distribution, $f(Y_{t+1}|\mathbf{Y}_{t-(p-1):t})$. In general this is a function of all 4 unknown hyperparameters. However, by utilizing the strong prior information about μ, we question the expert about $Y_{t+1}|\,(Y_t, \ldots Y_{t-(p-1)}) = (\mu, \ldots \mu)$ and deduce $f(Y_{t+1}|\,(Y_t, \ldots Y_{t-(p-1)}) = (\mu, \ldots \mu))$, which is a function of just variance hyperparameters n and w. We describe how we achieve this in more detail below

The distribution of Y_{t+1}, conditional on unknown parameters $\boldsymbol{\beta}$ and σ^2, and the series $\mathbf{Y}_{t-(p-1):t} = (Y_t, \ldots Y_{t-(p-1)} = y_t \ldots y_{t-(p-1)})$, is written as

$$Y_{t+1}|(Y_t, \ldots Y_{t-(p-1)}) = (y_t, \ldots y_{t-(p-1)}), \boldsymbol{\beta}, \sigma^2 \sim N(\mu + \sum_{i=1}^{p}(y_{t-(i+1)} - \mu)\beta_i, \sigma^2). \tag{6.7}$$

In particular, if we take the realizations of the previous p observations to be $(Y_t, \ldots Y_{t-(p-1)}) = (\mu, \ldots \mu)$, then (6.7) is independent of $\boldsymbol{\beta}$, and written as

$$Y_{t+1}|\,(Y_t, \ldots Y_{t-(p-1)}) = (\mu, \ldots \mu)\,, \sigma^2 \sim N(\mu\,, \sigma^2). \tag{6.8}$$

At this point (6.8) is conditional on σ^2 as well as the past. We will need to remove the conditioning on σ^2 in order to question our expert about $Y_{t+1}|\,(Y_t, \ldots Y_{t-(p-1)}) =$

$(\mu, \dots \mu)$. We do so by using Bayes theorem, before integrating over σ^2.

We can write the joint distribution of $\mathbf{Y}_{t-(p-1):t}|\boldsymbol{\beta}, \sigma^2$ as

$$f(\mathbf{Y}_{t-(p-1):t}\,|\,\boldsymbol{\beta}, \sigma^2) = \int f(Y_t\,|\,(\mathbf{Y}_{t-p:t-1},\, \boldsymbol{\beta}, \sigma^2) \times \dots$$
$$\times\, f(Y_{t-(p-1)}\,|\,\mathbf{Y}_{t-2p+1:t-p},\, \boldsymbol{\beta}, \sigma^2)\, f(\mathbf{Y}_{t-2p+1:t-p}|\boldsymbol{\beta}, \sigma^2) d\mathbf{Y}_{t-2p+1:t-p}, \quad (6.9)$$

which contains expressions involving $Y_t, \dots, Y_{t-(p-1)}$ and the additional random variables $Y_{t-p}, \dots, Y_{t-2p+1}$, and is dependent on $\boldsymbol{\beta}$.

We have chosen a very specific set of observations, $Y_t, \dots, Y_{t-(p-1)} = \mu, \dots \mu$, and these contain no information about $\boldsymbol{\beta}$. We have no concern about the preceding observations $Y_{t-p}, \dots, Y_{t-2p+1}$ that could in principle generate this set of observations, therefore the latter term and integral in (6.9) are dropped and we model $Y_t, \dots, Y_{t-(p-1)}$ as functions of σ^2 alone. It is important that we model these observations as functions of σ^2, since observing the sequence $Y_t, \dots, Y_{t-(p-1)} = \mu, \dots \mu$ will no doubt influence beliefs about Y_{t+1}, especially for large p. Thus, we have

$$f((Y_t, Y_{t-1} \dots, Y_{t-(p-1)}) = (\mu, \mu, \dots, \mu)|\,\boldsymbol{\beta}, \sigma^2) \propto \prod_{i=1}^{p} \sigma^{-2}. \qquad (6.10)$$

We find that since (6.10) is a function of σ^2 alone that it is independent of $\boldsymbol{\beta}$. We combine (6.10) with our prior on σ^2, (6.6), and update our beliefs about σ^2 in light of data, $\mathbf{Y}_{t-(p-1):t}$, via Bayes theorem. We have

$$f(\sigma^2|(Y_t, \dots Y_{t-(p-1)} = \mu, \dots \mu)) \sim w^*(n+p)/\chi^2_{n+p}, \qquad (6.11)$$

where

$$w^* = nw/(n+p). \qquad (6.12)$$

We take the product of (6.8) and (6.11) in order to get the joint distribution of

Y_{t+1} and σ^2, conditional on $\mathbf{Y}_{t-(p-1):t}$. After integrating over σ^2 we have

$$Y_{t+1} \,|\, Y, \ldots Y_{t-(p-1)} = \mu, \ldots \mu \sim t_{n+p}(\mu, w^*). \tag{6.13}$$

We are able to elicit summaries of $Y_{t+1} \,|\, Y_t, \ldots Y_{t-(p-1)}$, as discussed in section 6.1, in order to estimate the variance hyperparameters. We only need question our expert about the observable quantity Y_{t+1}.

We begin by eliciting our experts 25^{th} and 75^{th} percentiles of $Y_{t+1} \,|\, Y_t, \ldots Y_{t-(p-1)}$, and we calculate the interquartile range, which we denote k_1. Since μ is known (although we should check our experts median value is in fact μ), we can elicit the 25^{th} and 75^{th} percentiles without need to take account of errors in μ. However, in order to identify n and w uniquely we need additional information. We now consider two alternative approaches that provide this additional information without the need to assess tail probabilities.

6.4.1 Conditioning Method

Our first method adapts the methodology of *Garthwaite and Dickey*(18) for an autoregressive model. We suppose that $Y_{t+1} = y_{t+1}$ was observed, and further suppose that the series $(Y_{t'}, \ldots Y_{t'-(p-1)}) = (\mu, \ldots \mu)$ is observed at some future time. We adopt the dash notation to distinguish between the two distinct time periods. We want to know how the observation $Y_{t+1} = y_{t+1}$ affects our expert's beliefs about the random variable $Y_{t'+1}$.

We first update our beliefs about σ^2 in light of the observation $Y_{t+1} = y_{t+1}$;

$$\begin{aligned} \sigma^2 | Y_{t+1} = y_{t+1}, (Y_{t'}, \ldots Y_{t'-p}) &= (\mu, \ldots \mu) \sim w^{**}(n+p+1)/\chi^2_{n+p+1}, (6.14) \\ w^{**} &= ((y_{t+1} - \mu)^2 + nw)/(n+p+1). \quad (6.15) \end{aligned}$$

Now we consider our beliefs about $Y_{t'+1}$ given the new observation $Y_{t+1} = y_{t+1}$. We note that Y_{t+1} and $Y_{t'+1}$ are independent given $(Y_{t'}, \dots Y_{t'-(p-1)}) = (\mu, \dots, \mu)$ and parameters β and σ^2. Therefore, we can write

$$Y'_{t+1} \,|\, Y_{t+1} = y_{t+1}, (Y_{t'} \dots Y_{t'-(p-1)}) = (\mu \dots \mu), \sigma^2 \sim N(\mu, \sigma^2). \qquad (6.16)$$

We now remove the conditioning on σ^2. We take the product of (6.14) and (6.16) in order to obtain the joint distribution of $Y_{t'+1}$ and σ^2 conditional on $Y_{t+1} = y_{t+1}, (Y_{t'} \dots Y_{t'-(p-1)}) = (\mu \dots \mu)$. Now, integrating over σ^2 leaves

$$Y_{t'+1} \,|\, Y_{t+1} = y_{t+1}, (Y_{t'} \dots Y_{t'-(p-1)}) = (\mu \dots \mu) \sim t_{(n+p+1)}(\mu, w^{**}). \qquad (6.17)$$

We elicit summaries of this distribution. In fact, we only need one assessment, in addition to the previously elicited k_1, in order to uniquely identify n and w. We elicit our experts 25^{th} and 75^{th} percentiles, and calculate the interquartile range, which we denote as k_2.

We can write k_1 and k_2 as

$$\begin{aligned} k_1 &= w^{*1/2} q_{n+p}, & (6.18) \\ k_2 &= w^{**1/2} q_{n+p+1}, & (6.19) \end{aligned}$$

where q_{n+p} and q_{n+p+1} denote the respective interquartile ranges of standard t-distributions with $n+p$ and $n+p+1$ respective degrees of freedom. The solution to these simultaneous equations will uniquely identify n and w. In general we will have to implement a bivariate search procedure in order to identify these hyperparameters. However, if we take $y_{t+1} = \mu + 1/2(k_1 - \mu)$, we can solve the expression

$$\frac{k_1}{k_2} = \frac{q_{n+p}}{q_{n+p+1}} [\frac{n+p+1}{1/8q^2_{n+p} + (n+p)}]^{1/2}, \qquad (6.20)$$

for n. Given n, we then solve for w^* from (6.18) and hence w from (6.12).

6.4.2 Graphical Method

Our second method uses just our expert's first interquartile range, k_1. With just k_1 we do not have enough assessments to fit a unique distribution; in fact we have an infinite number of solutions to (6.18), and hence an infinite number of possible densities that match our elicited interquartile range. We confine our attention to integer values of n, thus greatly reducing the number of possible densities, although there are still infinitely many.

Some of the solutions to (6.18) will be more plausible than others. It is possible to distinguish between these different solutions more readily in our scaled t-distribution than is possible within the class of standard t-distributions. In general we will not know which subset of solutions to (6.18) are 'most plausible' in that they closely match the experts beliefs. However, for an arbitrary choice of n we can calculate w^* and plot the corresponding density

$$Y_{t+1} \,|\, (Y_t, \dots Y_{t-(p-1)}) = (\mu, \dots \mu) \sim t_{n+p}(\mu, w^*), \tag{6.21}$$

and ask our expert if our fitted density for Y_{t+1} approximates his beliefs.

If the density is consistent with the expert's beliefs, then we are within a subset of plausible values for n and w^*. If our density is inconsistent with the expert's beliefs, we can choose a different density. We modify n and w based on the expert's comments. Once we identify a plausible subset of densities we can attempt to select a single density from this subset.

Choice of a density is a simpler task if we have a point of reference; it is easier to reject one density in favour of another density, than to reject a density with no point of reference. However, we do need to strike a balance – it is not realistic to plot a large number of densities, and to have our expert choose a 'best density' from these. It is realistic to have our expert choose which density best matches his beliefs amongst a competing pair. When one density has small n and the other density has large n, this should be an easy choice to make.

If we make a series of comparisons, taking note of the expert's preferred density in each case, we can quickly converge toward a range of plausible values of n, and corresponding w^*. Since we only consider integer values of n, it is theoretically possible to converge to a single density.

We propose the following algorithm

1. Set $n_1 = p$ and $n_2 = 25$

2. Repeat

3. Calculate w_1^* and w_2^* from (6.18) and plot the two densities

4. If the expert chooses the density $t_{n_1}(\mu, w_1^*)$, then set $n_2 = n_2 - 1$.

   If the expert chooses the density $t_{n_2}(\mu, w_2^*)$, then set $n_1 = n_1 + 1$.

5. Terminate when $n_1 = n_2$

We demonstrate the algorithm via a simple example. We consider an autoregressive model with $p = 1$, $\mu = 2.5$ and elicited interquartile range of $k_1 = 0.75$. We show two possible comparisons that our expert may be faced with.

In *Figure* (6.2) we show densities with $n_1 = 3$ and $n_2 = 25$. In *Figure* (6.3) we show densities with $n_1 = 20$ and $n_2 = 25$. Our example highlights the fairly large differences between densities in the first few iterations of our algorithm. The two densities in *Figure* (6.2) are clearly distinguishable, so it should be a simple

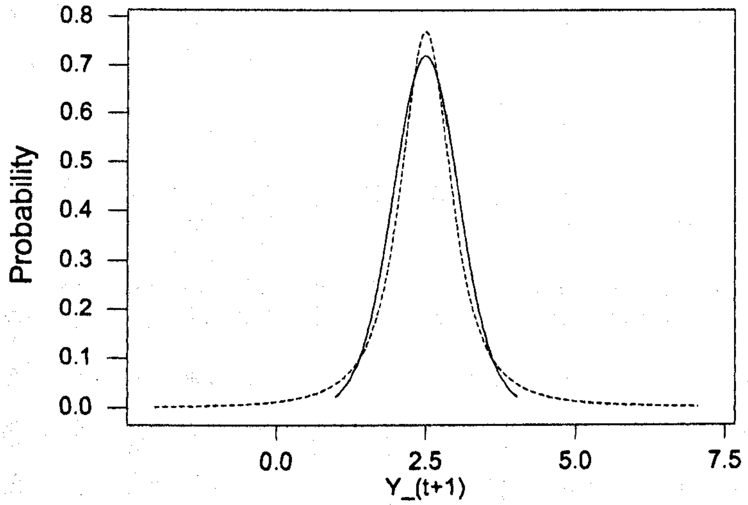

Figure 6.2: t-distributions with $n_1 = 3$ and $n_2 = 25$ degrees of freedom

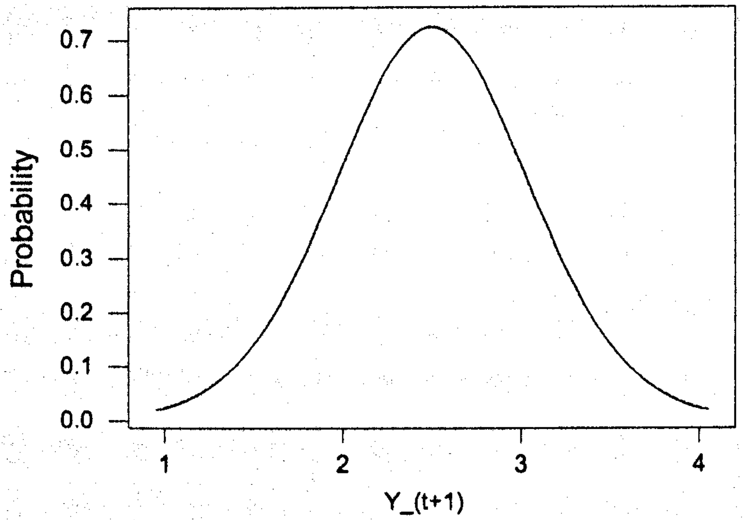

Figure 6.3: t-distributions with $n_1 = 20$ and $n_2 = 25$ degrees of freedom

task for our expert to choose his preferred density in the early iterations of our algorithm.

Our second plot, *Figure* (6.3), highlights that differences between t-distributions can only been seen in the extreme tails of the densities, even with our scaled parametrization, as n_1 increases towards n_2. It may not be realistic to expect an expert to make such fine judgements. In light of this we modify the last step of our algorithm to

5 Terminate when $n_1 = n_2$ or when our expert is indifferent as to n_1 or n_2.

At termination of the algorithm our uncertainty is reduced to a (hopefully relatively small) subspace of n (and resultantly w^*). We have to choose a single value from this region. *Kadane et al.*(35) encountered a similar problem in their work. They asked their expert more questions than our method requires, and as a result they obtained multiple estimates of n. To resolve this, they opted to take the arithmetic mean of their estimates. However, *Al-Awadhi and Garthwaite's*(1) empirical work suggested the geometric mean of these estimates is more stable.

Whilst taking the mean (arithmetic of geometric) of n_1 and n_2 is an option seemingly inline with previous work, we prefer to take n to be the smallest integer value contained in the subset. Our estimate is intended to counteract possible expert overconfidence.

Finally, we could speed our algorithm up by allowing larger jumps in the initial stages. At the first iteration of our algorithm, since the $n_1 = p$ and $n_2 = 25$ cases are so radically different, the expert is likely to have a strong preference for one of these densities based on how 'thick' they want the tails of the density to be. Supposing our expert initially had a strong preference for n_2, for our first modification to n_1 we might consider taking $n_1 = n_1 + c$ for $c > 1$. How large we should make c may depend on the strength of the expert's opinions - the extent to which he favours one density over another but this is difficult to quantify in terms of n. It may be to err on the side of caution and take c to be relatively small.

6.5 Expectation hyperparameters

6.5.1 Generating Series

In order to elicit the expectation hyperparameters we will have to ask our expert additional questions. These questions are of a similar form to those posed previously - given a short series we want to elicit beliefs about the one step ahead forecast, Y_{t+1}. Here we discuss a procedure for selecting our time series of length p, that we might think of as constituting 'design points'. This is a similar, albeit a more structured problem, to the one encountered in the linear models literature.

In the case of a linear model (6.5) and (6.6) remain unchanged, however (6.4) is replaced by

$$Y|\,\mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}^{\mathbf{T}}\boldsymbol{\beta}, \sigma^2), \tag{6.22}$$

where $\mathbf{X}$ is the *design point*, and Y the *response*.

We briefly review the literature relating to design points for the linear model, before looking at the problem of generating series for our autoregressive models.

Design Points

The most efficient procedure for estimating $\mathbf{b}$ and $\mathbf{U}$ in the linear models literature is given by *Garthwaite and Dickey*(18). They required just p design points, and at each design point they elicited the median and 25^{th} and 75^{th} percentiles (and hence were able to estimate the variance). They were able to determine the $p(p+1)/2$ elements of $\mathbf{U}$ with just p elicited variances.

Garthwaite and Dickey had their expert select the design points sequentially, subject to an increasing number of constraints. At stage i of their procedure, *Garthwaite and Dickey* fixed the first i inputs at values $X_1 = a_1, \ldots X_i = a_i$. They had their expert choose the values of the remaining $p - i$ inputs, subject to the condition that the variance of the prior predictive distribution of Y_i was minimized. They referred to this condition, as constrained minimum variance.

The special structure of the design points means that in effect, the design points in themselves contain some information about $\mathbf{U}$, and *Garthwaite and Dickey* were able to exploit this. However, despite the obvious advantages of such an approach, it has been criticized in a review of the literature by *Garthwaite, Kadane and O'Hagan*(20) since we have no information about the expert's inconsistencies. Additionally, there is no simple and practical adaptation of the methodology for an autoregressive model.

Other research in a linear models context (see for example *Kadane et al.*(35) and *Oakley* (48)), has recommended the use of at least $p(p+1)/2$ design points. If we have $> p(p+1)/2$ points, we can give feedback to our expert about his inconsistencies. A thorough procedure for selecting a space filling design requires the joint distribution of the inputs in order to choose the design points. However, $\mathbf{X}$ may not have a distribution in a linear models context, and if a distribution does indeed exist, this in itself may need to be elicited.

Kadane et al. elicited a range for each input dimension X_i, and generated the design point in dimension X_i, using a fairly crude discrete grid. Each dimension of $\mathbf{X}$ was sampled separately. This procedure obviously ignores any correlation that may exist between the inputs. However, the expert was able to reject each of the proposed design points as implausible and reselect, so the desired correlation structure may be generated, albeit inefficiently.

Time Series

An autoregressive model requires a more thorough procedure than a linear model. Each 'design point' is now a time series of length p. For each of our series, we require the expert to make assessments about Y_{t+1} given $Y_t, \ldots, Y_{t-(p-1)}$. We will require at least $p(p+1)/2$ series in order to determine the elements of $\mathbf{U}$.

As we stated above for the case of the linear model, a thorough procedure for selecting the 'design points' requires the joint distribution of our series $Y_t, \ldots, Y_{t-(p-1)}$. We elicited n and w in section 6.3, so we have information about σ^2, but we require the unknown $\boldsymbol{\beta}$ in order to state the joint distribution of $Y_t, \ldots, Y_{t-(p-1)}$.

We propose a procedure that makes use of the small amount of observed data $\mathbf{y_d}$. We stated in section 6.2 that a classical analysis was flawed since $\mathbf{y_d}$ does not contain enough observations. We can estimate $\boldsymbol{\beta}$ and σ^2, but the estimates, especially of σ^2, would be unreliable. However, we can use $\mathbf{y_d}$ in order to obtain crude maximum likelihood estimates of $\boldsymbol{\beta}$, which by convention, we denote $\hat{\boldsymbol{\beta}}$. We need not estimate σ^2, since the hyperparameters n and w are known.

We now employ the following algorithm, repeating for each design point.

1. Randomly select a continuous series of length p from $\mathbf{y_d}$. We denote these points $y_{t-2p+1}, \ldots, y_{t-p}$ respectively.
2. Generate a value for the variance, σ^2, from (6.6).
3. Generate our design point using

$$\begin{aligned}
Y_{t-(p-1)} &= \mu + (y_{t-p} - \mu)\hat{\beta}_1 + \ldots + (y_{t-2p+1} - \mu)\hat{\beta}_p + \epsilon_1, \\
Y_{t-(p-2)} &= \mu + (Y_{t-(p-1)} - \mu)\hat{\beta}_1 + \ldots + (y_{t-2p+2)} - \mu)\hat{\beta}_p + \epsilon_2, \\
&\vdots \\
Y_t &= \mu + (Y_{t-1} - \mu)\hat{\beta}_1 + \ldots + (y_{t-p} - \mu)\hat{\beta}_p + \epsilon_p,
\end{aligned}$$

where $\epsilon_i \sim N(0, \sigma^2)$.

Given the potential for error in our maximum likelihood estimate of $\boldsymbol{\beta}$, we allow our expert to reject any design point as implausible.

6.5.2 Estimating b

From (6.4), the expectation of Y_{t+1} conditional on $\boldsymbol{\beta}, \sigma^2$ and the data, is $\mu + (\mathbf{Y_t} - \mathbf{1}\mu)\boldsymbol{\beta}$. Removing the conditioning on the hyperparameters, $\boldsymbol{\beta}$ and σ^2, leaves us with

$$E[Y_{t+1}|\mathbf{Y}_{t-(p-1):t}] = \mu + (\mathbf{Y}_{t-(p-1):t} - \mathbf{1}\mu)\,\mathbf{b}. \qquad (6.23)$$

We wish to estimate $\mathbf{b}$ based on our expert's judgements.

We have $m \geq p(p+1)/2$ series, selected using the methodology of 6.5.1. At each of these we elicit a point estimate for Y^i_{t+1}, where Y^i_{t+1} denotes the predictive distribution of our i^{th} series at time $t+1$. In line with the discussion of section 6.2 we elicit our experts median, which we denote $\hat{y}^i_{t+1,0.5}$.

We treat our elicited medians as data. We assume our expert's elicitation errors are independent, with zero mean and common (but unknown) variance. By elicitation errors we mean where our expert's judgements do not exactly correspond with our model specification. Like *Kadane et al.*(35) and *Oakley*(48), this justifies a least squares estimate. We find

$$\hat{\mathbf{b}} = \{(\mathbf{Y} - \mathbf{1}\mu\mathbf{1}^T)^T(\mathbf{Y} - \mathbf{1}\mu\mathbf{1}^T)\}^{-1}(\mathbf{Y} - \mathbf{1}\mu\mathbf{1}^T)^T(\hat{\mathbf{y}}_{\mathbf{t+1,0.50}} - \mathbf{1}\mu), \qquad (6.24)$$

where $\hat{\mathbf{y}}_{\mathbf{t+1,0.50}}$ represents our vector of predictive medians, and $\mathbf{Y}$ represents our $m \times p$ matrix of time series.

A measure of how well our model fits the experts elicited medians is the vector of residuals $\mathbf{e} = \mathbf{y_{t+1,0.50}} - (\mathbf{Y} - \mathbf{1}\mu\mathbf{1^T})^\mathbf{T}\hat{\mathbf{b}}$. We examine the absolute value of the residuals to detect any inconsistent judgements. We feedback any inconsistent judgements, where our model and their judgement differ substantially, to our expert, and allow these judgements to be changed in light of this feedback. We re-estimate $\mathbf{b}$ if any medians are re-assessed.

6.5.3 Estimating U

We now move onto the difficult task of eliciting $\mathbf{U}$. From (6.4) we have that the variance of Y_{t+1}, conditional on $\boldsymbol{\beta}, \sigma^2$ and the data, is σ^2. We remove the conditioning on the hyperparameters $\boldsymbol{\beta}$ and σ^2, and find the variance of $Y_{t+1}|\mathbf{Y}_{t-(p-1):t}$ as

$$Var[Y_{t+1}|\mathbf{Y}_{t-(p-1):t}] = E[\sigma^2]\,\{1 + (\mathbf{Y}_{t-(p-1):t} - \mathbf{1}^T\mu)^T(\mathbf{U}/w)(\mathbf{Y}_{t-(p-1):t} - \mathbf{1}^T\mu)\}, \tag{6.25}$$

where $E[\sigma^2] = wn/(n-2)$ is known.

We could obtain a simpler expression than (6.25) by questioning the expert about the mean response and considering the variance, $Var[\bar{Y}_{t+1}|\mathbf{Y}_{t-(p-1):t}]$. By doing so we remove a factor of $E[\sigma^2]$ from (6.25), an approach which is similar in spirit to the work of *Garthwaite and Dickey*(18). However, it is difficult to consider one's beliefs about an expectation, especially so in a time series context, where an expectation is not observable. We only question our expert about observable quantities. Resultantly our questions relate to $Y_{t+1}|\mathbf{Y}_{t-(p-1):t}$.

As a consequence of questioning our expert about $Var[Y_{t+1}|\mathbf{Y}_{t-(p-1):t}]$ rather than $Var[\bar{Y}_{t+1}|\mathbf{Y}_{t-(p-1):t}]$, we have to enforce the logical constraint

$$Var[Y_{t+1}|\mathbf{Y}_{t-(p-1):t}] \geq Var[Y_{t+1}|\mathbf{Y}_{t-(p-1):t} = \mathbf{1}\mu], \tag{6.26}$$

since the right hand side of (6.26) does not depend on $\boldsymbol{\beta}$, and thus we have less uncertainty.

Now suppose we elicit the α percentile of our experts distribution $Y_{t+1}|\mathbf{Y}_{t-(p-1):t}$. We concentrate on the lower tail of Y_{t+1}, since in our application recent observations have been below the mean μ. We rely on symmetry to fit the upper tail, but implied values for the upper tail provide useful feedback.

From properties of t-distributions, we should find that for a 'perfect' expert – that is an expert with opinions entirely consistent with our chosen model,

$$Y_{t+1,0.5}|\mathbf{Y}_{t-(p-1):t} - t_{n,\alpha}\sqrt{Var[Y_{t+1}|\mathbf{Y}_{t-(p-1):t}]} - Y_{t+1,\alpha}|\mathbf{Y}_{t-(p-1):t} = 0. \qquad (6.27)$$

However, $Var[Y_{t+1}|\mathbf{Y}_{t-(p-1):t}]$ is unknown, since from (6.25) this is a function of the unknown $\mathbf{U}$. We aim to estimate $\mathbf{U}$ using quantile estimates from our expert and the relationship between (6.27) and (6.25).

If we have $m \geq p(p+1)/2$ series, with various values of $\mathbf{Y}_{t-(p-1):t}$ and the same 'perfect' expert, our m equations of the form (6.27), will allow us to solve uniquely for $\mathbf{U}$.

Of course, in reality no expert's judgements will be entirely consistent with a statistical model. If we take the minimum number of assessments, $m = p(p+1)/2$, we can still solve for a unique solution, with zero error. However, we require that $\mathbf{U}$ is positive definite, and if our expert has any inconsistencies, however small, we will not find a positive definite solution for $\mathbf{U}$ that satisfies our $p(p+1)/2$ equations.

Given that we cannot solve exactly for $\mathbf{U}$ subject to constraints (6.27), there is little computational benefit in taking $m = p(p+1)/2$ points. It is preferable to take $m > p(p+1)/2$, since we are able to assess our expert's inconsistencies, and allow him to modify his beliefs in light of feedback.

We now look to a method that selects an optimal, in some sense, value for $\mathbf{U}$, subject to $\mathbf{U}$ being positive definite. Like *Oakley*(48), we adopt a method of least squares in order to optimize for $\mathbf{U}$, although our method differs slightly since $E[\sigma^2]$ is known from our earlier questioning.

We minimize the expression

$$\sum_{i=1}^{m}(\hat{y}^i_{t+1,0.5} - t_{n,\alpha}\sqrt{Var[Y^i_{t+1}|\mathbf{Y}^i_{t-(p-1):t}]} - \hat{y}^i_{t+1,\alpha})^2, \tag{6.28}$$

where $\hat{y}^i_{t+1,\alpha}$ denotes the elicited α quantile. The variance term is derived from $\mathbf{U}$ using (6.25), where we numerically optimize for $\mathbf{U}$.

Oakley(47) used $\alpha = 0.75$, whilst in *Oakley*(48), a more complex sum of squares than (6.28) was used, with $\alpha_1 = 0.75$ and $\alpha_2 = 0.95$. We consider the following different sums of squares in order to estimate $\mathbf{U}$:

$$\sum_{i=1}^{m}(\hat{y}^i_{t+1,0.5} - t_{n,0.25}\sqrt{Var[Y^i_{t+1}|\mathbf{Y}^i_{t-(p-1):t}]} - \hat{y}^i_{t+1,0.25})^2, \tag{6.29}$$

$$\sum_{i=1}^{m}(\hat{y}^i_{t+1,0.5} - t_{n,0.05}\sqrt{Var[Y^i_{t+1}|\mathbf{Y}^i_{t-(p-1):t}]} - \hat{y}^i_{t+1,0.05})^2, \tag{6.30}$$

$$\begin{aligned}\sum_{i=1}^{m} 1/2\{(\hat{y}^i_{t+1,0.5} - t_{n,0.25}\sqrt{Var[Y^i_{t+1}|\mathbf{Y}^i_{t-(p-1):t}]} - \hat{y}^i_{t+1,0.25})^2 \\ +(\hat{y}^i_{t+1,0.5} - t_{n,0.05}\sqrt{Var[Y^i_{t+1}|\mathbf{Y}^i_{t-(p-1):t}]} - \hat{y}^i_{t+1,0.05})^2\}.\end{aligned} \tag{6.31}$$

We elicit the 25th percentile for each series using the method of bisection, discussed in section 6.2. For the 5th percentile, we ask our expert for a value that Y_{t+1} is highly unlikely to drop below and inform our expert we will interpret this value as the 5th percentile (see for example *Mosteller and Yountz* (44)). Obviously a training exercise with feedback to our expert will make this interpretation more plausible.

We denote the elicited variances for our i^{th} predictive distribution, Y^i_{t+1}, calculated using the 25^{th} and 5^{th} percentiles respectively, as $v^i_{0.25}$ and $v^i_{0.05}$ respectively. Given the estimates of $\mathbf{U}$ from (6.29)-(6.31), which we denote $\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2$ and $\hat{\mathbf{U}}_3$ respectively, we can obtain fitted values for the variances from (6.25). We denote the fitted variances for predictive distribution, Y^i_{t+1}, using our j^{th} estimate (for j = 1,2,3) of $\mathbf{U}$, as $\hat{v}^i_{0.25,j}$ and $\hat{v}^i_{0.05,j}$ respectively.

For a expert whose judgements are entirely consistent with our model, we should find

1. $e^i_1 = v^i_{0.75} - \hat{v}^i_{0.75,j} \approx 0$ and $e^i_2 = v^i_{0.95} - \hat{v}^i_{0.95,j} \approx 0 \; \forall i, j$
2. $v^i_{0.75} \approx v^i_{0.95}$
3. the differences, $v^i_{0.75} - v^i_{0.95}$, should be independent

If our three conditions above hold, we take $\hat{\mathbf{U}}_3$ as our estimate of $\mathbf{U}$, since this is the most robust estimate, based on the most judgements. However, we may find in practice that our expert is far more comfortable, and as a consequence more consistent, at estimating either the 25^{th} or 5^{th} percentile. Consequently, if the second and third conditions do not hold we examine the sums of squares (6.29) and (6.30), and select the α with lowest sum of squares in order to estimate $\mathbf{U}$.

We examine the differences between assessed and fitted variances, and allow re-assessment if any of these are large, before re-estimating $\mathbf{U}$.

Related Work

An alternative procedure for estimating $\mathbf{U}$ is given in *Kadane et al.*(35). They develop a mathematically appealing approach, and can estimate $\mathbf{U}$ without the need for a numerical search.

In addition to variances, their approach would require us to elicit correlations between $Y^i_{t+1}|\mathbf{Y}^i_{t-(p-1):t}$ and $Y^j_{t+1}|\mathbf{Y}^j_{t-(p-1):t}$ for all $i \neq j$. However, correlations are difficult to assess. *Clemen, Fischer and Winkler*(12) found that from a variety of methods, subjects performed best when stating a correlation directly. *Kadane et al.'s*(35) methodology requires the assessment of the more difficult conditional correlations, and as far as we are aware little empirical work has been done on the elicitation of these. Because of difficulties in the assessment of conditional correlations, *Kadane et al.'s*(35) methodology, whilst mathematically appealing, does not guarantee sensible values for their model hyperparameters in practice.

6.6 Example: Modelling Inflation

In chapter 1, when we gave an introduction to Government financial models, we explained that large projects often run over the course of decades. As a result, financial models contain estimates of various financial indices far into the future. In the financial model that we discuss in detail in the next chapter we require estimates of 2 different measures of inflation for thirty years into the future. We now describe the process of eliciting an economic expert's beliefs about one of these measures, the GDP deflator.

Inflation is a measure of how much prices are changing (almost always increasing) from one year to the next. Obviously prices changes are not universal – different commodities (i.e a loaf of bread; a family saloon car; a wide screen television) will have different rates of inflation in a given year. In order to have a single figure for inflation, we have to average over these commodity wise inflations. It is impractical to average over all commodities, so a "shopping basket" of goods is assessed and the average inflation from this shopping basket is taken to be the level of inflation.

The British Government uses three main measures of inflation; the Retail Price Index (RPI), the GDP Deflator, and the Consumer Prices Index (CPI). These measures differ in the content of the "shopping basket", and the way in which we average. The RPI and GDP Deflator are similar measures, they only differ in that the GDP Deflator does not include imported goods in the "shopping basket". In both cases a weighted arithmetic mean of commodity wise inflations is taken. The CPI excludes various housing costs from the "shopping basket" and a weighted geometric mean of commodity wise inflations is taken. Due to the differences in measurement, these three measures will differ at any given time, but the RPI and GDP Deflator are broadly similar, whilst CPI is approximately 0.75% below the other two measures. The Government target value for inflation in any given year is 2% as measured by CPI, which corresponds to a target value of 2.75% as measured by the GDP Deflator.

The Bank of England has been responsible for controlling inflation since 1997. The Bank of England have various powers in order to control inflation, but the major power is in setting baseline interest rates. Inflation and interest rates are negatively correlated, and the Bank of England use interest rates to control inflation.

- When inflation is high, a high interest rate will curb consumer spending and reduce inflation.
- A reduction in interest rates encourages consumer spending and inflation will increase.

Inevitably, there is a time lag between the Bank of England's actions, and the behavior of interest rates. Our economic expert assured us that if inflation strays significantly from the target value, the Bank of England will aim to be back on target in 2 years time, using small adjustments to base line interest rates in order to ensure continued stability in the economy.

The Bank of England produces short term forecasts for inflation, providing continuous time forecasts for 3 years in advance. In addition to the forecast they provide symmetric bounds of increasing uncertainty about the mean/median/mode in the form of a fan chart. Uncertainty, inevitably increases as a function of time.

The Bank of England uses a detailed economic model in order to produce its forecasts. Inflation is correlated with itself through time, but it is also a function of various economic factors (output gap, commodity prices (oil/gold etc ..), tightness of labour market etc ..). By using relevant economic data in its model, and adjusting its policy as a consequence, the Bank of England has greater certainty about inflation in the future than when using past values of inflation alone.

Whilst the Bank of England's analysis is useful in the short term, for our application we need to model inflation in the long term, when we will not have any relevant economic data. Discussions with our economic expert revealed that a stationary model, with known mean of 2.75% was appropriate. An autoregressive model captures the main features of the data. Given the parameters of the autoregressive model, we can find the joint distribution of inflation rates for the full period of our model.

A classical analysis would involve using past data to estimate the parameters of the model, using some criterion to determine the order of the autoregressive process. However, our economic expert advised us to discount all information before 1997, when the Bank of England was given responsibility for controlling inflation. In light of the lack of data, a Bayesian model which synthesizes the small amount of economic data with the subjective knowledge of an expert, is the only sensible procedure.

Our first task was to determine the order of the process. Given the information that when inflation strays significantly from the target value, the Bank of England will aim to be back on target in 2 years time, it seemed natural to take p, the order of the process, to be 2. Our expert agreed with this assessment.

In order to estimate the parameters of the process, we gave our expert different hypothetical future series consisting of 2 time points, indexed as $t-1$ and t respectively. Inflation is usually calculated at 3 monthly intervals, however we model the average inflation over the year, and hence our time points are years. We elicited our expert's beliefs about year $t+1$. Since the series were "snapshots" of the future, our expert had no other relevant economic data, just inflations for years $t-1$ and t. The process our expert used in each case was to use the data we provided to infer the economic situation of the time. Our expert then inferred the Bank of England's response to the economic situation, and provided his beliefs about year $t+1$.

Our first question, using the process described in section 6.4, aimed to elicit beliefs about the variance hyperparameters n and w. We took inflation to be at the target rate of 2.75% for two consecutive years, and elicited our experts median and upper and lower quartiles.

For the conditioning method we required more information. We instructed our expert that the series $(2.75\%, 2.75\%, 2.85\%)$ was previously observed. We wanted our expert to update his beliefs about $Y_{t'+1,0.25}$, $Y_{t'+1,0.5}$ and $Y_{t'+1,0.75}$ in light of this information. Our expert provided his median value, but he was unsure about the two quartiles and how they should differ from his previous assessments. We did not force an answer from him, and as a result did not obtain his quartiles.

Responses are tabulated in *Table* (6.1)

Y_{t-1}	Y_t	Y'_{t+1}	$Y_{t+1,0.5}$	$Y_{t+1,0.25}$	$Y_{t+1,0.75}$
2.75	2.75		2.75	2.55	2.95
2.75	2.75	2.85	2.75	-	-

Table 6.1: Forecasts based on hypothetical data

We had insufficient data to use the conditioning method. The expert was far more comfortable with the graphical approach. He had strong beliefs about the likely bounds of inflation in the following year. The algorithm terminated when he was unable to choose between t distributions with 22 and 25 degrees of freedom. Our parameter estimates are tabulated in *Table* (6.2).

	n	w
Conditioning Method	-	-
Graphical Method	20	0.0935

Table 6.2: Estimates of n and w

A further ten 'design points' were chosen using the methodology described in 6.5.1. For each series we asked our expert for his median. We show assessed medians and residuals, calculated using the methodology described in 6.5.2, in *Table* (6.3).

We note from the assessed medians that our expert gave assessments on a fairly course scale. When given feedback in the form of residuals the expert indicated that all his initial answers were given to the nearest 1/4 point and he modified his assessments to series 1, 7 and 10 after examining the residuals. These reassessments are tabulated in *Table* (6.4).

We calculated **b** as

Y_{t-1}	Y_t	$Y_{t+1,0.5}$	Residual
2.02	1.65	2.5	0.17498
2.22	1.67	2.25	- 0.0823
2.89	2.73	2.75	0.008
2.24	2.27	2.5	- 0.064
2.79	2.91	2.75	- 0.0619
2.47	2.03	2.5	0.0286
2.47	1.96	2.25	- 0.1942
2.00	2.32	2.5	- 0.0849
2.19	2.55	2.75	0.0763
2.84	3.18	2.75	- 0.1665

Table 6.3: Forecasts based on hypothetical data

Y_{t-1}	Y_t	$Y_{t+1,0.5}$
2.02	1.65	2.4
2.47	1.96	2.35
2.84	3.18	2.85

Table 6.4: Expert Re-assessments

$$\boldsymbol{b} = \begin{pmatrix} 0.396 \\ 0.013 \end{pmatrix},$$

which was in line with our experts verbally stated beliefs about how the Bank of England control inflation. With these coefficients, inflations significantly away from the target value of 2.75% would be quickly dragged back towards the target value.

Finally, we elicited the 25% and 5% percentiles for each of the series tabulated in *Table* (6.3). We show these along with medians in *Table* (6.5).

The two blanks in *Table* (6.5) represent values our expert had significant uncertainty about – he provided a range rather than a single figure answer. Rather than force an answer from our expert, we used the search procedure described in section 6.5.3, with these values missing. After calculating **U**, fitted values for

Y_{t-1}	Y_t	$Y_{t+1,0.5}$	$Y_{t+1,0.25}$	$Y_{t+1,0.05}$
2.02	1.65	2.4	2	1.65
2.22	1.67	2.25	2	1.7
2.89	2.73	2.75	2.55	2.25
2.24	2.27	2.5	2.27	-
2.79	2.91	2.75	2.5	2
2.47	2.03	2.5	2.25	2
2.47	1.96	2.35	-	1.95
2.00	2.32	2.5	2.3	2
2.19	2.55	2.75	2.5	2.1
2.84	3.18	2.85	2.6	2.25

Table 6.5: Forecasts and bounds

these blank cells were used to provide additional feedback.

We discussed three different sums of squares to minimize in order to estimate **U** in 6.5.3. The first used just the 25% percentile, the second used the 5th and the final sum of squares used both 25th and 5th percentiles. In our application, fitted variances showed our expert was consistent with his assessments of the 25th percentile, but not with the 5th. As a consequence we used just the 25th percentile in estimating **U**. We required no re-assessments.

We calculated **U** as

$$\boldsymbol{U} = \begin{pmatrix} 0.03 & 0.024 \\ 0.024 & 0.09 \end{pmatrix}.$$

These coefficients indicate some uncertainty about our second autoregressive coefficient, but little uncertainty about the first.

Our final course of feedback was to simulate hypothetical futures given our expert's judgements, and we show 5 such realizations in *Figure* (6.4). Taking GDP deflator to be 3.26% and 2.1% in 1998 and 1999 (the time frame used in the example was an artefact of the application described in chapter 7), we generated inflations for the period 2000 to 2006 and showed these to our expert. Some of these realizations were considered to be more plausible than others, however our expert stated these were all possible paths that inflation might take, and consistent with with his beliefs.

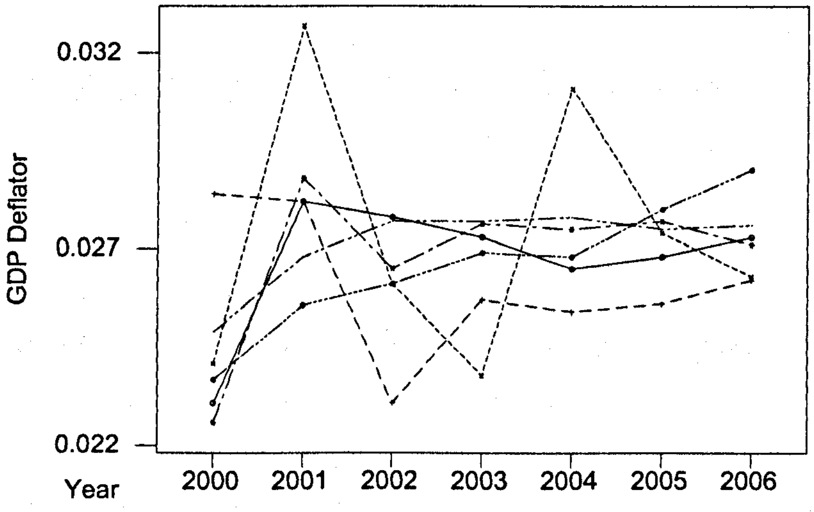

Figure 6.4: Simulated Inflations 2002-2006

6.7 Conclusions

In this chapter we considered elicitation of the parameters of a known order autoregressive model, with known mean, μ. We developed a two stage process for eliciting the hyperparameters of our distribution, first estimating variance hyperparameters, before assessing expectation hyperparameters.

We developed two methods for estimating the variance hyperparameters, so that we could avoid the assessment of tail probabilities. The conditioning method, whilst mathematically pleasing was of little use in our application. The graphical method we developed gave interactive feedback at each assessment. This method, whilst mathematically simple, allowed easy comparisons to be made and led to precise results in our application.

We developed methodology for estimating the expectation hyperparameters that required just assessed medians and at least one percentile. We estimated $\mathbf{b}$ using the expert's medians. The methodology we developed for estimating $\mathbf{U}$ allowed us to select which judgements to use – lower quartile, 5^{th} percentile, or both. In our application, we found our expert to be inconsistent with tail probability assessments, but precise with quartile assessments.

A final conclusion, based upon our limited practical experience, is to concur with various authors that experts are able to accurately assess measures of location such as a median or a mode for a symmetric distribution. Experts have far more difficulty when asked to provide percentiles, especially tail probabilities. We found that upper and lower quartile assessments were fairly consistent, but in line with other authors we found that tail probability assessments were more erratic. Whenever possible, current evidence suggests that the elicitation of tail probabilities should be avoided.

Chapter 7

PFI Example

7.1 Introduction

In this chapter we analyse a financial model with a high dimensional input vector, that arose from the Private Finance Initiative (PFI). The model has a scalar output, the Net Present Value (NPV) of the project. We use the methodology that we developed in chapter 5 to perform uncertainty and sensitivity analyses. We use the methodology of chapter 6 in order to model inflation in the long term – these inflations are inputs to the financial model.

7.2 MOD Main Building Redevelopment

We discussed the nature of PFI projects in detail in the introduction. We stated that PFI projects differ from conventional funding in that they involve a service being procured rather than an asset purchased. The MOD main building redevelopment project represents a typical example of this. The invitation to tender for the project was for a redeveloped and maintained building with a thirty year contract. Initially the project required a process of decant (staff relocating to other premises) before extensive destructive and reconstructive work, and the process of recant (repopulating the building). This initial phase was scheduled to last for 6 years, with decant beginning in 1999, and building works completed in 2004. Following the initial redevelopment phase of the project, the remaining years of the contract, until 2031, were for a maintained building. Maintenance incorporated everything from day to day tasks such as cleaning the building, to major repairs as building hardware needed replacing in the future. Under the terms of the contract the building was to be paid for uniformly, from the end of building works in 2004 until the PFI contract terminated in 2031.

Before the contract could be signed, the MOD were obliged to show that the terms of the PFI contract offered value for money. The Public Sector Comparator (PSC) had to involve the same project brief in order to allow for a like for like comparison with the proposed PFI deal. The PSC, therefore, had to estimate costs for decant, redevelopment, recant, maintenance and scheduled replacements for the period of the project.

7.3 Previous Analysis

The auditing firm Coopers and Lybrand were employed to create the PSC for this project. In consultation with experts they estimated hundreds of costs covering all aspects of the project. They also identified 33 risks (which are parameters of the financial model), each of which had uncertainty, represented by a probability distribution attached to it. Both discrete and continuous distributions probability distributions were used. The probability distributions of the risks were fitted using crude elicitation methodology and very strong assumptions.

As we explained in chapter 1, risks fall into two broad categories. These are adjustments for over-optimism and financial indices. For the MOD Main Building Redevelopment PSC, the first of these categories can be further sub categorized into *decant, redevelopment, operations, renovation, property, insurance, legislative changes and defects.* Since these all represent adjustments for overoptimism in the estimates of costs, in each case we are modelling beliefs about the magnitude of a unknown multiplier. In some instances (for example refurbishment overruns), the risks were estimated as a multiple of the total cost of a particular aspect of the project. In the absence of useful objective data from the current project, these multipliers were modelled using information on overruns from previous projects of a similar scope and scale. In other instances (e.g sale of a building, salvage costs), genuine objective data were available, and as as a result the risks were modelled directly as costs. Financial indices representing rates of inflation from one year to the next, were naturally modelled on a percentage scale.

We show the risks and the assumed probability distributions in *Table* (7.1).

In the MOD main building redevelopment PSC we have two specific categories of risk within *financial indices* (as seen in *Table* (7.1)). These both relate to inflation. When discussing inflation in the chapter 6 example, we stated that prices do not increase uniformly across all products and services from one year to the next. Therefore in calculating a measure of global inflation like the GDP deflator we average over these commodity wise inflations. Our second measure, TPI, is a measure of price increases from one year to the next for the specific commodities related to the project (for example building materials). If we take prices in 1997 to be our base year with index of 100, then over the long term our economic expert advised us that TPI and GDP deflator indices will increase at a similar rate, so for example in 2007 we would expect these indices to be similar. However, for a given year we would not in general expect GDP deflator and TPI

Area	Risk	Distribution
Decant	Cost overrun from budget	Triangular$(3, 12.1, 29.9)\%$
Refurbishment	Cost overrun from budget	Triangular$(0, 8, 31)\%$
Operations	Decant premises services	Normal $(2.5, 4.56)\%$
	Decant support services	Normal $(2.5, 4.56\%)$
	Long term premises services	Normal $(7.5, 4.56)\%$
	Long term support services	Normal $(5, 3.04\%)$
	Lease risk on building	Discrete $(p_x(X = 0) = 0.95,$ $p_x(X = 0.332065) = 0.05)\pounds M$
	Services risk	Triangular $(0, 1, 2)\%$
Renovation	Replacement risk	Triangular $(3, 12.1, 29.9)\%$
Property	Building dilapidation risk	Discrete $(p_x(X = 0) = 0.439,$ $p_x(X = 0.10225) = 0.511,$ $p_x(X = 1.0225) = 0.05)\pounds M)$
	Building dilapidation risk	Log-Normal $(1.0225, 3)\pounds M$
	Building sale risk	Normal $(0, 1.823)\pounds M$
	Salvage risk	Triangular $(-0.016, 0.217, 0.451)\pounds M$
Defect	Latent defect risk	Discrete $(p_x(X = 0) = 0.15,$ $p_x(X = 5) = 0.8,$ $p_x(X = 20) = 0.05)\pounds M)$
Insurance	Fittings insurance risk	Log-Normal $(0.038892, 0.116673)\pounds M$
	Other insurance risk	Log-Normal $(0.2325, 0.6975)\pounds M$
Legislation	Legislative decant risks	Triangular $(-7, 1, 13.5)\%$
	Legislative long-term risks	Triangular $(-7, 1, 13.5)\%$
Finance	GDP deflator risk 97/98	Triangular $(2.8, 2.81, 2.82)\%$
	GDP deflator risk 98/99	Triangular $(3.25, 3.26, 3.27)\%$
	GDP deflator risk 99/00	Triangular $(2.27, 2.4, 2.53)\%$
	GDP deflator risk 00/01	Triangular $(2.43, 2.7, 3.2)\%$
	GDP deflator risk 01/02	Triangular $(1.85, 2.5, 3.15)\%$
	GDP deflator risk 02/03	Triangular $(1.82, 2.5, 3.75)\%$
	GDP deflator risk 03/04	Triangular $(1.79, 2.5, 4.35)\%$
	TPI risk 97/98	Triangular $(8.99, 9, 9.01)\%$
	TPI risk 98/99	Triangular $(5.99, 6, 6.01)\%$
	TPI risk 99/00	Triangular $(1, 5, 8.7)\%$
	TPI risk 00/01	Triangular $(0, 5, 8.8)\%$
	TPI risk 01/02	Triangular $(-0.8, 4.5, 9.2)\%$
	TPI risk 02/03	Triangular $(-0.5, 2.5, 9.5)\%$
	TPI risk 03/04	Triangular $(0, 2.5, 9.5)\%$
	TPI risk 04/05	Triangular $(1.29, 2.5, 4.85)\%$

Table 7.1: Risks in original PSC

inflation to be the same figure, even though both series have the same long-term expectation. Since GDP deflator is an average over many inflations, it is more stable than TPI, with TPI exhibiting more variability. Due to the variability in inflations it is important that they are explicitly modelled.

The GDP deflator, which we described in the example of chapter 6 contributes 7 risks corresponding to the financial years 1997/98 - 2003/04, whilst TPI contributes 8 risks corresponding to the financial years 1997/98-2004/05. The financial indices shown in *Table* (7.1) were assumed to be 100% correlated on a rank correlation scale in the original PSC. Moreover, the PSC required inflations for the whole period of the contract. Due to the difficulty in modelling long term inflation, the auditing firm made the simplifying assumption that inflation as measured by the GDP deflator, for the period 2004/05 - 2031/32, was equal to the 2003/04 figure, and TPI inflation for 2005/06 - 2031/32 was equal to the 2004/05 figure. Under these strong assumptions, the 1997/98 GDP deflator figure uniquely determines all other inflations within the model, which, in effect, reduces the dimension of the input vector from 33 to 19 inputs. Note that the model requires financial indices from 1997/98 - 2031/32, even though the project was not due to commence until 1999. This is since the project was commissioned in 1997, and all costs are in 1997 prices.

7.3.1 Uncertainty

We observed the function at 200 design points, selected according to a Latin Hypercube design. We fitted a Gaussian Process model, as described in chapter 3, and found the posterior distribution of the model output. We calculated measures of uncertainty – the expectation, variance and distribution function using the methodology described in chapter 3.

The posterior expectation (with respect to $\eta(.)$) of the expected NPV was $\pounds 679.6M$ and the variance (with respect to $\eta(.)$) of the expected NPV was $\pounds 0.103M$. Even though the cost is measured in millions of pounds, the posterior distribution of the expected NPV is concentrated on a fairly small range of values.

The posterior expectation of the variance of the NPV was calculated as $\pounds 422.25M$, a large figure indicating substantial variability in the model.

For our final summary we calculated the posterior expectation (with respect to $\eta(.)$) of the distribution function, $F_{Y|\eta(.)}(s)$. We plot this in *Figure* (7.1). We note the heavy upper tail, which extends toward $\pounds 900M$, a figure around 1/3 greater than the expectation, whilst the lower end of the distribution is approx $\pounds 600M$, around 10% below the expectation.

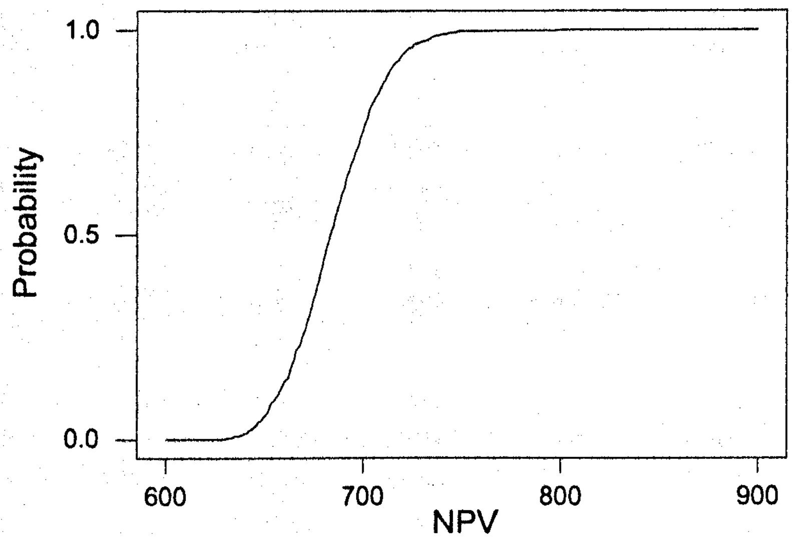

Figure 7.1: Posterior expectation of the distribution function

The MOD model was computationally cheap enough to allow uncertainty analysis to be performed via Monte Carlo so we could verify our estimates. This analysis was based upon 10000 runs of the model. The numerically evaluated expectation ($£679.55M$) and distribution function were close to our Bayesian estimates, and within a 50% credible interval in both cases. The numerically evaluated variance of $£435M$ also appeared to be consistent with our Bayesian estimate.

A analysis based on a Monte Carlo sample of 200 runs was also performed for comparison. Even at this small sample size, the expected NPV was reasonably accurate, albeit with a large standard error. However, the distribution function was inaccurate, and the variance varied greatly from one sample to the next. The Bayesian model performed far better at this sample size.

7.3.2 Sensitivity

We only briefly consider sensitivity analysis for the original PSC, in order to highlight the deficiencies in the modelling of inflation. It became apparent at an early stage that the uncertainty in the financial indices was the major driver of the uncertainty in $\eta(.)$.

In *Figure* (7.2) we show the main effect of GDP deflator risk 97/98. Due to the deterministic correlation structure, this is a reflection of the inflation effect as a whole rather than simply the main effect of GDP deflator risk 97/98. We note the quadratic trend as X increases from its minimum of $£32M$ to its maximum of $£49M$. The inflation main effect was responsible for 50.3% of the total variance.

7.4 Stochastic Inflation Model

Modelling inflation as we described in the previous section induces far greater correlations between inflations than one would expect. Clearly there will be some serial correlation, but not to the extent of a deterministic model.

An exploratory analysis showed the previously assumed correlation structure between GDP deflator and TPI inflations was erroneous. The two series are clearly correlated on an index scale, however on a percentage scale they can be taken to be independent. We highlight this in *Figure* (7.3), where we show percentage increase, and *Figure* (7.4), where we show the index scale. We use GDP deflator and TPI construction figures for the period 1994-2001, taken from the Office for National Statistics (ONS) website. The indices have base year of 100 in 1996.

The index scale *Figure* (7.4) shows that the two series are 100% correlated on a ranking scale, whilst *Figure* (7.3) suggests that on the percentage scale we have independence. Therefore since in the PSC, inflations were specified on the percentage scale, we regard these two series as independent.

The serial correlation within the two series can not be ignored. Using the methodology described in chapter 6 we were able to model these using $AR(2)$ models. We elicited an economic expert's beliefs and obtained the distributions

$$\begin{aligned} \sigma_1^{-2} &\sim \frac{20 \times 0.0935}{\chi^2_{22}} \\ \sigma_2^{-2} &\sim \frac{20 \times 0.17}{\chi^2_{22}} \end{aligned}$$

and

$$\beta_1|\sigma_1 \sim N\left(\begin{pmatrix} 0.396 \\ 0.013 \end{pmatrix}, \frac{\sigma_1^2}{0.0935}\begin{pmatrix} 0.03 & 0.024 \\ 0.024 & 0.09 \end{pmatrix}\right)$$

$$\beta_2|\sigma_2 \sim N\left(\begin{pmatrix} 0.396 \\ 0.013 \end{pmatrix}, \frac{\sigma_2^2}{0.17}\begin{pmatrix} 0.01 & 0.008 \\ 0.008 & 0.03 \end{pmatrix}\right)$$

By treating the first two years inflations as fixed and known for both GDP

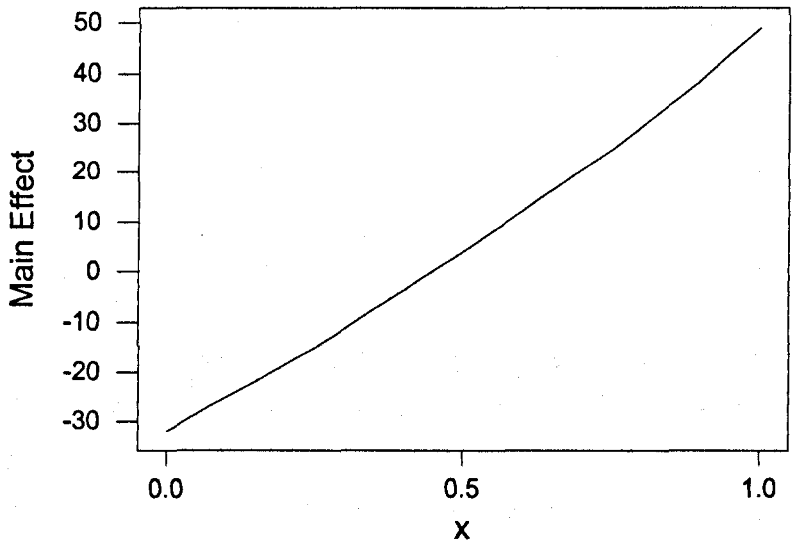

Figure 7.2: Inflation main effect

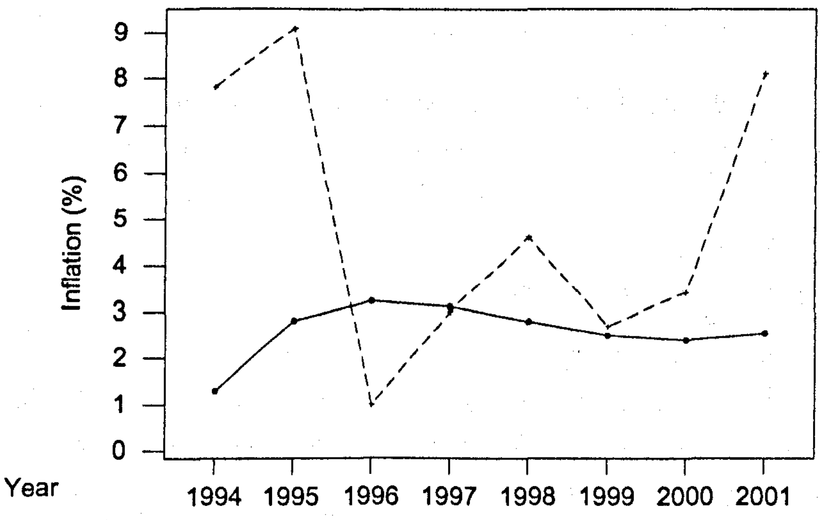

Figure 7.3: Inflations (percentage scale): GDP Deflator (solid); TPI (dashed)

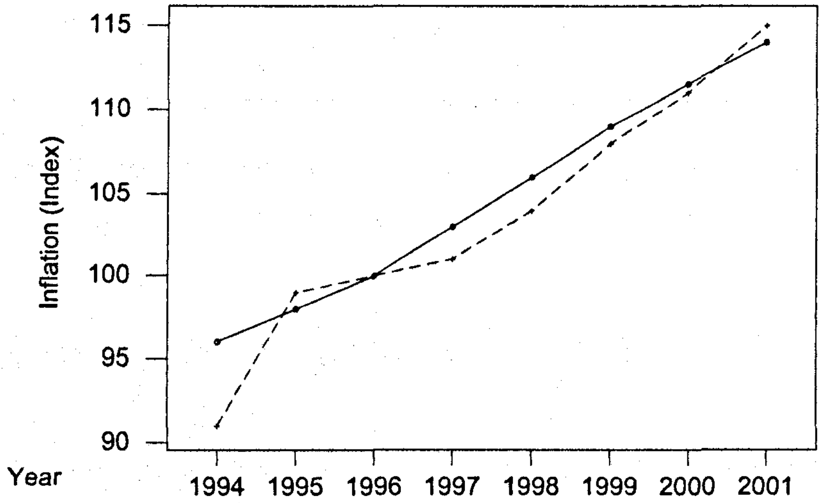

Figure 7.4: Inflations (index scale): GDP Deflator (solid); TPI (dashed)

deflator and TPI inflation, we are able to model the entire series of inflations until 2031/32 using our autoregressive models. Fixing the first 2 years has only a small effect, since the uncertainty in these is very small. Our more complex modelling of inflations has the effect of increasing the number uncertain inputs in the financial model from 33 (although the deterministic nature of inflations in the original model meant this was in effect just 19 inputs) to 88 inputs.

7.4.1 Uncertainty

In modifying the PSC so that we could model inflations using our autoregressive models, we were also able to modify the model so that we could observe additional outputs. Our working knowledge of the model allowed us to identify 7 additive groups and we were able to verify this was the case using the methodology developed in chapter 4. These groups are *capital expenditure, replacements, operational costs, legislation, defects, insurable risks, and inflation*, the sum of which came to the NPV of the project. We observed the function at 500 design points selected using a Latin Hypercube Design. The increased sample size used here is due to the increased complexity of the function. We fitted Gaussian Process models as described in section 4.3.1, and found the posterior distributions of $\eta(\mathbf{x}_{(j)})$ (for j = 1, ... 7, representing the 7 additive groups we identified).

We calculated measures of uncertainty using the methodology we developed in section 5.2. The posterior expectation (with respect to $\eta(.)$) of the expected NPV was calculated as $£674.15M$ and the variance of the expected NPV was $£0.45M$. The former figure is approx $£5M$ below the previously calculated expected NPV, whilst the latter figure reflects that we have more uncertainty about the expectation of the NPV than we had in the previous analysis, not unexpected given the large increase in the number of inputs. However, this variance is still small compared with the very large sums of money involved.

The posterior expectation of the variance of the NPV was calculated as $£345M$, a substantial drop (approx $£80M$) from the figure we calculated in the previous analysis (section 7.3.1). Our more appropriate handling of inflations is the sole reason for this. Due to the known additive decomposition of the model, we calculated the variance as a sum of 7 component variances. We show a breakdown of the variance into these 7 components in *Figure* (7.5).

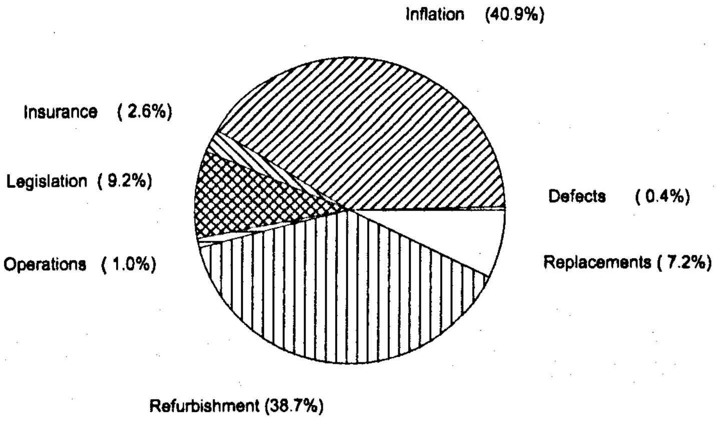

Figure 7.5: Variance components

We note from *Figure* (7.5) that inflation is still very important in driving the uncertainty, but with a much reduced influence from before. Inflation and refurbishment risks are the major contributors to the uncertainty, with substantial contributions from replacements and legislative risks. We also note the very small contribution from insurance, operations, and defects risks, explaining a total of just 4% of the variance. These groups will not be considered further.

Our final measure of assessing uncertainty is the distribution function. We plot the posterior expectation of the distribution function in *Figure* (7.6). The range of the plot is very similar to the distribution function we plotted in 7.2.1 (*Figure* (7.1)), however the upper tail is even more skewed now. The probability that the NPV is below £700M is very large, but there remains the small possibility that costs could spiral towards £900M.

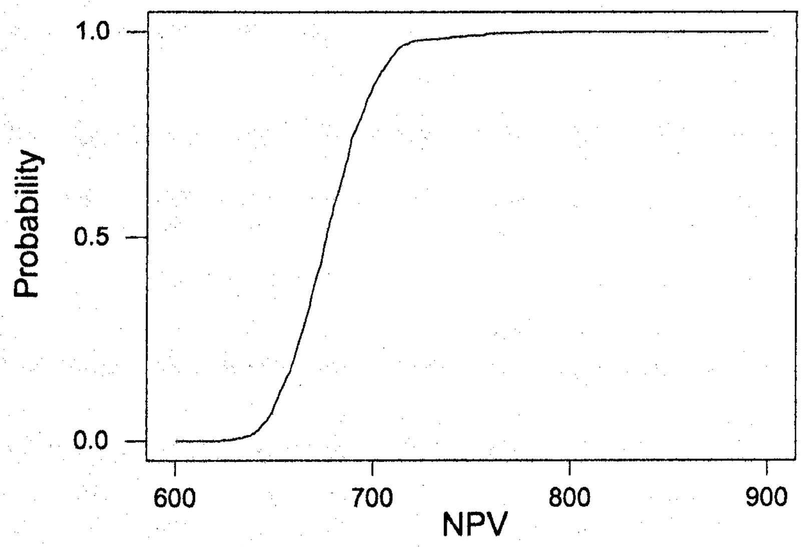

Figure 7.6: Distribution function

7.4.2 Sensitivity

Our sensitivity analysis produced markedly different results for a group comprised of refurbishment, replacements and legislation risks, and a second group consisting of just inflation risks. Consequently, we present results separately for these two groups.

We begin with inference with the first group, which comprised of refurbishment, replacements and legislation risks. All these inputs to the model were independent, and independent of inflations. We were able to perform a full variance based sensitivity analysis in addition to calculating main effects and interactions. We begin by plotting the main effects, standardized to be between 0 and 1. These are shown in *Figure* (7.7).

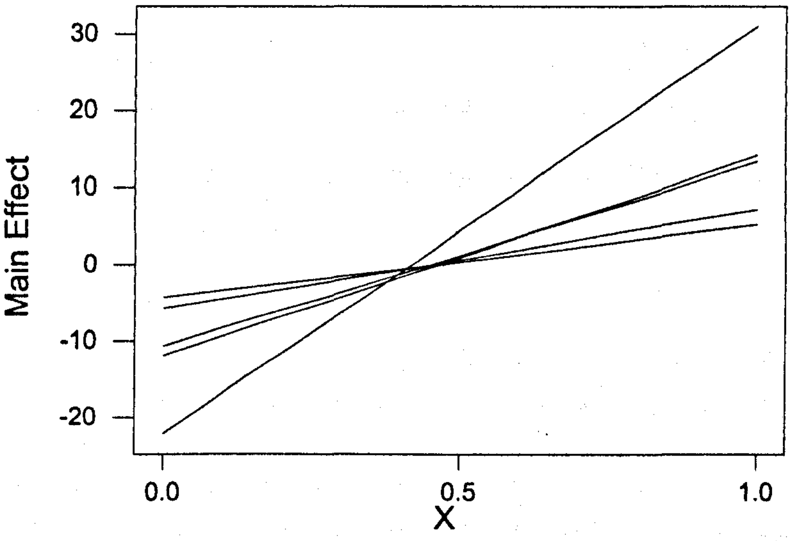

Figure 7.7: Main effects

We note that one main effect dominates; this is the capital redevelopment risk. The other large main effects were long term legislative risk and capital replacement risks, with small main effects for decant and short term legislative risks.

We can also calculate first order interactions. From the methodology developed in section 5.2, we know there are no interactions between redevelopment, replacement or legislative risks. Our analysis showed that any interactions within each group were also small. We found main effect variances, and calculated sensitivity indices of 0.365 for capital redevelopment risk, 0.021 for decant, 0.071 for redevelopment, 0.01 for short term legislation, and 0.081 for long term legislation.

The second group, comprised of inflations alone, had a far more complex structure due to the correlations between the inputs. All the TPI inflations were correlated, and all the GDP deflator inflations were correlated. A full variance based decomposition was not possible due to this structure, so we just show results for main effects and first order interactions.

The main effects of the inflations (both TPI and GDP deflator) were all relatively small. We plot 2 of these below, TPI 99/00 and GDP deflator 00/01.

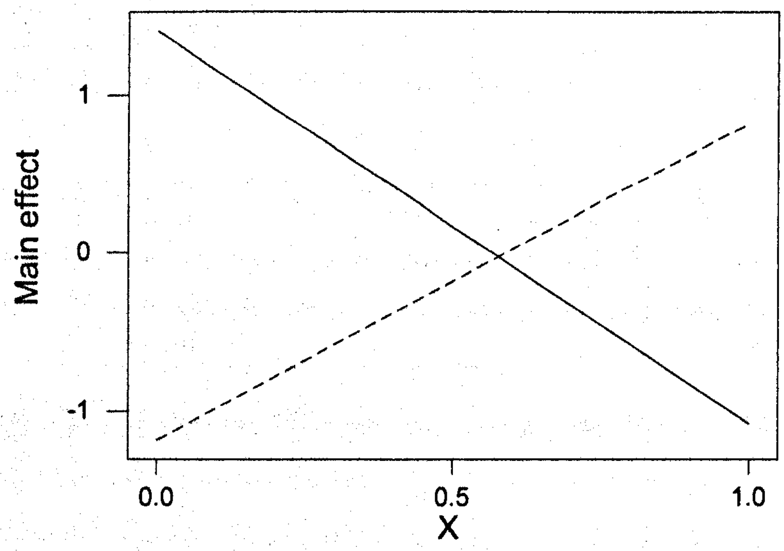

Figure 7.8: Main effects

We note the different sign on the gradients for the two main effects plotted in *Figure* (7.8). Large values of TPI 99/00 result in increases in the NPV, whilst the converse is true for GDP deflator 00/01. This is because the two inflations measure how prices on the project are increasing relative to prices in whole economy. If TPI = GDP deflator inflation, then the NPV would be unaffected by the magnitude of inflation. The two main effects plotted are representative of the behavior of the main effects of all years for these two measures of inflation. However our analysis showed that importance of the inflations in a given year was related to the capital expenditure in that year. Resultantly, the most important period of the project, with respect to the inflation risks, was the first 6 years of the project during the phase of major building works. The year of largest capital outlay was 02/03, and the main effects of TPI and GDP deflator were most pronounced in this year. Main effects after the first 6 years of the project, when redevelopment and hence the large capital investment was finished, were negligible.

Interactions were far more interesting. We found significant first order interactions between GDP deflator inflation in successive years, and as one might expect, large values of GDP deflator inflation in both years resulted in lower NPV of the project. We found that the first order interaction between GDP deflator inflations with a lag of d years between them, for $d \geq 1$, quickly decayed to zero for larger values of d, reflecting the autoregressive nature of the model. First order interactions between TPI inflations showed large values of TPI inflation in successive years had a compound effect, resulting in rises in the NPV. Lags between the TPI inflations resulted in similar behaviors to the GDP deflator inflations.

The most interesting interactions were between TPI and GDP deflator inflations. We show two of these here. In *Figure* (7.9) we show the first order interaction between X_1 = GDP deflator 00/01 and X_2 = TPI 99/00, and in *Figure* (7.10) we show the first order interaction between X_1 = GDP deflator 00/01 and $X_2 =$ TPI 00/01. From these two plots we can see that the NPV is affected by the relative effect of inflation, when one inflation is low and the other high, the relative inflation effect is greatest. From *Figure* (7.9) and *Figure* (7.10) we see the interaction between the two inflations is greatest within the same year. The largest interaction between these inflations is in 02/03, the year of largest capital expenditure.

7.5 Conclusions

In this chapter we have analysed a financial model that arose due to the Private Finance Initiative. We performed uncertainty and sensitivity analyses on the financial model. Our concluding remarks are in two parts; the first relating to the benefits of the methods we used; the second relating to the benefits or otherwise of the Private Finance Initiative.

7.5.1 Benefits of the analysis

In section 7.3 we presented results for the PSC model, created by *Coopers and Lybrand*, and showed the effect of the deterministic nature of the financial indices. In section 7.4 we demonstrated the assumptions made by *Coopers and Lybrand* were clearly erroneous and resulted in inflations having a greater influence than one would expect.

We also presented results for a revised model that made use of the methodology we developed in chapter 6. We cannot claim our method for modelling inflation is perfect, and time will no doubt highlight this. However, our model was an accurate reflection of an economic experts beliefs, and consistent with the Bank of England's short term model (symmetric about the target value and approximately

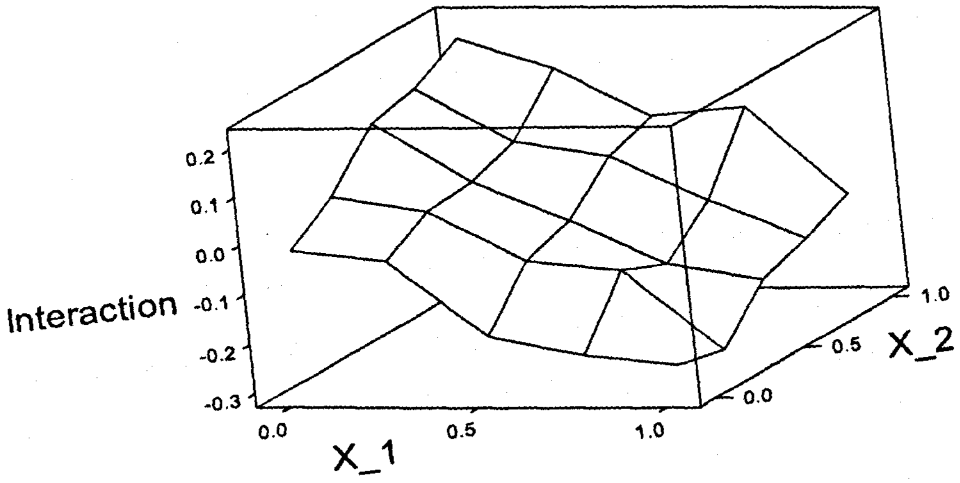

Figure 7.9: First order interaction

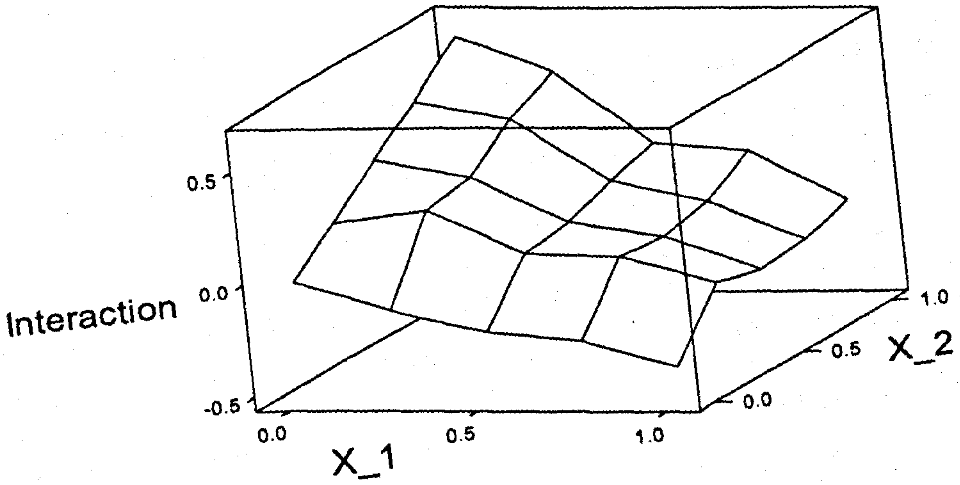

Figure 7.10: First order interaction

Gaussian). The net result was a far more plausible representation, that captured the main features of inflation – in particular the increase in TPI inflation relative to GDP deflator is more important than their magnitude. Using this model, we found the posterior expectation of the expected NPV was reduced by approx $\pounds 5M$, but the most striking difference was in the (posterior expectation of the) variance, which was reduced by $\pounds 90M$, a reduction of 20%.

We also demonstrated the methodology that we developed in chapter 5 – exploiting a known additive structure in order to produce more accurate results. An additive structure works most efficiently when the groups of inputs are of equal size and the sub functions of equal smoothness, so this example was by no means an 'optimal' case for our methodology – the inflations contained the majority of the inputs. The variances of each of the summaries we calculated would be lower than when using the Gaussian Process model we described in chapter 3 but we have not quantified this. The main benefit has been that the results of chapter 5 have allowed us to detect which interactions are exactly zero without need for further analysis, thus allowing us to direct resources to assessing non zero interactions.

7.5.2 The nature of PFI

Finally, we address the issue of PFI. The problem we are faced with is comparing a bid price from the private sector with a (usually) skewed distribution. Given the large variance and great uncertainty in the NPV, the bid price almost inevitably falls well within the bounds of the NPV distribution. Therefore any decision on how to fund a project is not simple. In a Bayesian setting, any decision can be addressed by utility, but this is not feasible in this case since the funding decision on a large project is made by a politician rather than an analyst, and political considerations as well as cost are taken into account.

For the MOD model the PFI bid was compared with the mean NPV, which is questionable given how skewed the distribution was. The PFI bid price was greater than the mean, and this was used to negotiate a discount before the deal was signed off. The PSC was used to show the public achieved value for money. Our model would have resulted in a greater discount. Building works have since been completed on this project, and the cost of these indicate that the risks were overestimated in the PSC.

We end this discussion by reiterating the concerns made by *House of Commons Treasury Committee*(30). For this project, and many other projects, public sector funding is simply not a realistic option. The PSC is simply created as a negotiating tool. It seems sensible to assume this has an effect on the magnitude of PFI bid prices.

Chapter 8

Discussion

In this thesis we have considered the uncertainty in Government financial models, a special type of computer model, that arises when we have uncertainty on (some of) the model inputs. The methodology we have developed has been in two distinct parts; we first developed methodology for function approximation, uncertainty analysis and sensitivity analysis for a decomposable computer model; we then considered the uncertainty in a particular set of inputs, inflations, and developed methodology for quantifying this uncertainty, based upon expert opinion. We will discuss these two components of the thesis separately before making some remarks about how both structures can be used for very large process models, with particular reference to the London Underground financial model.

In chapter 3 we examined a previously proposed but untried correlation function. The function had a positive semi-definite matrix $\mathbf{\Omega}$ of parameters, and resultantly, at least in principle, had far more flexibility than when modelling correlation as a product of 1 dimensional correlation functions (corresponding to $\mathbf{\Omega}$ diagonal). However, we found that as a consequence of greater flexibility we had a significantly increased computational burden in estimating $\mathbf{\Omega}$ from its posterior mode, resulting from a higher dimensional numerical search and a flatter posterior distribution. Furthermore, we discovered that whether or not our more flexible correlation function was worth the additional computational burden was related to the form of $\eta(\mathbf{x})$. For a function with additive or nearly additive inputs, that is where interactions are small compared with the main effects, the correlation was modelled well by a diagonal form for $\mathbf{\Omega}$, whilst for models with large interactions the more complex form could be worth the additional computational burden if a single evaluation of $\eta(.)$ was expensive.

We showed that matrix $\mathbf{\Omega}$ corresponded to a diagonal matrix of parameters on a transformed scale given by $\mathbf{z} = \mathbf{C}\mathbf{x}$, where $\mathbf{x}$ is $p \times 1$, $\mathbf{z}$ is $r \times 1$, $\mathbf{C}$ is $r \times p$, with the dimension r of the new coordinate system $\leq p$. This result demonstrates that particularly when we have interactions, the most efficient coordinate system with which to model $\eta(.)$ may not have orthogonal axes. We noted that in principle we could estimate $\mathbf{C}$ such that coordinate system $\mathbf{z}$ had an approximately diagonal matrix of parameters. However, specifying this transformation would be difficult even with expert knowledge since by writing $\eta(.)$ as the sum of a regression fit, $\mathbf{h}(\mathbf{x})^{\mathbf{T}}\boldsymbol{\beta}$, and a zero mean Gaussian Process, we explained in chapter 4 that the Gaussian process corrects the regression fit such that the model interpolates the data. Therefore $\mathbf{C}$ would depend upon the form of $\mathbf{h}(\mathbf{x})^{\mathbf{T}}\boldsymbol{\beta}$. More research would be needed in this area in order to identify if this approach is feasible in practice.

The developments we made in chapter 3 lead us to consider how we might better be able to model $\eta(.)$ when a decomposition into a sum of lower dimensional terms was possible. We found that for this case more efficient methods of modelling the correlation are available, and we developed methodology to exploit this. We considered both known and unknown decompositions of $\eta(.)$.

For a known decomposition of $\eta(.)$ we first developed the methodology for when each sub function was observable, finding posterior distributions for each sub function and developing a fast and accurate approximation to the posterior distribution of $\eta(\mathbf{x})$. We also developed methodology for when each sub function was not observable, exploiting a one at a time design structure – this latter case was only possible for an additive model and required very strong prior beliefs about $\eta(.)$. In chapter 5 we extended the calculations of *Oakley and O'Hagan*(51) so that we could calculate measures of uncertainty and sensitivity for models of this form. In general our measures will be more precise, with less uncertainty about posterior expectations than is the case in *Oakley and O'Hagan*, using fewer design points. We showed that a decomposition combined with at least partial independence of the inputs is a powerful result when calculating measures of sensitivity. In this case, not only can we obtain more precise inferences, but we can establish exactly which interactions are exactly zero without need for calculations. This result has particular importance when interpreting total effect sensitivity indices.

We also developed methodology for the case when we did not exploit a one at a time design. The methodology was developed for the case when we had uncertainty or even complete ignorance about a decomposition, and for partially additive models that were unable to exploit a one at a time design. In chapter 4 we went on to discuss methodology that could be used to verify if we had identified a correct decomposition, and finally we showed how we could use the parameters of the correlation functions in order to search for decompositions of $\eta(.)$. We were able to show that our methodology worked well for smooth functions. For rougher functions, as might be expected, we require more design points in order to search for secompositions, but our method is still practical. We were able to extend the calculations of *Oakley and O'Hagan* in order to perform uncertainty and sensitivity analyses with greater precision. We did this in chapter 5.

The use of structural information is a new innovation, and whilst the approach appears to show much promise, in order to adequately assess the full uses and limitations of the methodology, further research is required. We discuss some of the more important aspects that we have not developed in this thesis, below.

The first, and perhaps most important issue that we have not considered in the thesis is design points, and in particular how structural information might be utilized in order to develop more efficient designs. We discussed different criteria for choosing design points in chapter 3, with 4 design criteria that have been used in a computer models context discussed. A common problem to all criteria was that the parameters of the correlation function were unknown a priori. We explained in chapter 3 that a two phased approach to selecting design points had been proposed and implemented; the first phase to estimate the parameters of $c(.,.)$, and the second phase using some some design criterion that made use of these estimates. A decomposable model, where the decomposition is unknown could in principle use a similar two phased design. We could observe n' outputs selected using a Latin hypercube design in order to determine the decomposition of $\eta(.)$, and estimate the parameters of the correlation functions. We could then use some design criterion in order to select the remaining $n-n'$ design points. Further research is required in order to determine how large we would require n' to be in order to identify the decomposition however our limited amount of numerical work has indicated that this will be determined by the number of active dimensions, the complexity of the function and the number of terms in the decomposition.

A second area for further research is to investigate how well our method can discriminate between an additive or partially additive function, and a function which is very close to additive or partially additive, with a very small interaction. We described one example in chapter 4, in which we could successfully discriminate between these two cases. However, based on our small amount of numerical work we cannot claim our method will always be able to successfully discriminate. More research is required, especially for the more difficult partially additive case.

A related problem of interest is a function where we don't have a global interaction between inputs, a simple example of which is

$$\begin{aligned} y = \eta_1(x_{(1)}) + \eta_2(x_{(2)}) \qquad & x_1 \quad < c \\ y = \eta(x_1, x_2) \qquad & x_1 \quad > c. \end{aligned}$$

A problem like this might arise in a financial model when an unacceptable level of performance (as measured by one of the model inputs) results in financial penalties. Other examples will also exist in models representing physical systems (for example a model containing chemical interactions). An interesting area for future research would look into identifying how well our method could distinguish between an additive model and this scenario, especially if c was such that the additive model was correct for most values of x_1.

A final case of interest is not related to deterministic functions. Suppose we had a model of the form

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

for stochastic error term ϵ, taken to be normally distributed with zero mean and variance σ_ϵ^2. In this model the relationship between the inputs and the output cannot be described adequately by a parametric form, so a non parametric model is used. This non parametric regression model smoothes rather than interpolates the observed data. A model of this form was first proposed in a regression context by *O'Hagan*(52). We might have an interest in whether two (or more) of the inputs are interacting. Given the noise from the error term, this model would represent a difficult challenge even for low dimensional $\mathbf{x}$. Modifying our approach for this situation would be an interesting area for future research.

The second part of the thesis developed a Bayesian autoregressive model, and we did this in chapter 6. The motivation behind this research was to quantify the uncertainty in inflation in the long term, when only a small amount of data and the subjective knowledge of an expert was available. We developed methodology that required just medians and quartiles, although we did investigate the use of tail assessments. We found tail judgements to be unreliable. Evidence from the literature suggests that with sufficient training, subjects can improve their assessment of tail probabilities, although large scale empirical work has not been based on expert opinion. There is a real need for further empirical work in the area of training experts to assess tail probabilities.

We end this thesis by discussing how our methodology might be applicable for very large Government financial models, like the LU model, which contained thousands of inputs. The costs were so vast in this project, over the whole period of the 30 year contract, that the risks for a particular part of the project (i.e. track and signal replacements on the Northern Line) were modelled on a yearly basis in the financial model. A crude correlation structure was used in order to take into account the serial correlation in these risks between years. Individual aspects of the project were modelled independently. Therefore, in effect, the model is a sum of many lower dimensional models, and we have shown in this thesis that lower dimensional sub-functions may be modelled independently. For a model with this structure, even with thousands of inputs, the two components of this thesis could be in principle be used in order to model the cost of the project, and to perform uncertainty and sensitivity analyses. However, far more work would be required before attempting to apply our methodology to models of this scale.

Bibliography

[1] Al-Awadhi, S.A. and Garthwaite, P.H. (2001). Prior Distribution Assessment for a Multivariate Normal Distribution: An Experimental Study. *Journal of Applied Statistics*, 28.

[2] Alpert, M. and Raiffa, H. (1982). A progress Report on the Training of Probability Assessors. In D. Kahneman and P. Slovic and A. Tversky, editor, *Judgement Under Uncertainty: Heuristics and biases.* Cambridge University press.

[3] Ball, R. and Heafey, M., and King, D. (2000). Private Finance Initiative – good for the public purse or a drain on future generations? *Policy and Politics*, 29(1):95–108.

[4] Bayarri, M. and Berger, J. and Higdon, D. and Kennedy, M.C. and Kottas, A. and Paulo, R. and Sacks, J. and Cafeo, J. and Cavenish, J. and Tu, J. (2002). Validation of Computer Models. In D. Pace and S.Stevenson, editor, *Proceedings of the Workshop on Foundations for Validation in the 21st Century, Society for Modeling and simulation International.*

[5] Blight, B. J. N. and Ott, L. (1975). A Bayesian Approach to Model Inadequacy for Polynomial Regression. *Biometrika*, 42:79–88.

[6] Campolongo, F. and Braddock, R. (1997). The use of Graph Thoery in the Sensitivity Analysis of the Model Output: a new Screening Method. *Reliability Engineering and System Safety*, 64.

[7] Campolongo, F. and Kleijnen, J. and Andres, T. (2000). Screening Methods. In A. Saltelli and K. Chan and E.M. Scott, editor, *Sensitivity Analysis*. London: John Wiley and Sons.

[8] Campolongo, F. and Tarantola, S. and Saltelli, A. (1999). Tackling Quantitatively Large Dimensionality Problems. *Computer Physics Communications*, 117.

[9] Chan, K. and Saltelli, A. and Tarantola, S. (2000). Winding Stairs: A Sampling Tool to Calculate Sensitivity Indices. *Statistics and Computing*, 10(3).

[10] Chan, K. and Saltelli, A. and Tarantola, S. and Sobol', I.M. (2000). Variance Based Methods. In A. Saltelli and K. Chan and E.M. Scott, editor, *Mathematical and Statistical Methods for Sensitivity Analysis*. London: John Wiley and Sons.

[11] Chatfield, C., editor (2004). *The Analysis of Time Series: An Introduction.* Chapman and Hall.

[12] Clemen, R.T. and Fischer, G.W. and Winkler, R.L. (2000). Assessing Dependence: Some Experimental Results. *Management Science*, 46(8).

[13] Cukier, I.R. and Fortuin, C.M. and Schuler, K.E. and Petschek, A.G. and Schaibly, J.H. (1973). Study of the Sensitivity of Coupled Reactions Systems to Uncertainties in Rate Coefficients. *The Journal of Chemical Physics*, 59.

[14] Cukier, I.R. and Levine, H.B. and Schuler, K.E. (1978). Nonlinear Sensitivity Analysis of Multiparameter Model Systems. *Journal of Computational Physics*, 26.

[15] Currin, C. and Mitchell, T.J. and 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.

[16] Feller, W., editor (1966). *An Introduction to Probability and it's Applications.* New York:Wiley.

[17] Froud, J. (2003). The Private Finance Initiative: Risk, Uncertainty and the State. *Accounting, Organizations and Society*, 28(6).

[18] Garthwaite, P.H. and Dickey, J.M. (1988). Quantifying Expert Opinion in Linear Regression problems. *Journal of the Royal Statistical Society, series B*, 50(3).

[19] Garthwaite, P.H. and Dickey, J.M. (1992). Elicitation of prior Distributions for Variable Selection Problems in Regression. *Annals of Statistics*, 20.

[20] Garthwaite, P.H. and Kadane, J.B. and O'Hagan, A. (2005). Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association*, 100.

[21] Gosling, J.P. (2005). *Elicitation: A Nonparametric View.* PhD thesis, University of Sheffield.

[22] Haylock, R. (1997). *Bayesian Inference About Outputs of Computationally Expensive Algorithms with Uncertainty on the Inputs.* PhD thesis, University of Nottingham.

[23] Haylock, R. and O'Hagan, A. (1996). On Inference for Outputs of Computationally Expensive Algorithms with Uncertainty on the Inputs. In J. M. Bernado and J. O. Berger and A. P. Dawid and A. F. M. Smith, editor, *Bayesian Statistics 5.* Oxford: University Press.

[24] Heald, D. (1997). Privately Financed Capital in Public Services. *The Manchester School.*

[25] Heald, D. (2003). Value for Money Tests and Accounting Treatment in PFI Schemes. *Accounting, Auditing and Accountability Journal*, 16(3):342–371.

[26] Helton, J.C. and Davis, F.J. (2000). Sampling-based Methods. In A. Saltelli and K. Chan and E.M. Scott, editor, *Sensitivity Analysis.* London: John Wiley and Sons.

[27] Helton, J.C and Davis, F.J (2002). Illustration of Sampling-based Methods for Uncertainty and Sensitivity Analysis. *Risk Analysis* , 22(3).

[28] HM Treasury (2000). Technical Note Number 5: How to construct a Public Sector Comparator.

[29] Homma, T. and Saltelli, A. (1996). Importance Measures in Global Sensitivity Analysis of Model Output. *Reliability Engineering and System Safety,* 52.

[30] House of Commons Treasury Committee (2000). The Private Finance Initiative. Forth Report, Session 1999-2000.

[31] Iman, R.L. and Conover, W.J. (1982). A Distribution-free Approach to Inducing Rank Correlation Among Input Variables. *Communications, Statistics, Simulation and Computation B*, 11.

[32] Jansen, M.J.W. and Rossing, W.A.H. and Daamen, R.A. (1994). Monte-carlo Estimation of Uncertainty Contributions from Several Independent Multivariate Sources. In J. Gasman and G van Straten, editor, *Predictability and Nonlinear Modelling in Natural Sciences and Economics.* Kluwer Academic Publishers.

[33] Johnson, N.L. and Kotz, S. and Balakrishnan, N., editor (1994). *Continuous Univariate Distributions.* Wiley.

[34] Kadane, J. B. and Chan, N.H. and Wolfson, L.J. (1996). Priors for Unit Root Models. *Journal of Econometrics*, 75.

[35] Kadane, J. B. and Dickey, J.M. and Winkler, R.L. and Smith, W.S. and Peters, S.C. (1980). Interactive Elicitation of Opinion for a Normal Linear Model. *Journal of the American Statistical Association*, 75(372).

[36] Kennedy, M.C. and O'Hagan, A. (2001). Bayesian Calibration of Computer Codes. *Journal of the Royal Statistical Society: Series B*, 63:425–464. (Read before the Royal Statistical Society).

[37] Koda, M. and McRae and Seinfeld, J.H (1979). Automatic Sensitivity Analysis of Kinetic Mechanisms. *International Journal of Chemical Kinetics*, 11.

[38] Lord Lawson (2002). Presentation at the IAEE International Conference. Internet. URL : http://www.iaee.org/documents/a02lawson.pdf.

[39] McKay, M.D. (1997). Nonparametric Variance-based methods of Assessing Uncertainty Importance. *Reliability Engineering and System Safety*, 57:267–279.

[40] McKay, M.D. and Conover, W.J. and Beckman, R.J. (1979). Comparison of 3 methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21.

[41] McKay, M.D. and Morrison, J.D. and Upton, S.C. (1999). Evaluating prediction Uncertainty in Simulation Models. *Computr physics Communications*, 117:44–51.

[42] Monbiot, G. (2002). Very British Corruption. *Guardian.* 22 January.

[43] Morris, M.D. (1991). Factorial Sampling Plans for Preliminary Computational Experiments. *Technometrics*, 33.

[44] Mosteller, F. and Yountz, C. (1990). Quantifying Probabilistic Expressions. *Statistical Science*, 5(1). (with discussion).

[45] Neal, R. (1999). Regression and Classification Using Gaussian Process Priors. In J. M. Bernado and J. O. Berger and A. P. Dawid and A. F. M. Smith, editor, *Bayesian Statistics 6.* Oxford: University Press.

[46] Nelder, J.A. and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7:308–313.

[47] Oakley, J. (1999). *Bayesian Uncertainty Analysis for Complex Computer Codes.* PhD thesis, University of Sheffield.

[48] Oakley, J. (2002). Eliciting Gaussian Process Priors for Complex Computer Codes. *The Statistician*, 51:81–97.

[49] Oakley, J. and O'Hagan, A. (2002a). Bayesian Inference for the Uncertainty Distribution of Computer Model Outputs. *Biometrika*, 89(4).

[50] Oakley, J. and O'Hagan, A. (2002b). Uncertainty in Prior Elicitations. Technical report, University of Sheffield.

[51] Oakley, J. and O'Hagan, A. (2004). Probabilistic Sensitivity Analysis of Complex Models: A Bayesian Approach. *Journal of the Royal Statistical Society: Series B*, 66(3).

[52] O'Hagan, A. (1978). Curve Fitting and Optimal Design for Prediction. *Journal of the Royal Statistical Society: Series B*, 40:1–46.

[53] O'Hagan, A. (1991). Bayes Hermite Quadrature. *Journal of Statistical Planning and Inference*, 29:245–260.

[54] O'Hagan, A. (1992). Some Bayesian Numerical Analysis. In J. M. Bernado and J. O. Berger and A. P. Dawid and A. F. M. Smith, editor, *Bayesian Statistics 4*. Oxford: University Press.

[55] Public Services International Research Unit (2001). Uk Water Privitisation: A Briefing. Internet. URL : http://www.psiru.org/reports/2001-02-W-UK-over.doc.

[56] Raiffa, A. and Schlaifer, R., editor (2000). *Applied Statistical Theory.* Wiley.

[57] Sacks, J. and Schiller, S. (1988). Spatial Designs. In S.S. Gupta and J.O. Berger, editor, *Statistical Decision Theory and related Topics.* Springer, New York.

[58] Sacks, J. and Welch, W.J. and Mitchell, T.J. and Wynn, H.P. (1989). Design and Analysis of Computer Experiments. *Statistal Science*, 4:409–435.

[59] Sacks, J. and Welch, W.J. and Schiller, S.B. Designs for Computer Experiments. *Technometrics.*

[60] Saliby, E. and Pacheco, . (2002). An Empirical Evaluation of Sampling Methods in Risk Analysis Simulation: Quasi-Monte Carlo, Descriptive Sampling

and Latin Hypercube Sampling. In Yucesan, E. and Chen, C.H and Snowdon, J.L and Charnes, J.M, editor, *Proceedings of the 2002 Winter Simulation Conference.* I E E E, New York.

[61] Saltelli, A. and Tarantola, S. and Chan, K. (1999). A Quantitative Model Independent Method for Global Sensitivity Analysis of Model Output. *Technometrics*, 41(1).

[62] Sobol', I. M. (1993). Sensitivity Analysis for Nonlinear Mathematical Models. *Mathematical Modeling and Computational Experiment*, 1.

[63] Sobol', I.M. (2001). Global Sensitivity Indices for Nonlinear Mathematical Models and their Monte-Carlo Estimates. *Mathematics and Computers in Simulation*, 55(1).

[64] Welch, W.J. and Buck, R.J. and Sacks, J. and Wynn, H.P. and Mitchell, T.J. and Morris, M.D. (1992). Screening, Predicting, and Computer Experiements. *Technometrics*, 34.

[65] Wolfson, L.J. (1995). *The Elicitation of Priors and Utilities for Bayesian Analysis.* PhD thesis, Carnegie Mellon University.