The University of Southampton

Faculty of Social and Human Sciences

Optimal and efficient experimental design for nonparametric regression with application to functional data

Verity Alexandra Fisher

Doctor of Philosophy

December 2012

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

Statistics

Doctor of Philosophy

OPTIMAL AND EFFICIENT EXPERIMENTAL DESIGN FOR NONPARAMETRIC REGRESSION WITH APPLICATION TO FUNCTIONAL DATA

Verity Alexandra Fisher

Functional data is ubiquitous in modern science, technology and medicine. An example, which motivates the work in this thesis, is an experiment in tribology to investigate wear in automotive transmission.

The research in this thesis provides methods for the design of experiments when the response is assumed to be a realisation of a smooth function. In the course of the research, two areas were investigated: designs for local linear smoothers and designs for discriminating between two functional linear models.

Designs that are optimal for minimising the prediction variance of a smooth function were found across an interval using two kernel smoothing methods: local linear regression and Gasser and Müller estimation. The values of the locality parameter and run size were shown to affect the optimal design. Optimal designs for best prediction using local linear regression were applied to the tribology experiment. A compound optimality criterion is proposed which is a weighted average of the integrated prediction variance and the inverse of the trace of the smoothing matrix using the Gasser and Müller estimator. The complexity of the model to be fitted was shown to influence the selection of optimal design points. The robustness of these optimal designs to misspecification of the kernel function for the compound criterion was also critically assessed.

A criterion and method for finding T-optimal designs was developed for discriminating between two competing functional linear models. It was proved that the choice of optimal design is independent of the parameter values when discriminating between two nested functional linear models that differ by only one term. The performance of T-optimal designs was evaluated in simulation studies which calculated the power of the test for assessing the fit of one model using data generated from the competing model.

Contents

List of Figures

List of Tables

Declaration Of Authorship

I, Verity Alexandra Fisher, declare that the thesis entitled 'Optimal and efficient experimental design for nonparametric regression with application to functional data' and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the theis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed:

Date:

Acknowledgements

This work was supported by a research studentship from the Engineering and Physical Sciences Research Council.

I would like to thank my supervisors, David Woods and Susan Lewis, for their invaluable advice and assistance throughout my studies.

I would also like to thank my friends for putting up with all my statistical monologues in their free time. Finally I would like to thank Sue, Simon, Oliver and Jessica Fisher for their unwavering support and always being at the other end of the phone.

Chapter 1

Introduction

Increasingly, data from experiments in science and technology are used to investigate complex systems where the response function cannot be adequately approximated by a simple regression function such as a low-order polynomial. Then, nonparametric regression is preferred where no assumption is required on the form of the regression function. The research in this thesis provides methods for the design of experiments when the response is assumed to be a realisation of a smooth function.

In the course of the research, the following two issues are addressed:

(i) Little research is available on how to design an experiment to obtain as much information as possible from available resource. The first broad aim of the research is to give methods of finding highly efficient or optimal designs for nonparametric regression.

(ii) In many experiments, functional data are collected where the response from each run is a realisation of a smooth function of a continuous variable, such as time, as opposed to a scalar value. Then, each function may require estimation using nonparametric regression. An example of such an experiment is given in Section 1.3. In practical applications, we may need to make a decision about which of two models provides the better description of a response. The second broad aim of this research is to provide methods of designing experiments that enable discrimination between two competing functional linear models.

1.1 Models and linear smoothers

1.1.1 Modelling and estimation for nonparametric smoothing

We consider the nonparametric regression model which describes a response variable by

$$y_j = g(x_j) + \epsilon_j, \text{ for } j = 1, \ldots, n, \tag{1.1}$$

where g is the unknown regression function, x_j is the value of the single explanatory variable, and the ϵ_j are independent error random variables which are identically distributed with constant variance σ^2. This model is often called the 'fixed design model', see for example Wand and Jones (1995, ch. 5). It is different from the 'random design model' in which the observations are regarded as a random sample from a bivariate distribution. The work in this thesis considers only fixed design models.

Estimation of $g(x)$ in (1.1) is through a linear smoother, $\hat{g}(x)$, which is a weighted linear combination of the observations y_j expressed as

$$\hat{g}(x) = \sum_{j=1}^{n} S_j(x) y_j, \tag{1.2}$$

where $S_j(x)$ are the smoothing weights (see, for example, Ramsay and Silverman, 2005, ch. 4). A simple example of a linear smoother is linear regression. Further examples can be found in Buja, Hastie and Tibshirani (1989), Wand and Jones (1995, ch. 5), Simonoff (1996, ch. 5) and Ramsay and Silverman (2005, ch. 4).

1.2 Functional linear models

If an experiment produces functional data, then a functional linear model may be used to describe the response functions. This model is written in matrix form as

$$\boldsymbol{Y}(t) = X\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t), \tag{1.3}$$

where $\boldsymbol{Y}(t)$ is an $n \times 1$ vector of response functions, X is the $n \times p$ model matrix, $\boldsymbol{\beta}(t)$

is an $n \times 1$ vector of parameters and $\boldsymbol{\epsilon}(t)$ is an $n \times 1$ vector holding realisations from a stochastic process with mean zero and covariance function $\gamma(s,t)$; for $a \leq s \leq t \leq b$ and $[a,b] \subset \mathbb{R}$.

1.3 Example from tribology

To motivate the theory and results in this thesis, we consider a common type of experiment from the EPSRC National Centre for Advanced Tribology (nCATS) at the University of Southampton. The experiment is a pilot study to assess how six factors affect the wear of a pin and disc assembly when a given lubricant is used to lubricate the surface of the disc.

The experiment involved 16 runs, each having a different combination of values of six factors. Each of the factors is listed below, together with the two factor levels used in the experiment.

- disc material: silicone or steel
- pin material: silicone or steel
- addition of soot: 0% or 10%
- level of oxidation: 0 or 10 hours
- addition of H_2S0_4: 0 or 25mM (millimolar)
- level of moisture: 0% or 2.5%

In addition, the tribologists ran four model checking runs. In this thesis, where we label runs from the experiment we use the labels from the randomised order of the 20 runs.

Figure 1.1 shows a schematic of the pin and disc equipment. The gimbal arm suspends the pin over the disc. The disc spins and the combined wear on the pin and the disc is measured by a Linear Variable Displacement Transformer at a large number of equally spaced discrete time points (referred to as the time index). The first 500 observations are typically discarded as 'burn in'. In our particular motivating experiment, all observations after the 2400th were disregarded as, for some runs, the equipment was erroneously left on after the experiment had finished, producing spurious results. Figure 1.2 shows wear data produced by this experiment.

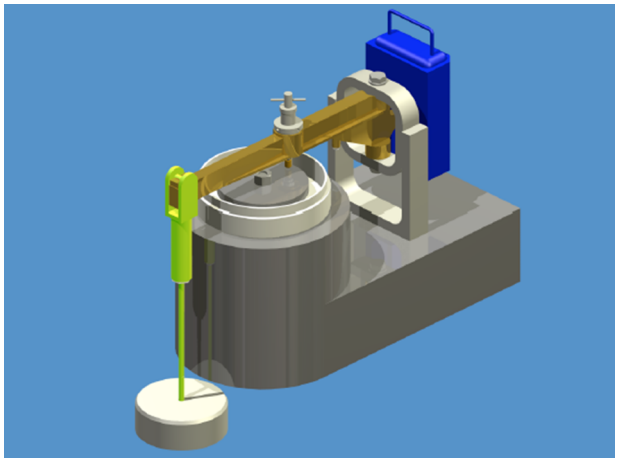

Figure 1.1: Schematic of the pin and disc equipment.

The aim of the experiment is to predict the value of the response, that is, the profile over the interval [500, 2400] of the combined wear on the pin and disc for a given combination of values of the six factors. This is to be achieved by using an optimal design consisting of the 'best' subset of points $\{x_1, \ldots, x_n\}$ selected from the interval [500, 2400].

1.4 Design preliminaries

The following section describes terms and ideas from the field of design of experiments which will be used later in the thesis. A detailed account of the optimal design of experiments, including various optimality criteria, with application to linear and non-linear models can be found in Atkinson, Donev and Tobias (2007, ch. 10).

There are two approaches to design specification:

(a) A continuous, or approximate, design which is represented by a measure ξ on a design region χ and written as

$$\xi = \left\{ \begin{array}{cccc} x_1 & x_2 & \ldots & x_s \\ w_1 & w_2 & \ldots & w_s \end{array} \right\}, \tag{1.4}$$

where, without loss of generality, x_i $(i = 1, \ldots, s)$ are the s distinct design points and w_i $(i = 1, \ldots, s)$ are the associated design weights. Each distinct design point is called a support point and its weight specifies the proportion of total experimental effort to be expended at that point. Since ξ is a measure, $\int_\chi \xi \, \mathrm{d}x = 1$ and the design weights are restricted to $0 < w_i \leq 1$, $i = 1, \ldots, s$, with $\sum_{i=1}^{s} w_i = 1$.

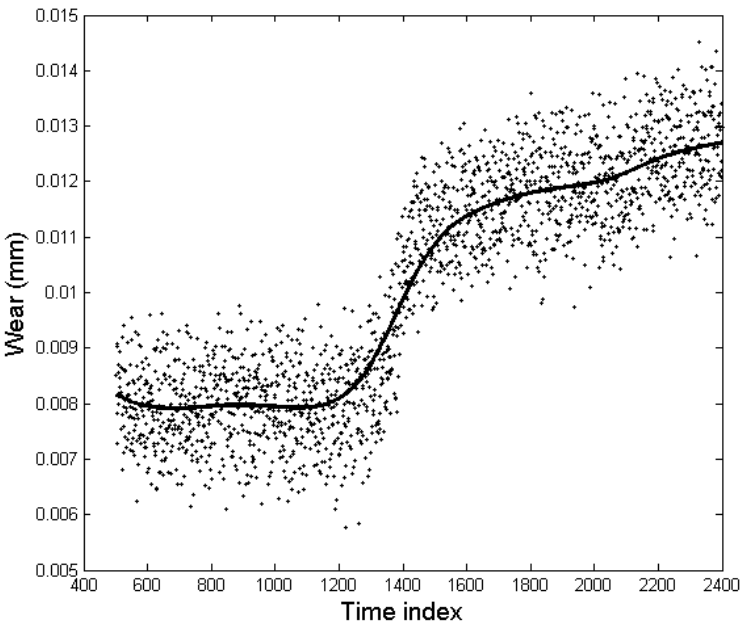

Figure 1.2: Plot of data from one run of the wear experiment with an example locally linear smooth fit.

For example, for the continuous design

$$\xi = \left\{ \begin{array}{ccc} 0 & 1 & 2 \\ 0.3 & 0.3 & 0.4 \end{array} \right\},$$

in $n = 10$ runs, the design used for an experiment would have three runs at each of $x = 0$ and 1, and four runs at $x = 2$. When nw_i is non-integer, for some $i = 1, \ldots, s$, then the value must be rounded to provide a realisable design with an integer number of occurrences of each x_i in the design; see Fedorov and Hackl (1997, ch. 1).

(b) An exact design which has s support points and n design points and may be written

$$\xi_n = \left\{ \begin{array}{cccc} x_1 & x_2 & \ldots & x_s \\ r_1 & r_2 & \ldots & r_s \end{array} \right\}, \tag{1.5}$$

where r_i is the integer number of observations taken at the ith distinct design point, x_i, and $\sum_{i=1}^{s} r_i = n$.

For simplicity, we usually write an exact design as $\{x_1, x_2, \ldots, x_n\}$ where the x_i are not necessarily distinct.

For both approaches, we define an optimal design as one which optimises an objective function $\Psi(\xi)$ or $\Psi(\xi_n)$, for a continuous or exact design, respectively.

1.5 Objectives and overview of thesis

This thesis has two specific objectives. The first is to develop methods for the efficient design of experiments for local linear smoothing and for Gasser and Müller kernel smoothing. The second objective is to develop a T-optimality criterion and derive analytical results to enable the design of experiments for efficient discrimination between two functional linear models.

In Chapter 2, we introduce common linear smoothers and discuss the existing literature on experimental design. We then give, in Chapter 3, methods of finding designs for prediction using the local linear estimator for two cases:
(i) prediction at a finite number of points in the design region, and
(ii) prediction across the whole of an interval within the design region.

In Chapter 4, we find optimal and efficient designs for the Gasser and Müller estimator when the purpose of the experiment is prediction across an interval. We compare the performance of designs found using the uniform and the Gaussian kernels (defined in Section 2.1).

Chapter 5 develops, for the first time, optimal designs for 'best' discrimination between two functional linear models using a T-optimality criterion developed for this class of models. Chapter 6 evaluates the work and methods in this thesis and highlights avenues for future work.

Chapter 2

Background to linear smoothing and design of experiments literature

This chapter provides an introduction to local linear smoothing and a review of the literature on the design of experiments for these smoothing methods.

In Section 2.1, we give a brief description and background for nonparametric methods of local linear smoothing, in particular kernel smoothing, which can be used to estimate the function, $g(x)$, in model (1.1). The estimate $\hat{g}(x)$ is calculated using (1.2), where different forms of smoothing weights, $S_j(x)$, are used for different types of local smoothing. In Section 2.2, we discuss the limited literature on optimal design for local fitting using the local linear estimator.

2.1 Local fitting

Generally, local fitting describes methods of estimating $g(x)$ such that observations at points closer (or more local) to x have larger influence on $\hat{g}(x)$. Popular local smoothing methods use kernel regression, spline functions and wavelets. In this section we consider two approaches using kernel regression estimators: local polynomial estimators and the Gasser and Müller estimator.

Kernel regression methods, a form of local linear smoothing, were first considered by Nadaraya (1964) and Watson (1964) and later modified by Priestly and Chao (1972) and Gasser and Müller (1979, 1984).

This type of local linear smoother is defined through the choice of smoothing weights, $S_j(x)$, see equation (1.2), which determine the distribution of the weights assigned to each observation y_j in $\hat{g}(x)$. The form of $S_j(x)$ depends on a pre-specified constant h, known as the bandwidth.

A choice of smoothing weight distribution which incorporates the bandwidth may be obtained from using a kernel function, $K(\cdot)$. Such functions are symmetric and have the property that $\int K(u)\,\mathrm{d}u = 1$. Some widely used kernel functions are given below and shown in Figure 2.1:

$$\text{Uniform:} \quad K(u) = \begin{cases} 0.5 & \text{if } |u| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{Epanechnikov:} \quad K(u) = \begin{cases} 0.75(1-u^2) & \text{if } |u| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{Gaussian:} \quad K(u) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\} \quad -\infty < u < \infty,$$

The jth smoothing weight is defined as $K(u_j)$ where $u_j = x - x_j$. Hence both the kernel and bandwidth affect the degree of locality in $\hat{g}(x)$ in the following sense:

- If $|x - x_j| \leq h$, then observation y_j has a substantial weight $S_j(x)$, which is a monotonically decreasing function of $|x - x_j|$.
- If $|x - x_j| > h$, then $S_j(x) = 0$ or decreases monotonically with $|x - x_j|$.

The Epanechnikov kernel function, has desirable asymptotic properties and so is a popular kernel function choice (Simonoff, 1996, ch. 5). However, in this thesis, we use the uniform and Gaussian kernel functions. The Gaussian kernel, which is widely used, is not truncated and so $S_j(x) \neq 0$ for all j. We use the uniform kernel, together with the Gaussian kernel, in a study (Sections 3.5.3 and 4.6) of the robustness of designs to the choice of kernel function. We consider the uniform kernel as its kernel function has a significantly different form from that of the Gaussian kernel.

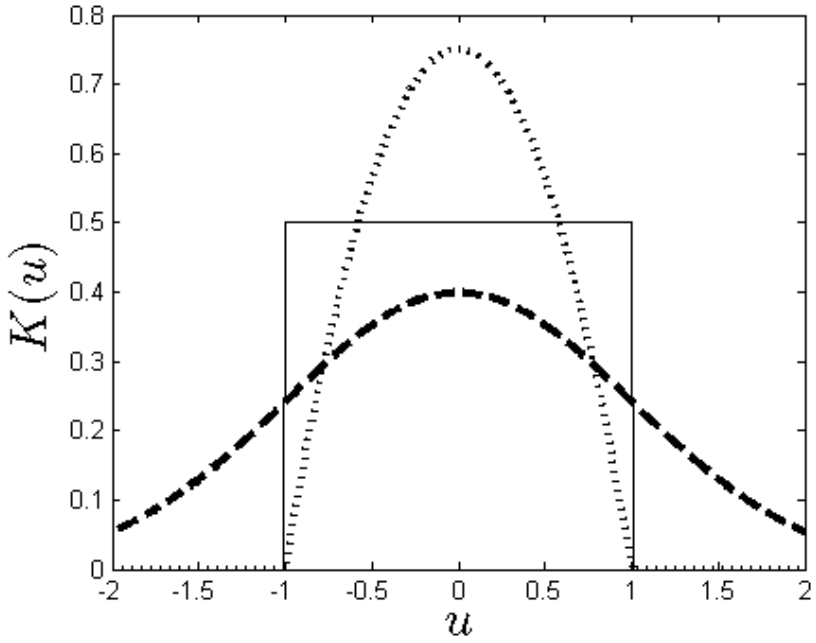

Figure 2.1: Plot of (–) uniform, $(\cdots)$ Epanechnikov and (- -) Gaussian kernel functions

2.1.1 Choice of bandwidth size

An important choice in the estimation of a function $g(x)$ is the size of the bandwidth, h, which is related to the complexity of the function. (In contrast, in a parametric model the complexity is controlled by increasing or decreasing the number of parameters in the model). The estimation of a function with many features (e.g. turning points) requires a smaller bandwidth, so that features are not lost or weakened by observations made at points at some distance from x having undue influence on $\hat{g}(x)$. However, even though a small bandwidth provides a less biased estimate of $g(x)$, in that $\mathrm{E}(\hat{g}(x))$ is very close to $g(x)$, this is attained at the cost of a high variance (Ramsay and Silverman, 2005, p. 78). Conversely, larger bandwidths include more data points in the prediction at x, which provides a prediction with low variability but potentially high bias.

Various methods have been developed to choose the optimal bandwidth for use in prediction. For example, Fan and Gijbels (1995) discussed data-driven bandwidth selection based on a residual squares criterion which is relatively simple to compute and can be 'plugged in' to $K(\cdot)$. In some cases there may be a need for a varying bandwidth (Simonoff, 1996, ch. 5). For example, suppose a function is relatively simple, for small x, and then, for larger x, has a sharp peak. Estimation of the function would benefit from a larger bandwidth initially to avoid oversmoothing, leading to a prediction which is too 'wiggly', followed by a smaller bandwidth to avoid undersmoothing, leading to a prediction which is 'too smooth', see Section 6.2.1.1 for further discussion.

2.1.2 Local polynomial estimators

The use of local polynomial estimators, local weighted regression (loess) or moving local regression was introduced by Pelto, Elkins and Boyd (1968) and Cleveland (1979). Müller (1996) provided a clear explanation of local fitting. A smooth function can often be approximated by a simpler function over a small region of the design space. The method of local weighted regression fits a pth degree polynomial to data locally using weighted least squares. Each observation, y_j $(j = 1, \ldots, n)$, is assigned a particular weight calculated using the kernel function K, where more weight is given to an observation at a design point closer to the prediction point x. Again, the locality of the smoothing is controlled by the bandwidth, h.

The local linear estimator with $p = 1$ is widely used, and given by

$$\hat{g}(x) = \frac{1}{nh} \frac{\sum_{j=1}^{n} \{\hat{s}_2(x;h) - \hat{s}_1(x;h)u_j\} K(\frac{u_j}{h}) y_j}{\hat{s}_2(x;h)\hat{s}_0(x;h) - \hat{s}_1(x;h)^2}, \tag{2.1}$$

where, for $r = 0, 1, 2$, $\hat{s}_r(x;h) = \frac{1}{nh} \sum_{k=1}^{n} u_k^r K(\frac{u_k}{h})$ and

$$S_j(x) = \frac{1}{nh} \frac{\{\hat{s}_2(x;h) - \hat{s}_1(x;h)u_j\} K(\frac{u_j}{h})}{\hat{s}_2(x;h)\hat{s}_0(x;h) - \hat{s}_1(x;h)^2}.$$

Note that if we set $K(\frac{u_j}{h}) = h$ for all u_j in (2.1) then we find $\hat{g}(x)$ reduces to

$$\hat{g}(x) = \frac{\sum_{j=1}^{n} u_j^2 \sum_{j=1}^{n} y_j - \sum_{j=1}^{n} u_j \sum_{j=1}^{n} u_j y_j}{n \sum_{j=1}^{n} u_j^2 - (\sum_{j=1}^{n} u_j)^2},$$

the ordinary least squares estimator of $\beta_0(x)$, the intercept parameter in linear regression.

A further special case is the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964), where $p = 0$,

$$\hat{g}(x) = \frac{\sum_{j=1}^{n} K(\frac{u_j}{h}) y_j}{\sum_{j=1}^{n} K(\frac{u_j}{h})}, \tag{2.2}$$

with

$$S_j(x) = \frac{K(\frac{u_j}{h})}{\sum_{j=1}^{n} K(\frac{u_j}{h})}, \tag{2.3}$$

In a foundation paper, Fan (1992) established, through both theory and simulation methods, advantages of the local linear estimator ($p = 1$) over the alternative Nadaraya-Watson (2.2) and the Gasser and Müller (Section 2.1.3) smoothers. The author proved, under certain conditions, that this estimator was the 'best' among all linear smoothers when using the optimal bandwidth chosen by minimising the mean expected squared error. Here 'best' means having the lowest maximum expected squared error over a class of true regression functions. Fan (1992) showed this best case was achieved by the local linear estimator with the Epanechnikov kernel function. He argued against the use of the Nadaraya-Watson estimator because it had infinite expected squared error for all kernels, as a result of it potentially having infinite bias.

In general, choosing an odd value for p, the degree of the local polynomial fitted, results in the order of the bias being the same for boundary and interior points. Therefore, $p = 1$ or $p = 3$ were suggested by Wand and Jones (1995, ch. 5). In the work presented in Chapters 3 and 4 we use $p = 1$ because the resulting estimator is much quicker to compute and has often adequate bias and boundary properties for small enough h.

For a truncated kernel such as the uniform or Epanechnikov, it is important that there are always sufficient points in the interval $[x - h, x + h]$ to fit a polynomial of the required degree. For example, we require at least two points within the bandwidth of our prediction point to fit a straight line, otherwise it is impossible to estimate the slope. Also the greater the number of design points lying within a distance of h of x, the lower the variance of the prediction.

2.1.3 The Gasser and Müller estimator

The Gasser and Müller estimator, see Gasser and Müller (1979, 1984), is given by

$$\hat{g}(x) = \sum_{j=1}^{n} \left[\frac{1}{h} \int_{\bar{x}_{j-1}}^{\bar{x}_j} K\left(\frac{v - x}{h} \right) \mathrm{d}v \right] y_j, \tag{2.4}$$

where $\bar{x}_j = (x_{j+1} + x_j)/2$ for $1 \leq j < n$, $\bar{x}_0 = x_1$ and $\bar{x}_n = x_n$ and

$$S_j(x) = \frac{1}{h} \int_{\bar{x}_{j-1}}^{\bar{x}_j} K\left(\frac{v-x}{h}\right) \mathrm{d}v. \tag{2.5}$$

The simplest weights are given by the uniform kernel.

Ramsay and Silverman (2005, p. 75) comment that the Gasser and Müller weights are fast to compute, deal better with unequally spaced data points and have good asymptotic properties in comparison to the Nadaraya-Watson estimator. Unlike the local linear estimator, the Gasser and Müller estimator has boundary bias problems in a manner similar to the Nadaraya-Watson estimator (Fan, 1992) which are not addressed in this thesis. The Gasser and Müller estimator is used in this thesis as we make predictions with unequally spaced points and, in some cases, it is possible to get analytic results using the uniform kernel.

2.2 Experimental design for local fitting

The literature provides three main approaches to experimental design for local fitting. Cheng, Hall and Titterington (1998) developed a sequential approach using the local linear estimator to find optimal 'design densities', from which the required number of design points for an experiment is drawn at random. At each step, both the optimal design density and asymptotic optimal bandwidths are calculated, by minimising the integrated mean squared error. This approach has the advantage of mathematical tractability: at each step, the optimal design density for the next step has a closed form solution. They obtained numerical results on approximate efficiencies to illustrate the gains of the optimal design densities over the uniform design density.

For experiments when sequential designs cannot be applied, Biedermann and Dette (2001) proposed a minimax approach to find optimal design densities using the Gasser and Müller estimator. For a specified class of 'true' functions, $g(x)$, and certain error distributions, they found design densities that minimised the maximum of the asymptotic integrated mean squared error in conjunction with using the optimal bandwidth. They numerically investigated the performance of these optimal design densities via asymptotic relative efficiency when either the form of $g(x)$ of the variance function (or both) were misspecified.

The disadvantage of the design density methods is that for small to moderate sized designs, there is large variability in the realised designs and hence in the achieved efficiencies. Hence we have not pursued these methods.

The work in this thesis builds upon optimal design strategies introduced by Müller (1992, 1996) who found continuous designs that enable 'best' prediction at q distinct points in the design region. Specifically, designs were found that minimised a weighted sum of the variances of the estimator $\hat{\beta}_0(x_i)$, for prediction at points x_i, $i = 1, ..., q$, see Section 3.3.3.1. This is a special case of the linear optimality criterion which selects support points to minimise the objective function

$$\Psi(\xi) = tr \sum_{i=1}^{q} A_i M_i(\xi_n)^{-1}, \tag{2.6}$$

where $A_i = a_i A$ with a_i a scalar, and A a $p \times p$ matrix with every element zero except for the $(1, 1)$ element which is 1. The $p \times p$ matrix, M_i, is the information matrix for the linear model at point x_i for $i = 1, ..., q$, given by

$$M_i = X_i^T W_i X_i, \tag{2.7}$$

where X_i is the design matrix and W_i is the matrix of kernel weights. Note that a different X_i is required for each of local quadratic and local linear regression; the value w_i changes according to the kernel function used. See Chapter 3 for more details.

This criterion was applied by Müller (1992) to a number of simple examples of finding optimal designs for predicting a response at nine equally spaced points in the interval $[-1, 1]$ when the design region consisted of the same nine points:

(a) Using the Nadaraya-Watson estimator and the uniform kernel with bandwidth $h = 0.1$. This is a very simple example. Since one design point is required to be within $h = 0.1$ of each of the prediction points, there is only one point to choose: the prediction point itself. Hence the optimal design is given by the set of nine prediction points.

(b) Using the local linear estimator with two nearest neighbour weight functions (see Cleveland (1979) and McLain (1971)). The weight functions were calibrated to ensure an equivalent degree of smoothing was enforced throughout, by fixing the equivalent degrees of freedom in the model.

Fedorov et al. (1999) also found designs which minimised a type of linear optimality criterion in which the objective function was a function of the 'mean cross product error' matrix, R, instead of a function of the information matrix, where R is proportional to $E\left[(\hat{g}(x_i) - g(x_i))(\hat{g}(x_k) - g(x_k))\right]$. The methods of Fedorov et al. (1999) differ from those of Müller (1992) because the error term is split into approximation error and random error

in order to analyse model misspecification through the bias of the fit. In order to find designs under this criterion, information about the the true model is required to calculate the bias. Both local linear and local quadratic true models were considered for three levels of local bias. Such designs are specific to the particular true model assumed.

Designs were found for making predictions at $q = 1, 11$ and 21 equally spaced points in the interval $[-1, 1]$ using the 101-point design region: $\{-1, -0.98, -0.96, \cdots, 0.98, 1\}$.

Designs were found numerically using the local linear and quadratic estimators and two types of weight function: a constant weight function over $[-1, 1]$ and the Gaussian kernel with standard deviation of 1/6. The authors did not explicitly define the bandwidth and enforced the locality of the fit through the choice of the standard deviation in the Gaussian kernel.

The authors found that an optimal design for any of $q = 1, 11$ and 21 prediction points when there was zero bias, constant weight function and for the local linear estimator had only two support points at -1 and 1 with $w_1 = w_2 = 0.5$. Optimal designs had more support points when the Gaussian weight function was used instead of the constant weight function. The optimal designs were not given explicitly, but plots presented indicated that they were roughly the equally weighted points:

- $\{-0.45, 0.45\}$ for predicting at a single point
- $\{-1.00, -0.55, -0.20, 0.20, 0.55, 1.00\}$ for prediction at 11 points
- $\{-1.00, -0.55, -0.20, 0.20, 0.55, 1.00\}$ for prediction at 21 points.

The support points for 11 and 21 points were almost equally weighted with slightly more weight at -0.55 and 0.55. In Chapter 3, we show how designs such as these can be obtained by a more general approach.

Throughout this thesis we concentrate on models where the variance outweighs the bias due to the assumed complexity of the model. Hence we follow the strategy of Müller (1992, 1996) where only 'stochastic disturbance' is defined. This approach is more appropriate than that of Fedorov et al. (1999) when we have little or no information about the function we wish to estimate.

Chapter 3

Optimal designs for local linear estimation

This chapter focuses on optimal design for 'best' prediction of a function g using the local linear estimator, a type of kernel smoother. We find designs $\xi_n = \{x_1, ..., x_n\}$, composed of ordered points, that maximise the average of the reciprocal prediction variances at q prediction points $x_1^*, ..., x_q^* \in \mathbb{R}$. A similar problem was considered by Müller (1992), Müller (1996) and Fedorov et al. (1999) for local linear smoothing. In Section 3.4 we obtain more general results for prediction across a continuous interval in $\mathbb{R}$.

In Section 3.5 we demonstrate our methodology on a Tribology experiment (Section 1.3), and find optimal designs for the local linear estimator to enable accurate prediction of the functional response from each treatment. We assess the performance for prediction of these optimal designs using the average mean squared error. We also investigate the robustness of the optimal designs to choice of bandwidth.

3.1 The local linear estimator

Recall, from Section 1.1.1, that a linear smoother, $\hat{g}(x)$, estimates the value of $g(x)$ through a linear combination of y_j as

$$\hat{g}(x) = \sum_{j=1}^{n} S_j(x) y_j,$$

where $S_j(x)$ is the smoothing weight for observation y_j for predicting at x.

The method of local weighted regression, introduced in Section 2.1.2, fits a pth degree polynomial to data locally using weighted least squares. Each observation, y_j $(j = 1, \ldots, n)$, is assigned a particular weight calculated using the kernel function K, where more weight is given to an observation at a design point x_j closer to the prediction point x^*.

Suppose that the $(p+1)$th derivative of $g(x)$ exists in a small neighbourhood about a point x^*. Then, from the Taylor series expansion of $g(x)$ about x^*

$$\begin{aligned} g(x) &\approx g(x^*) + g^{(1)}(x^*)(x - x^*) + \frac{g^{(2)}(x^*)}{2}(x - x^*)^2 + ... + \frac{g^{(p)}(x^*)}{p!}(x - x^*)^p \\ &= \beta_0(x^*) + \beta_1(x^*)(x - x^*) + ... + \beta_p(x^*)(x - x^*)^p, \end{aligned} \tag{3.1}$$

where $g^{(p)}(x)$ denotes the pth derivative. If we set $u = x - x^*$ we obtain

$$g(x) = \beta_0(x^*) + \beta_1(x^*)u +, \ldots, +\beta_p(x^*)u^p.$$

On, setting $x = x^*$, we see that the problem of estimating $g(x^*)$ is equivalent to that of estimating $\beta_0(x^*)$.

It follows directly from (3.1) that, regardless of the degree of the local polynomial, $\hat{g}(x^*) = \hat{\beta}_0(x^*)$ and $\hat{\boldsymbol{\beta}}(x^*) = (\hat{\beta}_0(x^*), ..., \hat{\beta}_p(x^*))^T$ minimises

$$\sum_{j=1}^{n} (y_j - \beta_0(x^*) - \beta_1(x^*)u_j - ... - \beta_p(x^*)u_j^p)^2 K(u_j),$$

where $u_j = x_j - x^*$ for $j = 1, \ldots, n$.

The form of the local polynomial estimator for $p = 0$ and the local linear estimator $(p = 1)$ used in this chapter are given by (2.2) and (2.1) respectively.

3.1.1 Local linear regression and weighted least squares

We now formulate the prediction variance for the local linear estimator, $(p = 1)$ using weighted least squares.

From Wand and Jones (1995, p. 114) the local linear estimator estimates the regression function at a specific point x^* by locally fitting a first degree polynomial to the data, i.e. the observations at x_j, $j = 1, \ldots, n$. Weighted least squares regression is used to correct for unequal error variance. For prediction at x^*, we assume the model

$$\boldsymbol{Y} = X\boldsymbol{\beta}(x^*) + \boldsymbol{\epsilon}(x^*), \tag{3.2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector, X is the model matrix, in terms of u_i,

$$X = \begin{bmatrix} 1 & u_1 \\ 1 & u_2 \\ \vdots & \vdots \\ 1 & u_n \end{bmatrix},$$

with $u_j = x_j - x^*$, $\boldsymbol{\beta}(x^*)$ is a $(p+1) \times 1$ vector and $\boldsymbol{\epsilon}(x^*) \sim N(0, W^{-1}\sigma^2)$. Here the matrix of smoothing weights, W, is given by

$$W = \frac{1}{h} \begin{bmatrix} K\left(\frac{u_1}{h}\right) & \ldots & \ldots & 0 \\ \vdots & K\left(\frac{u_2}{h}\right) & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \ldots & \ldots & K\left(\frac{u_n}{h}\right) \end{bmatrix}.$$

Then the estimator $\hat{\boldsymbol{\beta}}(x^*) = (\hat{\beta}_0(x^*), \hat{\beta}_1(x^*))^T$ is given by

$$\hat{\boldsymbol{\beta}}(x^*) = (X^T W X)^{-1} X^T W \boldsymbol{Y}. \tag{3.3}$$

The estimator of the response at point x can be found, using $u = x - x^*$, as

$$\hat{y} = \hat{\boldsymbol{\beta}}_0(x^*) + \hat{\boldsymbol{\beta}}_1(x^*)u.$$

Hence, if $x = x^*$ then $u = 0$ and $\hat{y} = \hat{\boldsymbol{\beta}}_0(x^*)$.

From equations (3.2) and (3.3), the variances of the estimated parameters are

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}(x^*)) &= \sigma^2(X^TWX)^{-1}X^TWW^{-1}WX(X^TWX)^{-1} \\ &= \sigma^2(X^TWX)^{-1}, \end{aligned} \tag{3.4}$$

using $\epsilon \sim N(0, W^{-1}\sigma^2)$. The information matrix, i.e. the inverse of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}(x^*)$, is given by $M(\xi_n) = X^TWX$.

3.2 D_s-optimality for prediction at a single point

In this section, we find designs which maximise the reciprocal variance of $\hat{g}(x^*) = \hat{\beta}_0(x^*)$. This leads us to consider D_s-optimality (Atkinson et al., 2007), which finds designs that minimise the variance of a subset of model parameter estimators whilst regarding the remaining parameters as nuisance parameters.

For our problem, $\beta_0(x^*)$ is the parameter of interest and $\beta_1(x^*)$ is the nuisance parameter. The information matrix can be expressed, in the notation of Atkinson et al. (2007), as

$$\begin{aligned} M(\xi_n) &= X^TWX \\ &= \begin{bmatrix} M_{11}(\xi_n) & M_{12}(\xi_n) \\ M_{21}(\xi_n) & M_{22}(\xi_n) \end{bmatrix} \qquad (3.5) \\ &= \frac{1}{h}\begin{bmatrix} \sum_{j=1}^n K\left(\frac{u_j}{h}\right) & \sum_{j=1}^n u_j K\left(\frac{u_j}{h}\right) \\ \sum_{j=1}^n u_j K\left(\frac{u_j}{h}\right) & \sum_{j=1}^n u_j^2 K\left(\frac{u_j}{h}\right) \end{bmatrix} \\ &. \qquad (3.6) \end{aligned}$$

The D_s-optimality criterion for β_0 seeks a design to maximise the determinant

$$|M_{11}(\xi_n) - M_{12}(\xi_n)M_{22}^{-1}M_{12}^T(\xi_n)| = \frac{|M(\xi_n)|}{|M_{22}(\xi_n)|}.$$

For our problem,

$$|M| = |X^TWX| = \left[\frac{1}{h}\sum_{j=1}^n K\left(\frac{u_j}{h}\right)\right]\left[\frac{1}{h}\sum_{j=1}^n u_j^2 K\left(\frac{u_j}{h}\right)\right] - \left[\frac{1}{h}\sum_{j=1}^n u_j K\left(\frac{u_j}{h}\right)\right]^2,$$

and

$$|M_{22}| = \frac{1}{h}\sum_{j=1}^{n} u_j^2 K\left(\frac{u_j}{h}\right).$$

We can now formulate a specific D_s-criterion for our problem.

Criterion 3.1. *Design ξ_n^* is D_s-optimal for predicting at a single point using the local linear estimator if it maximises the objective function*

$$\Psi(\xi_n) = \frac{\sum_{j=1}^{n} K\left(\frac{u_j}{h}\right)\sum_{j=1}^{n} u_j^2 K\left(\frac{u_j}{h}\right) - [\sum_{j=1}^{n} u_j K\left(\frac{u_j}{h}\right)]^2}{h\sum_{j=1}^{n} u_j^2 K\left(\frac{u_j}{h}\right)}, \tag{3.7}$$

where $u_j = x_j - x^*$.

We give (in Theorem 3.1) sufficient conditions for a design to be optimal under Criterion 3.1. To do this, we first prove two results.

Lemma 3.1. *For any kernel function and design that is symmetric about x^* and has at least two design points, the objective function* (3.7) *is*

$$\Psi(\xi_n) = \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{u_j}{h}\right). \tag{3.8}$$

Proof. It is possible to write (3.7) as

$$\Psi(\xi_n) = \frac{1}{h}\left(\sum_{j=1}^{n} K\left(\frac{u_j}{h}\right) - \frac{[\sum_{j=1}^{n} u_j K(\frac{u_j}{h})]^2}{\sum_{j=1}^{n} u_j^2 K(\frac{u_j}{h})}\right).$$

For a symmetric design, since the kernel function is symmetric about 0, $\sum_{i=1}^{n} u_j K\left(\frac{u_j}{h}\right) = 0$ and hence

$$\frac{[\sum_{j=1}^{n} u_j K(\frac{u_j}{h})]^2}{\sum_{j=1}^{n} u_j^2 K(\frac{u_j}{h})} = 0. \tag{3.9}$$

Therefore

$$\Psi(\xi_n) = \frac{1}{h}\sum_{i=1}^{n} K\left(\frac{u_j}{h}\right).$$

□

Note that a symmetric design is not necessarily optimal. There may exist non-symmetric designs for which

$$\frac{[\sum_{j=1}^{n} u_j K(\frac{u_j}{h})]^2}{\sum_{j=1}^{n} u_j^2 K(\frac{u_j}{h})} > 0$$

and

$$\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{u_j}{h}\right) < \Psi(\xi_n).$$

For prediction at a single point, it is possible to find an upper bound for the objective function analytically. This upper bound then provides a sufficient condition for a design to be optimal under Criterion 3.1.

Lemma 3.2. *An upper bound, U, for objective function* (3.7) *is given by*

$$U = \frac{n}{h}K(0) \geq \max_{\xi_n}(\Psi(\xi_n)).$$

Proof. By definition, $K(0)$ is the maximum value of K. Hence, we can re-write the kernel $K\left(\frac{u_j}{h}\right)$ as $[K(0) - f(u_j)]$ with the function f satisfying $f(x) \geq 0$ for all x and $f(0) = 0$. Hence from Lemma 3.1, equation (3.7) can be written as

$$\Psi(\xi_n) = \frac{1}{h}\left(\sum_{j=1}^{n}[K(0) - f(u_j)] - \frac{[\sum_{j=1}^{n} u_j K(\frac{u_j}{h})]^2}{\sum_{j=1}^{n} u_j^2 K(\frac{u_j}{h})}\right).$$

Now,

$$\frac{[\sum_{j=1}^{n} u_j K\left(\frac{u_j}{h}\right)]^2}{\sum_{j=1}^{n} u_j^2 K\left(\frac{u_j}{h}\right)} \geq 0, \tag{3.10}$$

because $K\left(\frac{u_j}{h}\right) \geq 0$ for all u_j, $j = 1, ..., n$. Therefore

$$\Psi(\xi) = \frac{1}{h}\left(\sum_{j=1}^{n}[K(0) - f(u_j)] - \frac{[\sum_{j=1}^{n} u_j K\left(\frac{u_j}{h}\right)]^2}{\sum_{j=1}^{n} u_j^2 K\left(\frac{u_j}{h}\right)}\right) \tag{3.11}$$

$$\leq \frac{1}{h}\sum_{j=1}^{n}[K(0) - f(u_j)]$$

$$= \frac{1}{h}\sum_{j=1}^{n} K(0) - \frac{1}{h}\sum_{j=1}^{n} f(u_j)$$

$$\leq \frac{n}{h}K(0), \tag{3.12}$$

where n is the number of design points.

□

Theorem 3.1. *A sufficient condition for design ξ_n^* to be D_s-optimal under Criterion 3.1 for prediction at a single point is $\Psi(\xi_n^*) = nK(0)/h$.*

Proof. Proof follows directly from Lemma 3.2

□

3.2.1 Results

We now find D_s-optimal designs using Criterion 3.1 for the uniform and Gaussian kernels. Here and throughout this chapter, where designs are found numerically, we use the 'fminsearch' routine in MATLAB to minimise $-\Psi(\xi_n)$. This routine uses the Nelder-Mead simplex algorithm (Nelder and Mead, 1965) as described by Lagarias et al. (1998), a direct search method which does not use numerical or analytic gradients in the optimisation.

3.2.1.1 Optimal designs using the uniform kernel

Recall from Section 2.1 that the uniform kernel is

$$K(v) = \begin{cases} 0.5 & \text{if } |v| \leq h, \\ 0 & \text{otherwise.} \end{cases}$$

When this kernel is used, the objective function (3.7) becomes

$$\Psi(\xi_n) = \frac{1}{2h}\left[\sum_{j=1}^{n} 1_A(u_j) - \frac{[\sum_{j=1}^{n} u_j 1_A(u_j)]^2}{\sum_{j=1}^{n} u_j^2 1_A(u_j)}\right], \tag{3.13}$$

where

$$1_A(u_j) = \begin{cases} 1 & \text{if } u_j \in A, \\ 0 & \text{otherwise,} \end{cases}$$

with

$$A = \{u_j; |u_j| \leq h\}. \tag{3.14}$$

We can see that (3.13) is maximised when $|u_j| \leq h$, so that $1_A(u_j) = 1$, for all $j = 1, \ldots, n$, by minimising

$$\frac{[\sum_{j=1}^{n} u_j 1_A(u_j)]^2}{\sum_{j=1}^{n} u_j^2 1_A(u_j)}. \tag{3.15}$$

As (3.15) is greater than or equal to zero, it is minimised by any design ξ_n satisfying

(i)

$$\begin{aligned} \sum_{j=1}^{n} u_j = \sum_{j=1}^{n} (x_j - x^*) &= 0 \\ \Leftrightarrow \sum_{j=1}^{n} x_j &= nx^* \\ \Leftrightarrow \bar{x} &= x^*, \end{aligned} \tag{3.16}$$

where $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$, and

(ii)

$$\sum_{j=1}^{n} u_j^2 > 0$$

$$\Leftrightarrow \sum_{j=1}^{n} (x_j - x^*)^2 > 0 \tag{3.17}$$

for all $j = 1, \ldots, n$.

Substituting (3.16) into (3.17) gives

$$\sum_{j=1}^{n} (x_j - \bar{x})^2 > 0. \tag{3.18}$$

These arguments show that the average of the design points must equal x^* and all design points cannot be equal. This leads to the following corollary

Corollary 3.1. *For the uniform kernel and prediction at x^* a design $\xi_n = \{x_1, \ldots, x_n\}$ with $n \geq 2$ that satisfies*

(i) $|x_j - x^*| \leq h$

(ii) $\bar{x} = x^*$

(iii) $\sum_{j=1}^{n} (x_j - \bar{x})^2 > 0$

has $\Psi(\xi_n) = \frac{n}{2h}$, is D_s-optimal under Criterion 3.1.

Proof. Conditions (i)-(iii) imply that

$$\Psi(\xi_n^*) = \frac{nK(0)}{h} = \frac{n}{2h}. \tag{3.19}$$

and the result follows from Theorem 3.1. □

To confirm these results, optimal designs were also found numerically. We minimised $-\Psi(\xi_n)$ for a selection of values for h, x^* and n and found in every case that $\Psi(\xi_n)$ had a maximum value of $n/2h$. All the optimal designs found, as expected, satisfied $\bar{x} = x^*$ and (3.18)with all n design points within h of x^*.

Note that for truncated kernels, such as the uniform, it is necessary to have at least two design points within h of x^*, otherwise it is impossible to make a prediction. If there are more than two design points, an optimal design under Criterion 3.1 has all its points within h of x^*; then all observations contribute to the prediction. This results in greater accuracy than if only observations at two design points are used.

The case of predicting at $x^* = 0$ with the uniform kernel and $h = 1$, ensuring a constant weight function for $x_j \in [-1, 1]$, is very similar to a problem considered by Fedorov et al. (1999). These authors used the mean cross product error, assuming there was no bias term (see Section 2.2 for details) with the constant weight function over $[-1, 1]$. They found the approximate optimal design with equally weighted support points, at $x \pm 1$, for the discrete design region $\{-1, -0.98, -0.96, \cdots, 0.98, 1\}$. From Corollary 3.1, this design is also D_s-optimal under Criterion 3.1 for n even. This is one of many designs that satisfy Corollary 3.1.

3.2.1.2 Optimal designs for the Gaussian kernel

The Gaussian kernel is not truncated and hence $K(v) > 0$ for all v. The Gaussian kernel is defined as

$$K(v) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{v^2}{2}\right\} \quad -\infty < v < \infty,$$

When using the Gaussian kernel the upper bound on the objective function, $nK(0)/h$ from Lemma 3.2, cannot be attained. To see this, recall from (3.11), that

$$\Psi(\xi_n) = \frac{1}{h}\left(\sum_{j=1}^{n}[K(0) - f(u_j)] - \frac{[\sum_{j=1}^{n} u_j K(\frac{u_j}{h})]^2}{\sum_{j=1}^{n} u_j^2 K(\frac{u_j}{h})}\right).$$

A design for the Gaussian kernel cannot achieve the bound U of Lemma 3.2. This is because $f(u_j) = 0$ implies $u_j = 0$, as $K(v) < K(0)$ for $v \neq 0$, and hence $[\sum_{j=1}^{n} u_j K(\frac{u_j}{h})]^2 = 0$ and $\sum_{j=1}^{n} u_j^2 K(\frac{u_j}{h}) = 0$, leading to the ratio

$$\frac{[\sum_{j=1}^{n} u_j K(\frac{u_j}{h})]^2}{\sum_{j=1}^{n} u_j^2 K(\frac{u_j}{h})}$$

being indeterminate.

Numerical search found optimal designs that were symmetric about x^*, and therefore by

Lemma 3.1 satisfy (3.8). However, these designs have the property that $f(u_j) \neq 0$. The design points are positive such that u_j is very close to zero, but not exactly zero, for all j. In other words, all the design points are of the form, $x_j = x^* \pm \delta$, where $\delta > 0$ is small.

Our results differ from those of Fedorov et al. (1999) for predicting at zero with the Gaussian kernel function, again using the mean cross product error. These authors used the normal kernel with standard deviation 1/6 with the bias term set to be zero. D_s-optimal designs from Criterion 3.1 have points placed very close to zero, the point of prediction. The optimal designs from Fedorov et al. (1999) have two support points $\{-0.45, 0.45\}$. Hence, unlike with the uniform kernel, the D_s-optimal design differs from the design from Fedorov et al. (1999) due to their different objective functions.

3.3 D_s-optimality for prediction at a finite number of points

In this section, building on Section 3.2, we define a design criterion for when we wish to predict the response at more than one point. We then find, for the uniform and Gaussian kernels, the minimum number of design points needed. We also find designs for different values of n and h for each kernel.

An optimal design for predicting at several distinct points $x_1^*, \ldots, x_q^*$ simultaneously is not necessarily the union of q designs, each optimal for predicting at one of the points. This is due to the fact that some design points may influence the prediction at several different prediction points.

To find designs to predict at several points, we use a compound criterion. Atkinson et al. (2007, p. 266), explained that minimising the sum of the variances could give a large variance too much prominence. A better alternative is to maximise the product of the reciprocal variances of $\hat{\beta}_0(x_i^*)$ for $i = 1, \ldots, q$. This is equivalent to maximising the compound objective function

$$\Psi(\xi_n) = \sum_{i=1}^{q} \log \frac{|M_i(\xi_n)|}{|M_{22_i}(\xi_n)|}, \tag{3.20}$$

where M_i is the information matrix for the local linear estimator for predicting at one point $x = x_i$ (see Section 3.2), and M_{22_i} is the partition of information matrix, M_i, for nuisance parameter $\beta_1(x_i^*)$ corresponding to M_{22} in (3.5). Therefore, analogous to Criterion 3.1, we

find designs using the following criterion.

Criterion 3.2. *Design ξ_n^* is compound D_s-optimal for prediction at $x_1^*, \ldots, x_q^*$ using the local linear estimator if it maximises*

$$\Psi(\xi_n) = \sum_{i=1}^{q} \log \left(\frac{\sum_{j=1}^n K(\frac{u_{ij}}{h}) \sum_{j=1}^n u_{ij}^2 K(\frac{u_{ij}}{h}) - [\sum_{j=1}^n u_{ij} K(\frac{u_{ij}}{h})]^2}{h \sum_{j=1}^n u_{ij}^2 K(\frac{u_{ij}}{h})} \right), \tag{3.21}$$

where $u_{ij} = x_j - x_i^*$.

We consider designs for the uniform and Gaussian kernels separately.

3.3.1 Uniform kernel: minimum number of design points and the corresponding optimal design

In this section we consider the minimum number, n_{min}, of design points required to make a prediction at q ordered points, $x_1^* < \ldots < x_q^*$, and identify a D_s-optimal design. Recall that when $q = 1$, we only required two design points to make a prediction using the local linear estimator. When predicting at $q > 1$ points with the uniform kernel, we require two design points within h of each prediction point x_i^* $(i = 1, ..., q)$.

When using the uniform kernel, only design points x_j satisfying $|x_j - x^*| \leq h$ influence the prediction at x^*. Define $[x_i^* - h, x_i^* + h]$ as the *prediction interval* for x_i^*. For two neighbouring prediction points, x_i^* and x_{i+1}^*, there are two different scenarios: the intervals $[x_i^* - h, x_i^* + h]$ and $\left[x_{i+1}^* - h, x_{i+1}^* + h\right]$ either intersect, or they do not. These two scenarios can be defined, respectively, as 'overlapping' interval and 'disjoint' interval prediction.

In the latter case, the design points that influence the prediction at x_i^* do not influence the prediction at x_{i+1}^*. Hence, four design points are required, a pair on each interval, and the pairs can be chosen independently. When the prediction intervals intersect, the position of the design points used for predicting at x_i^* will affect the position of design points used for predicting at x_{i+1}^*. Then we require fewer than four design points. This is discussed in Section 3.3.1.2.

3.3.1.1 Disjoint prediction intervals

It is clear that in order to predict at q points which have non-intersecting prediction intervals, we require at least two design points for each of the q points. Therefore, the minimum number of design points required is $n_{min} = 2q$. A D_s-optimal design, under Criterion 3.2, with this minimum number of points and the uniform kernel is one which has two distinct design points for each x_i^*, and satisfies (3.16) and (3.17) for each interval.

If there are $n > 2q$ design points available, we must first ensure that two distinct design points are allocated in the interval corresponding to each x_i^*. We then must decide how to allocate the remaining points. An optimal design must satisfy (3.16) and (3.17) for each prediction point. Then if $n_i \geq 2$ design points are within h of x_i^*, from (3.19), the objective function (3.20) takes the value

$$\Psi(\xi_n) = \sum_{i=1}^{q} \log\left(\frac{n_i}{2h}\right). \tag{3.22}$$

Designs found numerically with $n > 2q$ design points demonstrated that if q divides n exactly, then $n_1 = ... = n_q = n/q$. If q does not divide n exactly but divides $n - r$ where $r < q$, then $(n-r)/q$ points were used for $q-r$ intervals and $(n-r+q)/q$ points used for r intervals. It does not matter which of the r intervals have an extra point, each possible allocation gives an optimal design. Once again the n/q, $(n-r)/q$ or $(n-r+q)/q$ design points in each interval must satisfy (3.16) and (3.17).

Example 1

Suppose that $q = 2$ with $x_1^* = 0$, $x_2^* = 1$, $h = 0.2$ and $n = 6$. The optimal design found numerically for $n_1 = 3$ and $n_2 = 3$ had an objective function value of 4.03; setting $n_1 = 2$ and $n_2 = 4$ gave a corresponding value of 3.91. Therefore we can see that equally dividing points is preferred under Criterion 3.2. For both choices of n_1, n_2, any design satisfying (3.16) and (3.17) on both $[-0.2, 0.2]$ and $[0.8, 1.2]$ is optimal for predicting at x_1^* and x_2^*.

Example 2

Consider $q = 2$ with $x_1^* = 0$, $x_2^* = 1$, $h = 0.2$ and $n = 7$. Numerical results show that setting $n_1 = 3$ and $n_2 = 4$ or $n_1 = 4$ and $n_2 = 3$ gives an objective function value of 4.32, whereas setting $n_1 = 2$ and $n_2 = 5$ gives the value of 4.14. Again, in all three cases optimal designs have points $x_1, \ldots, x_7$ satisfying (3.16) and (3.17) in each interval for predicting at x_1^* and x_2^*.

3.3.1.2 Overlapping prediction intervals

If the prediction intervals around any two prediction points intersect, then $n_{min} < 2q$. There are two different cases which must be considered:

1. The intersection of two prediction intervals is itself an interval.
2. The intersection of two prediction intervals is a single point.

In this section, designs denoted 'optimal' have not been proven to be optimal. However we give firm intuitive reasoning on why these designs may be optimal. Numerical results from Section 3.3.3 confirm our conjecture.

Two simple examples for each of cases 1 and 2 above are as follows.

Example 3

Take the simplest case where we have to predict at two points, x_1^* and x_2^*, and

$$[x_1^* - h, x_1^* + h] \cap [x_2^* - h, x_2^* + h] = [x_2^* - h, x_1^* + h] \neq \emptyset.$$

Then it is possible to predict at both x_1^* and x_2^* with any two design points x_1 and x_2 such that $x_1, x_2 \in [x_2^* - h, x_1^* + h]$.

Example 4

The degenerate case of two overlapping intervals is when $x_2^* - h = x_1^* + h$. Here we require three design points, x_1, x_2 and x_3, with the unique three point D_s-optimal design having $x_1 = x_1^* - h$, $x_2 = x_1^* + h$ and $x_3 = x_2^* + h$. This design is the only design to satisfy (3.16) and (3.17) on each interval.

We now investigate four cases for predicting at q points, carefully defining when more than two prediction intervals are dependent.

Case (i): prediction intervals with intersection of the form $[a, b]$, $a < b$

We know that there must be at least two design points within h of each prediction point. Therefore two points are required within each disjoint intersection of prediction intervals. We only need to consider the disjoint intersections, as any overlapping intersections can be treated as a single interval.

Example 5

Suppose that predictions are required at five points $x_1^*, x_2^*, x_3^*, x_4^*, x_5^*$ where the prediction intervals for x_1^*, x_2^*, x_3^* have intersection $[x_3^* - h, x_1^* + h]$, and the prediction intervals for x_4^*, x_5^* have intersection $[x_5^* - h, x_4^* + h]$.

The intersections of these two prediction intervals are disjoint when $[x_3^* - h, x_1^* + h] \cap [x_5^* - h, x_4^* + h] = \emptyset$, that is, when $x_5^* - h > x_1^* + h$. If these prediction intervals were not disjoint then their intersection would result in a single prediction interval $[x_3^* - h, x_1^* + h] \cap [x_5^* - h, x_4^* + h] = [x_5^* - h, x_1^* + h]$.

The union of all disjoint intersections is given by

$$\bigcup_{k=0}^{l-1} \left[x^*_{a_{k+1}} - h, x^*_{a_k+1} + h \right],$$

where l is the total number of disjoint intersections. We define $a_0 = 0$, $a_l = q$ and a_{k+1} as the largest integer such that

$$\bigcap_{i=a_k+1}^{a_{k+1}} [x_i^* - h, x_i^* + h] \neq \emptyset.$$

Note that $a_{k+1} > a_k$. Our definition of the union of disjoint intersections ensures that no two disjoint intersections may involve the same prediction interval. This avoids a design having more design points than necessary. The minimum number of design points required is $2l$, since we require two design points per disjoint intersection.

Example 6

Here we consider an example when $q = 5$ prediction points: $x_1^* = 0$, $x_2^* = 0.2$, $x_3^* = 0.35$, $x_4^* = 0.8$ and $x_5^* = 0.9$, and $h = 0.2$. Then there are two disjoint intersections of prediction intervals $[0.15, 0.2]$ and $[0.7, 1]$. Prediction intervals for x_1^*, x_2^* and x_3^* form one intersection; the other intersection is formed by the intersection of prediction intervals for x_4^* and x_5^*. We have that $a_0 = 0$, $a_1 = 3$ and $a_2 = 5$. Since there are two disjoint intersections, we require at least four points for predicting at $x_1^*, ..., x_5^*$: two points in $[0.15, 0.2]$ and two points in $[0.7, 1]$.

The locations of optimal design points with these intersections need to be determined numerically. For all examples of case (i), including Example 6, we found that design points

were placed at each end of the intersection intervals. That is, the compound D_s-optimal design under Criterion 3.2 is given by

$$\left\{x^*_{a_{k+1}} - h, x^*_{a_k+1} + h : k = 0, ..., l-1\right\}.$$

In Example 6, the optimal design had points at $x_1 = 0.15, x_2 = 0.2, x_3 = 0.7$ and $x_4 = 1$.

Case (ii): intersecting prediction intervals in the form of $q-1$ distinct points

Now we consider the situation where q prediction points are equidistant and the distance between consecutive points is exactly h. Then the set of $q-1$ points which occur on the boundaries of the intervals can be defined as

$$\{x^*_1 + (2k-1)h : k = 1, ..., q-1\} = \bigcup_{k=0}^{q-1} \Big([x^*_1 + (2k-1)h, x^*_1 + (2k+1)h] \cap [x^*_1 + (2k+1)h, x^*_1 + (2k+3)h] \Big). \tag{3.23}$$

The minimum number of design points required to predict at $x^*_1, \ldots, x^*_q$ is $q+1$ and these design points are given by the set (3.23) augmented by the two points at the end of the first and last prediction interval. i.e. the set of design points is

$$\{x^*_1 + (2k-1)h : k = 0, ..., q\}.$$

This design is uniquely optimal since it is the only design satisfying (3.16) and (3.17) for each prediction interval.

Example 7

Consider $q = 3$ prediction points: $x^*_1 = 0.2$, $x^*_2 = 0.6$ and $x^*_3 = 1$ and $h = 0.2$. The prediction intervals intersect at 0.4 and 0.8. Therefore the optimal design is given by $x_1 = 0$, $x_2 = 0.4$, $x_3 = 0.8$ and $x_4 = 1.2$.

Case (iii): intersecting prediction intervals in the form of distinct points and one interval $[a, b]$, $a < b$

Here we combine cases (i) and (ii). The first $q-1$ consecutive prediction intervals intersect at $q-2$ distinct points and the intersection between the $(q-1)$th and qth prediction intervals is an interval. We would expect to require q design points to predict the response at the $q-2$ points from the first $q-1$ intersections. However, the intersection of the $(q-1)$th and qth prediction intervals must also be taken into account.

The union of the intersections of these q intervals is

$$\bigcup_{k=1}^{q-1} \bigcap_{i=k}^{k+1} [x_i^* - h, x_i^* + h] = \bigcup_{k=1}^{q-1} \left[x_k^* + h, x_{k+1}^* - h\right],$$

which is equal to

$$\{x_1^* + (2k-1)h : k = 1, ..., q-2\} \cup \left[x_q^* - h, x_{q-1}^* + h\right],$$

where

$$\left[x_q^* - h, x_{q-1}^* + h\right] = \left[x_{q-1}^* - h, x_{q-1}^* + h\right] \cap \left[x_q^* - h, x_q^* + h\right].$$

The minimum number of design points required is $n_{min} = q + 1$. If $x_{q-1}^* + h \neq x_q^*$ the first q design points are given by

$$\{x_1^* + (2k-1)h : k = 0, ..., q-1\}.$$

Then the two design points in the optimal design for predicting at x_q^* will be $x_{q-1}^* + h$ (from the prediction of x_{q-1}^*) and $x_q^* + [x_q^* - (x_{q-1}^* + h)] = 2x_q^* - x_{q-1} - h$. These two design points are equidistant from x_q^*. This design is optimal since (3.16) and (3.17) are satisfied for all q prediction points.

However, if $x_{q-1}^* + h = x_q^*$ the first $q-1$ design points are given by

$$\{x_1^* + (2k-1)h : k = 0, ..., q-2\}.$$

Using the above argument for $x_{q-1}^* + h \neq x_q^*$, the two remaining design points would be placed at $x_{q-1}^* + h$, included in (3.3.1.2), for predicting at x_q^*. However, using this design it would be impossible to make a prediction at x_q^* as we only have one distinct design point in $[x_q^* - h, x^*q + h]$. The best we can do is put two points very close, but not equal, to $x_{q-1}^* + h$, so that (3.16) and (3.17) are almost satisfied for all q (see Section 3.2.1.2 for a similar argument). Therefore the two remaining design points are placed at $x_{q-1}^* + h - \delta$ and $x_{q-1}^* + h + \delta$ for small $\delta > 0$.

Example 8

For making predictions at $q = 3$ points: $x_1^* = 0.2$, $x_2^* = 0.6$ and $x_3^* = 0.85$, and $h = 0.2$, the prediction intervals for x_1^* and x_2^* intersect at 0.4 and the prediction intervals for x_2^* and x_3^* intersect on $[0.65, 0.8]$. Therefore the D_s-optimal design is given by $x_1 = 0$, $x_2 = 0.4$, $x_3 = 0.8$ and $x_4 = 0.9$.

Example 9

If we change Example 8 so that $x_1^* = 0.2$, $x_2^* = 0.6$ and $x_3^* = 0.8$, again with $h = 0.2$, the problem is slightly different. Then $x_2^* + h = x_3^* = 2x_3^* - x_2^* - h$. Naively, this suggests that two design point are placed at 0.8, giving design points $x_1 = 0$, $x_2 = 0.4$, $x_3 = 0.8 - \delta$ and $x_4 = 0.8 + \delta$ for small $\delta > 0$.

Case (iv): Internal prediction intervals intersect at points, and the intersection of the prediction intervals for the first two and last two points are both intervals

In this case $q - 2$ consecutive prediction intervals for prediction at $x_2^*, \ldots, x_{q-1}^*$ intersect at $q - 3$ distinct points. There are interval intersections between the prediction intervals for x_1^* and x_2^*, and between the $(q-1)$th and qth prediction intervals.

The union of intersection of these q intervals is given by (3.24) and can be simplified as

$$\{x_1^* + (2k-1)h : k = 2, ..., q-2\} \cup [x_2^* - h, x_1^* + h] \cup \left[x_q^* - h, x_{q-1}^* + h\right],$$

where

$$[x_2^* - h, x_1^* + h] = [x_1^* - h, x_1^* + h] \cap [x_2^* - h, x_2^* + h] \tag{3.24}$$

and

$$\left[x_q^* - h, x_{q-1}^* + h\right] = \left[x_{q-1}^* - h, x_{q-1}^* + h\right] \cap \left[x_q^* - h, x_q^* + h\right].$$

The minimum number of design points required is $n_{min} = q + 1$. If $x_{q-1}^* + h \neq x_q^*$ and $x_1^* + h \neq x_2^*$ design points $x_2, \ldots, x_{q-1}$ are given by

$$\{x_1^* + (2k-1)h : k = 1, ..., q-1\}. \tag{3.25}$$

The two design points for predicting at x_q^* will be $x_{q-1}^* + h$, included in (3.25), and $2x_q^* - x_{q-1} - h$ as in case (iii). The two design points for predicting at x_1^* are $x_1^* - (x_2^* - h - x_1^*) = 2x_1^* - x_2^* + h$ and $x_2^* - h$, included in (3.25). This design is optimal since (3.16) and (3.17) are satisfied for all q prediction points.

However, as in case (iii), if $x_{q-1}^* + h = x_q^*$ and $x_1^* + h = x_2^*$ then the design points $x_3, \ldots, x_{q-2}$ are given by

$$\{x_1^* + (2k-1)h : k = 1, ..., q-1\}.$$

Under the above argument, for $x_{q-1}^* + h \neq x_q^*$ and $x_1^* + h \neq x_2^*$ there would be two design points placed at each of $x_2^* - h$ and $x_{q-1}^* + h$. This prevents us from making a prediction at x_1^* and x_q^* since we only have one distinct design point for each prediction. The best we can do is put two points very close to $x_2^* - h$ and $x_{q-1}^* + h$, but not equal, so (3.16) and (3.17) are almost satisfied for all q. Therefore the remaining design points are placed at $x_2^* - h - \delta$, $x_2^* - h - \delta$, $x_{q-1}^* + h - \delta$ and $x_{q-1}^* + h + \delta$ for small $\delta > 0$. This design does not satisfy the sufficient conditions in Corollary 3.1 for predicting at either x_1^* and x_q^*.

Example 10

Consider making predictions at $q = 4$ points: $x_1^* = 0.2$, $x_2^* = 0.5$, $x_3^* = 0.9$ and $x_4^* = 1.2$, with $h = 0.2$. The optimal design is given by $x_1 = 0.1$, $x_2 = 0.3$, $x_3 = 0.7$, $x_4 = 1.1$ and $x_5 = 1.3$.

Example 11

To make predictions at $x_1^* = 0.3$, $x_2^* = 0.5$, $x_3^* = 0.9$ and $x_4^* = 1.2$, with $h = 0.2$, the problem is slightly different. Now $x_1^* + h = x_2^* = 2x_2^* - x_1^* - h$ suggesting that two design points are put at 0.3. Therefore we have design points $x_1 = 0.3 - \delta$, $x_2 = 0.3 + \delta$, $x_3 = 0.7$, $x_4 = 1.1$ and $x_5 = 1.3$, for small $\delta > 0$.

In general, it is possible to predict at q points which have overlapping prediction intervals in a number of different ways. A combination of results from cases (i)-(iv) can be applied to find the minimum number of design points and to give a compound D_s-optimal design; either established analytically or through numerical search.

3.3.2 Minimum number of design points required to predict at q points using the Gaussian kernel

The minimum number of design points required for predicting at q points with the Gaussian kernel is two. This is because $K(u_{ij}) > 0$ for all $\infty < u_{ij} < \infty$ and hence each design point influences the prediction at all q points. However, these predictions may not be very accurate when a large number of predictions is required and $n = 2$.

3.3.3 Prediction at q points for different n and h

In this section, we find designs for predicting at q points when more than the minimum number of design points is available, i.e. $n > n_{min}$. Optimal designs are found under Criterion 3.2, again using the Nelder-Mead simplex algorithm. We present two sets of optimal designs for each of the uniform and Gaussian kernels: designs for predicting at $x_1^* = 0, x_2^* = 0.5$ and designs for predicting at $x_1^* = 0, x_2^* = 0.6, x_3^* = 0.8, x_4^* = 1.1$. Tables 3.1 and 3.2 give designs for prediction with the uniform kernel, and Tables 3.3 and 3.4 for prediction with the Gaussian kernel. All tables present optimal designs for $h = 0.2, 0.5, 0.75, 1$ and $n = 2, 3, 4, 5, 6, 7, 8$, where possible.

In this section, the designs found numerically have not been proven to be optimal. However, they will be at least highly efficient under Criterion 3.2, and for brevity we denote them as optimal.

3.3.3.1 Optimal designs for the uniform kernel.

In this section, designs are found for $n > n_{min}$ for predicting at q points where a design may have disjoint or overlapping prediction intervals, depending on the value of the bandwidth, h.

When $h = 0.2$, the prediction intervals for predicting at $\{0, 0.5\}$ are disjoint. Therefore a design is optimal if it satisfies (3.16) and (3.17) for each interval and $|u_{ij}| \leq h$ for every

prediction point, x_i^*. Table 3.1 gives one optimal design for each n when $h = 0.2$; however there are infinitely many optimal designs. Since the prediction intervals for 0 and 0.5 are disjoint, four points are required to make a prediction. For this reason, there is no optimal design given for $n = 2, 3$.

For $h = 0.5, 0.75, 1$, the prediction intervals for predicting at $x_1^* = 0, x_2^* = 0.5$ intersect. For several values of n and $h = 0.5, 0.75$, two optimal designs were found for predicting at $x_1^* = 0, x_2^* = 0.5$, both of which are given in Table 3.1. These designs, ξ_n^* and $-\xi_n^*$, have the property that $-\xi_n$ is the design composed of the reflections of the points from ξ_n^* in the line $x = 0.25$.

Unlike when predicting at $x_1^* = 0, x_2^* = 0.5$, there was only one optimal design found for predicting at $x_1^* = 0, x_2^* = 0.6, x_3^* = 0.8, x_4^* = 1.1$; see Table 3.2. This is perhaps because the prediction points are not equally spaced, so we cannot find two optimal designs which are reflections of each other. For prediction at $x_1^* = 0, x_2^* = 0.6, x_3^* = 0.8, x_4^* = 1.1$ with $h = 0.2$, at least six design points are required as there are three disjoint intersections of overlapping prediction intervals. Hence no optimal designs were found for $n = 4, 5$.

It was difficult to find optimal designs for $h = 0.5$ when predicting at $x_1^* = 0, x_2^* = 0.6, x_3^* = 0.8$ and $x_4^* = 1.1$ for large values of n, see Tables 3.2. Of the values for h investigated, $h = 0.5$ is the smallest value of h which gives overlapping prediction intervals when predicting at the above four points. However, the pattern of overlap was complicated for this value of h. This led to a complicated objective function and the optimisation routine did not always converge.

In general, for both sets of prediction points, as h is reduced an optimal design has more support points. This is explained by the fact that only points very close to x_i^* influence the prediction. Hence we require more distinct design points to account for the increased complexity when making a prediction with smaller h.

Comparison to the optimal designs of Müller (1992)

We now compare our results to those of Müller (1992) for the Nadaraya-Watson estimator. Recall that this estimator is the local polynomial estimator with $p = 0$ and the prediction is given by (2.2). Using this estimator, only one point is required to lie within h of x^* for a prediction to be possible.

Müller (1992) found approximate designs for making predictions at nine points, equally spaced on the interval $[-1, 1]$, where the design region consisted of the same nine points. The Nadaraya-Watson estimator and the uniform kernel were used with $h = 0.1$. We found designs using the same set up as Müller (1992), except that the design region was

n	$h = 0.2$	$h = 0.5$
2	-	0.00 0.50
3	-	-0.05 0.05 0.95
4	-0.02 0.02 0.43 0.57	-0.50 0.21 0.29 1.00
5	-0.20 0.01 0.19 0.31 0.69	-0.50 0.00 0.25 0.50 1.00
6	-0.20 0.01 0.19 0.35 0.56 0.59	-0.50 0.00(2) 0.50(2) 1.00
7	-0.20 0.01 0.19 0.40	(a) -0.50 0.00(2) 0.13 0.50 1.00(2)
	0.45 0.46 0.70	(b) -0.50(2) 0.00 0.37 0.50(2) 1.00
8	-0.20 -0.07 0.10 0.17	(a) -0.50 0.00(3) 0.30 0.50 1.00(2)
	0.36 0.42 0.53 0.70	(b) -0.50(2) 0.00 0.20 0.50(3) 1.00
12	-0.20 -0.06 0.05(2) 0.07 0.10	-0.50 -0.49 0.00(3) 0.06
	0.31(2) 0.50 0.51 0.67 0.70	0.39 0.44 0.45 1.00(3)
15	-0.20 -0.07 0.03(2) 0.04	-0.50(3) 0.02 0.05
	0.05 0.12 0.30 0.41(2)	0.10(2) 0.22 0.40(2)
	0.43 0.52 0.61 0.62 0.70	0.43 0.49 1.00(3)

	$h = 0.75$	$h = 1$
2	-0.25 0.75	-0.50 1.00
3	(a) -0.25 0.27 1.25	-0.50 0.25 1
	(b) -0.75 0.23 0.75	
4	(a) -0.75 -0.21 0.75(2)	-0.50(2) 1.00(2)
	(b) -0.25(2) 0.71 1.25	
5	(a) -0.75 -0.25 0.53 0.75(2)	-0.50(2) 0.25 1.00(2)
	(b)-0.25(2) -0.03 0.75 1.25	
6	(a) -0.75 -0.25(2) 0.75(3)	-0.50(3) 1.00(3)
	(b) -0.25(3) 0.75(2) 1.25	
7	(a) -0.75 -0.25(2) 0.75(4)	(a) -0.50(3) 0.18 1.00(2) 1.50
	(b) -0.25(4) 0.75(2) 1.25	(b) -1.00 -0.50(2) 0.33 1.00(3)
8	-0.75 -0.25(3) 0.75(3) 1.25	-0.50(4) 1.00(4)

Table 3.1: D_s-optimal designs under Criterion 3.2 for predicting at $x_1^* = 0$ and $x_2^* = 0.5$, using the uniform kernel and differing numbers of design points and values for h. Number of repetitions of a design point in parentheses; (a) and (b) indicate designs that are reflections.

n	$h = 0.2$	$h = 0.5$
4	-	$\pm$0.30 0.90 1.29
5	-	-0.50 0.30 0.39 1.10(2)
6	$\pm$0.19 0.58 0.62 0.98 1.22	-0.50 0.30(2) 0.89 1.10 1.30
7	$\pm$0.98 0.40 0.61 0.79 1.00 1.20	-0.50 0.30(3) 0.93 1.10 1.30
8	$\pm$0.11 0.60(3) 0.10(2) 1.30	-0.50 0.30(3) 0.89 1.10(2) 1.30
	$h = 0.75$	$h = 1$
2	0.35 0.75	0.10 1.10
3	-0.15 0.35 1.35	-0.20 0.27 1.60
4	-0.15 0.05 0.75 1.35	-0.20 0.20 1.41 1.60
5	-0.15 0.05 0.35 1.35(2)	-0.20 0.10 0.12 1.60(2)
6	-0.75 0.05 0.35 0.75 1.35(2)	-0.20 0.10(2) 1.60(3)
7	-0.75 0.05(2) 0.75(2) 1.35(2)	-0.20 0.10(3) 1.60(3)
8	-0.75 0.05(2) 0.35 0.63 1.35(3)	-0.20(2) 0.10(2) 0.25 1.60(3)

Table 3.2: D_s-optimal designs under Criterion 3.2 for predicting at $x_1^* = 0, x_2^* = 0.6, x_3^* = 0.8$ and $x_4^* = 1.1$, using the uniform kernel and differing numbers of design points and values for h. Number of repetitions of a design point in parentheses.

the whole interval $[-1, 1]$. We found exact designs instead of approximate designs.

We compared our designs with those of Müller (1992) for $h = 0.1$. The optimal design from Müller (1992) consisted of support points at each of the nine prediction points, all equally weighted. This is due to the fact that one point is required to be within $h = 0.1$ of each of the prediction points and there is only one such point in the discrete design space: the prediction point itself.

The design found by Müller (1992) was optimal under our criterion and set-up. In contrast, we set the prediction region to be the interval $[-1, 1]$. Therefore an optimal design was any set of points with one point lying within 0.1 of each prediction point, again equally weighted.

3.3.3.2 Optimal designs for the Gaussian kernel.

The Gaussian kernel is different from the uniform kernel in that it is not truncated. The minimum number of design points required to make a prediction is two, as these are sufficient to fit a straight line on the whole design region (see Section 3.3.1.2).

We consider predicting at the sets of points $x_1^* = 0, x_2^* = 0.5$ and $x_1^* = 0, x_2^* = 0.6, x_3^* = 0.8, x_4^* = 1.1$ for $n = 2, 3, 4, 5, 6, 7, 8, 12, 15$ and $h = 0.2, 0.5, 0.75, 1$.

Table 3.3 shows the results for predicting at $\{0, 0.5\}$. We see that, for each value of

n, the distinct design points become further apart as h increases. This is in line with our expectation, as increasing the bandwidth gives a design point further away from a prediction point more influence on the prediction at that point. We also notice that for even values of n each optimal design has only two distinct design points, with $n/2$ points at $0 - a$ and $n/2$ points at $0.5 + a$. Note that for a given h, the value of a is constant. When n is odd, there are two optimal designs, both with only two distinct design points: $(n-1)/2$ points at $0 - c$ and $(n+1)/2$ points at $0.5 + b$ or $(n+1)/2$ points at $0 - b$ and $(n-1)/2$ points at $0.5 + c$ ($b \leq a \leq c$).

For example, for $h = 0.5$ we see $a = 0.15$. If $n = 5$, then $c = 0.09$ and $b = 0.23$. However, if $n = 15$ then $c = 0.13$ and $b = 0.18$. As n increases, b increases towards a and c decreases towards a. There is an exception when $n = 3$ for $h = 0.5, 0.75, 1$. In these cases, a design point is put at 0.25, half-way between the prediction points and then two points are placed equidistant from 0 and 0.5.

We see similar patterns when predicting at the four points $x_1^* = 0, x_2^* = 0.6, x_3^* = 0.8, x_4^* = 1.1$, see Table 3.4. Once again the distinct design points become more spread out as h increases. It is also noticeable that for smaller h we have more support points than for larger h. This is explained by the fact that the prediction is more local for small h. Only points very close to x_i^* have a large amount of influence. Hence we require more support points to account for the increase in complexity driven by h.

Comparison of optimal designs with those of Fedorov et al. (1999)

We compare the designs found by the two approaches on an example for predicting at eleven equally spaced points in the interval $[-1, 1]$. Fedorov et al. (1999) set the design region to be the discrete set:

$$\{-1, -0.98, -0.96, \cdots, 0.98, 1\},$$

whereas again, our design region was the interval $[-1, 1]$. We compared the design from Criterion 3.2 for $h = 0.25$, with the corresponding design of Fedorov et al. (1999) when the standard deviation is $1/6$.

For $n = 15$, our design has points

$$\{\pm 0.98(2), \pm 0.53(3), \pm 0.16(2), 0.00\},$$

with 7 distinct or support points, where (2) indicates 2 repetitions of the design point. The optimal design in Fedorov et al. (1999) had support points $\{-1, -0.55, -0.2, 0.2, 0.55, 1\}$,

with roughly 50% more weight at -0.55 and 0.55 than at the other support points. The designs from the two methods are at least qualitatively similar.

Note, that it was computationally easier to find optimal designs using the Gaussian kernel than for the uniform kernel as fewer iterations of the optimisation routine were required.

n	$h = 0.2$	$h = 0.5$
2	-0.02 0.52	-0.15 0.65
3	(a) -0.01(2) 0.53	-0.26 0.25 0.76
	(b) -0.03(2) 0.50	
4	-0.02(2) 0.52(2)	-0.15(2) 0.65(2)
5	(a) -0.01(3) 0.52(2)	(a) -0.09(3) 0.73(2)
	(b) -0.02(2) 0.51(3)	(b) -0.23(2) 0.59(3)
6	-0.02(3) 0.52(3)	-0.15(3) 0.65(3)
7	(a) -0.01(4) 0.52(3)	(a) -0.10(4) 0.71(3)
	(b) -0.02(3) 0.51(4)	(b) -0.21(3) 0.60(4)
8	-0.02(4) 0.52(4)	-0.15(4) 0.65(4)
12	-0.02(6) 0.52(6)	-0.15(6) 0.65(6)
15	-0.01(8) 0.52(7)	-0.13(8) 0.68(7)
	-0.02(7) 0.51(8)	-0.18(7) 0.63(8)
	$h = 0.75$	$h = 1$
2	-0.25 0.75	-0.33 0.83
3	-0.37 0.25 0.87	-0.47 0.25 0.97
4	-0.25(2) 0.75(2)	-0.33(2) 0.83(2)
5	(a) -0.15(3) 0.87(2)	(a) -0.20(3) 0.98(2)
	(b) -0.37(2) 0.65(3)	(b) -0.48(2) 0.70(3)
6	-0.25(3) 0.75(3)	-0.33(3) 0.83(3)
7	(a) -0.17(4) 0.83(3)	(a) -0.24(4) 0.93(3)
	(b) -0.33(3) 0.67(4)	(b) -0.43(3) 0.74(4)
8	-0.25(4) 0.75(4)	-0.33(4) 0.83(4)
12	-0.25(6) 0.75(6)	-0.33(6) 0.83(6)
15	(a) -0.21(8) 0.78(7)	(a) -0.28(8) 0.88(7)
	(b) -0.28(7) 0.71(8)	(b) -0.37(7) 0.79(8)

Table 3.3: D_s-optimal designs under Criterion 3.2 for predicting at $x_1^* = 0$ and $x_2^* = 0.5$, using the Gaussian kernel and differing numbers of design points and values for h. Number of repetitions of a design point in parentheses; (a) and (b) indicate designs that are reflections.

n	$h = 0.2$	$h = 0.5$
2	0.00 1.10	0.00 1.09
3	0.00 0.61 1.10	-0.09 0.72 1.23
4	-0.01 0.54 0.80 1.15	-0.05 0.07 1.09(2)
5	-0.01 0.60(2) 1.10(2)	-0.03(2) 0.78 1.16(2)
6	0.00(2) 0.61(2) 1.10(2)	-0.09(2) 0.39 1.11(3)
7	-0.01(2) 0.57(2) 0.79 1.13(2)	-0.01(3) 0.86 1.13(3)
8	-0.01(2) 0.60(3) 1.10(3)	-0.05(3) 0.49 1.11(4)
12	-0.01(3) 0.58(4) 0.76 1.11(4)	-0.02(5) 0.63 1.11(6)
15	0.00(5) 0.61(5) 1.10(5)	0.00(7) 1.08(8)
	$h = 0.75$	$h = 1$
2	-0.05 1.21	-0.11 1.32
3	-0.22 0.82 1.27	-0.38 0.91 1.29
4	-0.05(2) 1.21(2)	-0.11(2) 1.32(2)
5	-0.15(2) 1.12(3)	-0.27(2) 1.18(3)
6	-0.18(2) 0.31 1.20(3)	-0.11(3) 1.32(3)
7	-0.12(3) 1.14(4)	-0.22(3) 1.22(4)
8	-0.17(3) 0.69 1.18(4)	-0.31(3) 1.15(5)
12	-0.13(5) 1.13(7)	-0.24(5) 1.21(7)
15	-0.15(6) 1.12(9)	-0.27(6) 1.18(9)

Table 3.4: D_s-optimal designs under Criterion 3.2 for predicting at $x_1^* = 0, x_2^* = 0.6, x_3^* = 0.8$ and $x_4^* = 1.1$, using the Gaussian kernel and differing numbers of design points and values for h. Number of repetitions of a design point in parentheses.

3.4 Prediction across an interval

We now investigate designs for predicting across an entire continuous interval $[-1, 1]$ rather than at a discrete set of points. We define a new objective function which is the continuous version of (3.21), and will be used in a compound D_s-optimality criterion to find designs.

$$\begin{aligned}\Psi(\xi_n) &= \frac{1}{h}\int_{-1}^{1} \log\left(\sum_{j=1}^{n} K\left(\frac{x_j - x^*}{h}\right) - \frac{[\sum_{j=1}^{n}\left(\frac{x_j-x^*}{h}\right)K\left(\frac{x_j-x^*}{h}\right)]^2}{\sum_{j=1}^{n}\left(\frac{x_j-x^*}{h}\right)^2 K\left(\frac{x_j-x^*}{h}\right)}\right)\mathrm{d}x^* \\ &= \int_{-1}^{1}\log\left[L(x^*)\right]\mathrm{d}x^*, \end{aligned} \tag{3.26}$$

where

$$L(x^*) = \frac{1}{h}\left[\sum_{j=1}^{n} K\left(\frac{x_j - x^*}{h}\right) - \frac{\left[\sum_{j=1}^{n}\left(\frac{x_j-x^*}{h}\right)K\left(\frac{x_j-x^*}{h}\right)\right]^2}{\sum_{j=1}^{n}\left(\frac{x_j-x^*}{h}\right)^2 K\left(\frac{x_j-x^*}{h}\right)}\right].$$

Note that, although we do not restrict the design region to $[-1, 1]$, we will see that most points in the resulting designs are in, or close to, the interval $[-1, 1]$.

We again find designs for prediction using both the uniform and Gaussian kernels. The integral in (3.26) is analytically intractable for each of these kernels. Therefore we implement a numerical quadrature scheme to approximate this integral for each kernel. We have used Legendre-Gauss quadrature to calculate the optimal weights and abscissae to approximate (3.26), with the abscissae given by the roots of the Legendre polynomials. Details of Gauss quadrature methods in general can be found in Golub and Welsch (1969). The approximation to (3.26) involves a weighted sum of the objective function at p_a abscissa values over the integration region and is given by

$$\Psi(\xi_n) \approx \sum_{i=1}^{p_a} \kappa_i \log\left[L(x_i^*)\right], \tag{3.27}$$

where x_i^* are chosen as solutions to the Legendre polynomials and κ_i are Legendre-Gauss weights.

Criterion 3.3. *An optimal design ξ_n^* for prediction across the interval $[-1,1]$ for the local linear estimator maximises*

$$\Psi(\xi_n) \approx \sum_{i=1}^{p_a} \kappa_i \log\left[L(x_i^*)\right].$$

Initially $p_a = 500$ was chosen to give a very accurate approximation to (3.26). However, for some larger values of n we found that it was impossible to run the optimisation algorithm for enough iterations to converge to an optimal design. We therefore decided to choose p_a large enough to produce an accurate approximation but also small enough to enable the optimisation to be performed in a reasonable time.

In order to choose an appropriate value of p_a, we generated 500 random designs by random selections of n points from $[-1,1]$. We then evaluated (3.27) for various values of p_a for each design and compared the results (see Figure 3.1 for $p_a = 25, 500$). A high correlation between the values of the objective functions for designs under (3.27) for different p_a suggests we can use the smaller value of p_a for design selection.

For some small values of n and h, for example, Figures 3.1(a) and 3.1(b), we see that $p_a = 25$ does not produce sufficiently accurate results as there is not a strong correlation between the objective function values for low $-\Psi(\xi_n)$ when calculated using $p_a = 25$ and $p_a = 500$. In these cases, we would not get the same optimal design. Hence, optimal designs for these values of h and n were calculated using $p_a = 500$. When $h = 0.1$, it seems $p_a = 25$ is sufficient for $n \geq 7$, see Figure 3.2.

In general, for other values of h, we use an approximation with 25 abscissa points

$$\Psi(\xi_n) \approx \sum_{i=1}^{25} \kappa_i \log\left[L(x_i^*)\right]. \tag{3.28}$$

3.4.1 Optimal designs for predicting on $[-1,1]$ using the uniform kernel

Table 3.5 gives optimal designs for Criterion 3.3 for predicting on the interval $[-1,1]$ using the uniform kernel. It is important to note that for each value of h it was only possible to find optimal designs for certain values of n. For example, when $h = 0.2$ we require at least eleven points to predict over the whole interval of length 2. These eleven points are equally spaced and ensure that there are at least two design points within h of any point in the

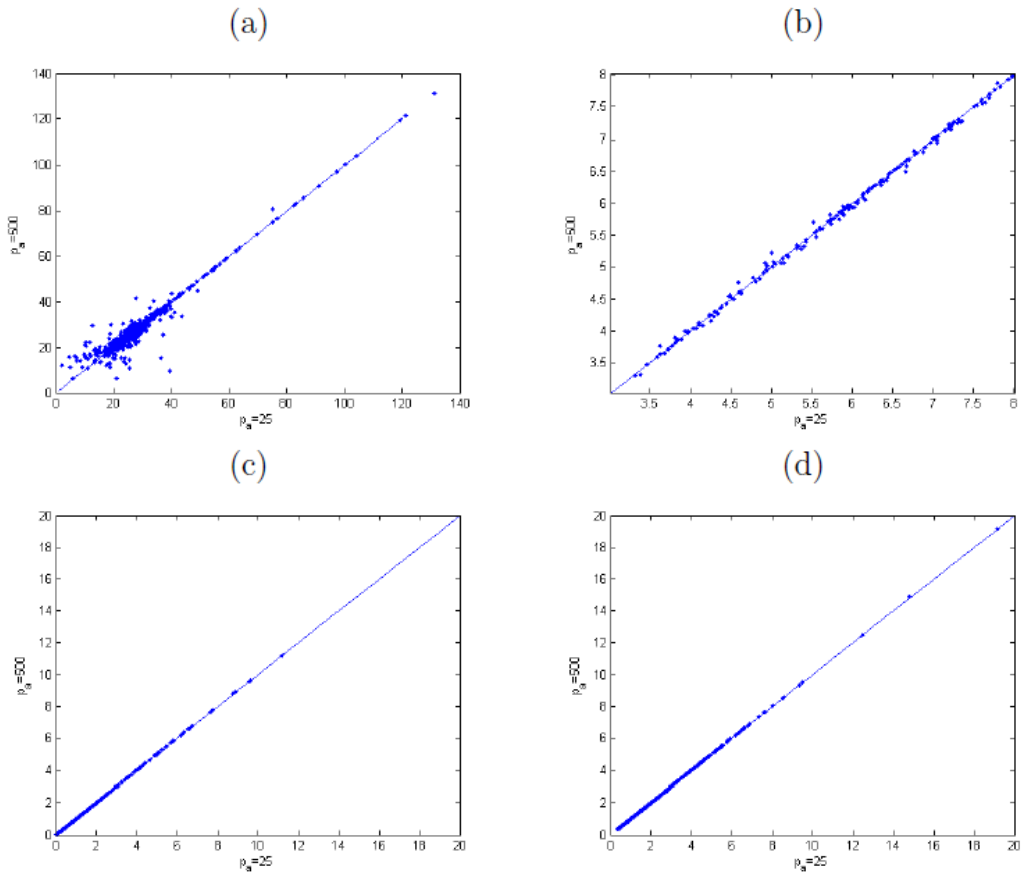

Figure 3.1: Comparison of objective function values $(-\Psi(\xi_n))$ for 500 random designs with $p_a = 25$ and $p_a = 500$. (a) $n = 3$ and $h = 0.1$, (b) $n = 3$ and $h = 0.2$, (c) $n = 3$ and $h = 0.5$ and (d) $n = 3$ and $h = 0.75$.

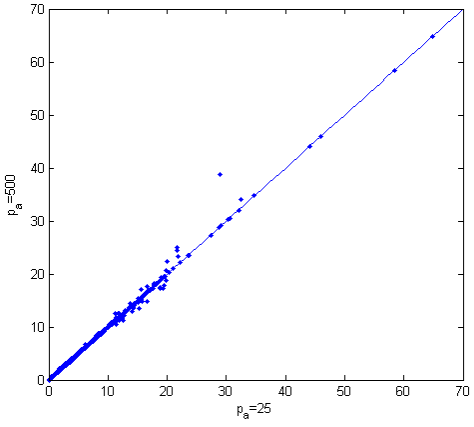

Figure 3.2: Comparison of objective function values $(-\Psi(\xi_n))$ for 500 random designs with $p_a = 25$ and $p_a = 500$ for $n = 7$, $h = 0.1$.

interval $[-1, 1]$. Otherwise, there are not two design points available for prediction over every part of the interval. Therefore $n = 12$ is the first optimal design given in Table 3.5. In general the designs are symmetric, or close to symmetric, with points spread across the interval. For smaller n and larger h, the numerical optimisation was more straightforward and faster.

For some values of n and h the designs were not quite symmetric. This may be because the optimum has not been identified exactly. However, this is not necessarily the case. In Table 3.5, for example, when $n = 7$ and $h = 0.5$ the optimal design found is $\xi_n^* = \{-1.01, -0.57, -0.25, 0.11, 0.30, 0.64, 1.01\}$ with $\Psi(\xi_n^*) = 3.26$. The value of the objective function was calculated for a number of symmetric designs to see if there was an obvious improvement to be made. However, the best symmetric design found was

$$\xi_n = \{-1.01, -0.60, 0.28, 0.00, 0.28, 0.60, 1.01\}$$

with $\Psi(\xi_n) = 3.23$.

3.4.2 Optimal designs for prediction on an interval with the Gaussian kernel

Table 3.6 gives compound Optimal designs under Criterion 3.3 for predicting over the interval $[-1, 1]$ using the Gaussian kernel. Firstly, we notice that the optimal designs are symmetric about zero. This is not unexpected, as both the interval we are predicting on and the kernel function are symmetric about zero. In a similar manner to predicting at $q = 4$ points in Section 3.3.3, fewer support points are required as h increases for fixed n. Secondly, it is noticeable that as h increases, more design points are placed near the ends of the interval. These points are more influential in predicting at points closer to the centre of the interval for larger h.

A uniform kernel design from Section 3.4.1 can be quantitavely compared to a design using the Gaussian kernel by calculating the efficiency of the 'uniform kernel' design for prediction with the Gaussian kernel. Two examples are considered; (i) $h = 0.5$ and $n = 5$, and (ii) $h = 0.5$ and $n = 15$. The efficiency is calculated as

$$\text{Eff} = \exp\left\{\Psi_G(\xi^u) - \Psi_G(\xi^G)\right\},$$

where $\Psi_G(\xi^u)$ and $\Psi_G(\xi^G)$ are the values of the objective function, calculated with the Gaussian kernel, using (i) ξ^u, the optimal design under Criterion 3.3 with the uniform

n	$h = 0.2$
12	-1.13 -0.85 -0.64 ±0.43 -0.27 -0.04 0.07
	0.28 0.65 0.86 1.12
15	-1.11 -0.93 -0.75 -0.54 -0.39 -0.28 -0.15
	0.00 0.19 0.33 0.40 0.56 0.76 0.95 1.12
	$h = 0.5$
5	±1.00 ±0.50 0.00
6	±0.99 ±0.55 ±0.20
7	±1.01 -0.57 -0.25 0.11 0.30 0.64
8	±1.04 ±0.71 ±0.40 ±0.16
12	-1.12 -0.85 -0.62 -0.44 -0.26 -0.05
	0.06 0.28 0.46 0.65 0.87 1.13
15	-1.12 -0.96 -0.78 -0.57 -0.41 -0.32 -0.19
	-0.03 0.16 0.28 0.38 0.51 0.76 0.93 1.11
	$h = 0.75$
4	-0.94(2) -0.25(2)
5	±1.09 ±0.53 0.00
6	±1.10 0.64 ± 0.13 0.63
7	±1.10 -0.82 -0.23 -0.01 0.22 0.81
8	±1.17 ±0.83 -0.46 -0.10 0.09 0.47
12	±1.36 -1.20 -0.82 -0.49 -0.34 -0.07
	0.02 0.31 0.51 0.93 1.08
15	-1.40 ±1.20 -1.03 -0.74 -0.47 -0.27 -0.05
	0.01 0.08 0.24 0.51 0.71 1.05 1.39
	$h = 1$
3	±1.00 0.00
4	±1.1221 ±0.32
5	±1.16 ±0.52 0.01
6	-1.19 ±-0.65 ±-0.29 1.18
7	±-1.30 -0.84 ±0.36 0.01 0.85
8	±1.29 -0.89 ±0.48 -0.20 0.23 0.91
12	-1.36 -1.18 -0.80 -0.49 -0.36 -0.12
	0.07 0.33 0.52 0.93 1.09 1.34
15	-1.31 -1.19 -1.05 -0.75 -0.52 -0.38 -0.28
	0.11 0.15 0.37 0.50 0.64 1.02 1.16 1.35

Table 3.5: D_s-optimal designs under Criterion 3.3 for predicting over the interval $[-1, 1]$ using a uniform kernel and differing numbers of design points and values for h. Number of repetitions of a design point in parentheses.

n	$h = 0.2$	$h = 0.5$
2	±0.16	±0.65
3	±0.72 0.00	±0.88 0.00
4	±0.88 ±0.31	±0.96 ±0.31
5	±0.92 ±0.46 0.00	±1.00 ±0.53 0.00
6	±0.93 ±0.54 ±0.18	±0.88(2) 0.00(2)
7	±0.95 ±0.59 ±0.30 0.00	±0.92(2) ±0.27 0.00
8	± 0.96 ± 0.64 ± 0.39 ± 0.12	±0.95(2) ±0.50 0.00(2)
12	±0.98 ±0.85 ±0.52(2) ±0.17(2)	±0.88(4) 0.00(4)
15	±0.95(2) ±0.61(2) ±0.43 ±0.20(2) 0.00	±0.88(5) 0.00(5)

	$h = 0.75$	$h = 1$
2	±0.77	±0.87
3	±0.98 0.00	±1.09 0.00
4	±0.86 ±0.68	±0.87(2)
5	±0.88(2) 0.00	±0.99 0.00
6	±0.85(2) ±0.60	±0.87(3)
7	±0.85(3) 0.00	±0.95(3) 0.00
8	±0.85(3) ±0.52	±0.87(4)
12	±0.84(5) ±0.33	±0.87(6)
15	±0.81(7) 0.00	-0.83(8) 0.91(7)

Table 3.6: D_s-optimal designs under Criterion 3.3 for predicting over the interval $[-1, 1]$ using a Gaussian kernel and differing numbers of design points and values for h. Number of repetitions of a design point in parentheses.

kernel and (ii) ξ^G, the optimal design under Criterion 3.3 using the Gaussian kernel.

In case (i), the uniform optimal design is $\xi^u = \{-1.00, -0.50, 0.00, 0.50, 1\}$, the Gaussian optimal design is $\xi^G = \{-1.00, -0.53, 0.00, 0.53, 1\}$ and the efficiency is 0.998. In case (ii), the uniform optimal design is

$$\begin{aligned}\xi_u^* = [&- 1.12, -0.96, -0.78, -0.57, -0.41, -0.32, -0.19, \\ &- 0.03, 0.16, 0.28, 0.38, 0.51, 0.76, 0.93, 1.11],\end{aligned}$$

the Gaussian optimal design is $\xi_G^* = \{\pm 0.88(5), 0.00(5)\}$ and the efficiency is 0.932, where (5) indicates 5 repetitions of the design point. In these two examples the uniform kernel designs perform well.

3.5 Application to the tribology experiment

The designs found in Section 3.4 can be applied to experiments such as the tribology example introduced in Section 1.3. Recall that the tribology experiment consisted of 16 runs, each of which resulted from the application of a different treatment. The response of interest is the combined wear of the disc and pin, see Figure 1.1. The aim was to predict this wear over the time interval $[501, 2400]$, where a time point will be denoted by x. We can predict the wear over this interval using the whole dataset, i.e. 1900 observations. However, in this section, we demonstrate the application of the design methods in this chapter by choosing a small subset of the observations with which to predict the wear. We chose the Gaussian kernel for prediction, as it was possible to find designs with large n for this kernel. We assess the performance of the designs from Section 3.4, for Criterion 3.3 for $n = 15, 20, 25, 30$ design points. To do this, a comparison is made of the smooth fit produced by the optimal design with the smooth fit produced using the whole dataset. We also compare the optimal designs to uniform designs composed of equally spaced points over the interval $[-1, 1]$. Comparisons are made in terms of mean squared error.

Optimal designs were found for bandwidths $h = 0.2, 0.5$ and $n = 15, 20, 25$, and for $h = 0.1, 0.3$ and $n = 15, 20, 25, 30$ (as the optimisation was faster for $h = 0.1$). The bandwidths and designs were transformed from the interval $[-1, 1]$ to $[501, 2400]$. For instance, $h = 0.1, 0.2, 0.3$ and $h = 0.5$ on the transformed interval $[501, 2400]$ correspond to smoothing parameters of 95, 190, 285 and 475. For each run, predictions were made using designs with all combinations of n and h. Immediately we saw that $h = 0.5$ was too large for all datasets since many features of the data were oversmoothed. Therefore we chose to investigate designs for $h = 0.1, 0.2$ and 0.3 where the selection of h was done 'by eye'. Designs for each of these bandwidths can be found in Table 3.7. Note that these designs were calculated specifically for this application and cannot be found in Table 3.6.

The results are illustrated using two runs, run 2 and run 19, which exhibit very different features. Figures 3.3 and 3.4 show the smoothed fits for these runs using bandwidths 0.2 and 0.1 respectively, for both the whole dataset and data from the corresponding optimal designs. These bandwidths were chosen to allow enough locality to describe features of the data. The difference in form of each run can be attributed to the different levels of factor settings for each run (see Chapter 5). Figure 3.3 shows that the smooth fit calculated from the whole dataset and from the data corresponding to the design points for run 2 are very similar for $n = 20$ and $n = 25$ (plots (b) and (c)). However, plot (a) shows that 15 design points were not sufficient.

For run 19, Figure 3.4 shows that the smooth fit calculated from the whole data and that

n	$h = 0.1$
15	±0.98 ±0.81 ±0.67 ±0.53 ±0.40 ±0.27 ±0.13 0.00
20	±0.99 ±0.87 ±0.75 ±0.66 ±0.55 ±0.45 ±0.35 ±0.25 ±0.15 ±0.05
25	±0.99 ±0.94 ±0.77(2) ±0.63 -0.57 -0.49 -0.39
	-0.34 -0.22 -0.19 -0.06 -0.02 0.10 0.15 0.25 0.32 0.41 0.48 0.58
30	0.99 ±0.98 -0.81 ±0.80 ±0.71 ±0.60(2) ±0.49 -0.44 -0.36
	-0.31 -0.21 -0.19 -0.07 -0.06 0.05 0.08 0.17 0.23 0.29 0.38 0.43 0.80 0.98
	$h = 0.2$
15	±0.95(2) ±0.61(2) -0.43 ±0.20(2) 0.00 0.42
20	±0.95 ±0.94 -0.94 ±0.57(3) ±0.38 ±0.18 ±0.16 ±0.13 0.93
25	±0.96(3) ±0.79 ±0.55 -0.55 ±0.54 -0.53 ±0.23
	±0.22 -0.19 -0.16 0.00 0.15 0.20 0.54 0.54
	$h = 0.3$
15	±0.93(3) ±0.41(3) ±0.14 0.00
20	±0.93(4) ±0.41(4) -0.16 ±0.01 0.17
25	±0.93(5) -0.42 ±0.41(3) -0.41 -0.18 -0.02 ±0.01 0.20 0.40 0.43
30	±0.93(6) ±0.42(2) ±0.41(2) ±0.40 -0.39 -0.24 ±0.00 0.01 0.03 0.19 0.42

Table 3.7: Further D_s-optimal designs under Criterion 3.3 for predicting over the interval $[-1, 1]$ using a Gaussian kernel and differing numbers of design points and values for h. Number of repetitions of a design point in parentheses.

from the design points are quite different, even for large values of n. Although small h can cause the fit to be more 'wiggly', thus undersmoothing the data, the value of $h = 0.1$ was chosen as no other value of h produced a fit which could predict the steep incline around $x = 1300$. However, for each of $n = 15, 20, 25, 30$ the data were undersmoothed elsewhere on the interval. This is due to the fact that, with this bandwidth, only a small number of design points have a significant influence on the prediction at each point.

Another issue is the variability, in terms of the signal to noise ratio of the data, see Figure 3.5. This leads to the performance of the designs being very unstable. Placing a design point at observation x_{t+1} rather than at x_t has the potential to make a big difference in the prediction over the interval. Figure 3.6 shows the autocorrelation in the errors from a smooth fit with $h = 0.1$ for run 19. For n given observations $y_1, \ldots, y_n$, the lag k autocorrelation is given by

$$r_k = \frac{(n-1)\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{(n-k)\sum_{t=1}^{n}(y_t - \bar{y})^2},$$

where

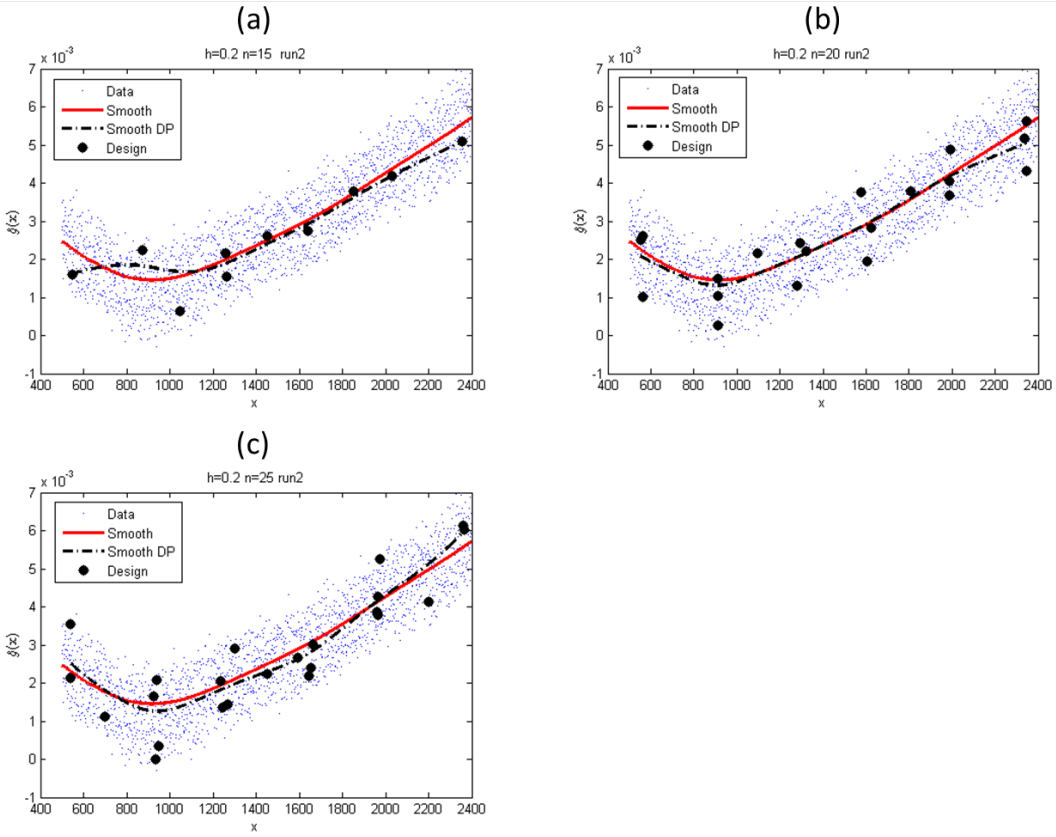

Figure 3.3: Run 2: data (small dot) and design points (large dot), with the smooth fit using whole data (-) and smooth fit using design points (-.) (a) $n = 15$, (b) $n = 20$ and (c) $n = 25$.

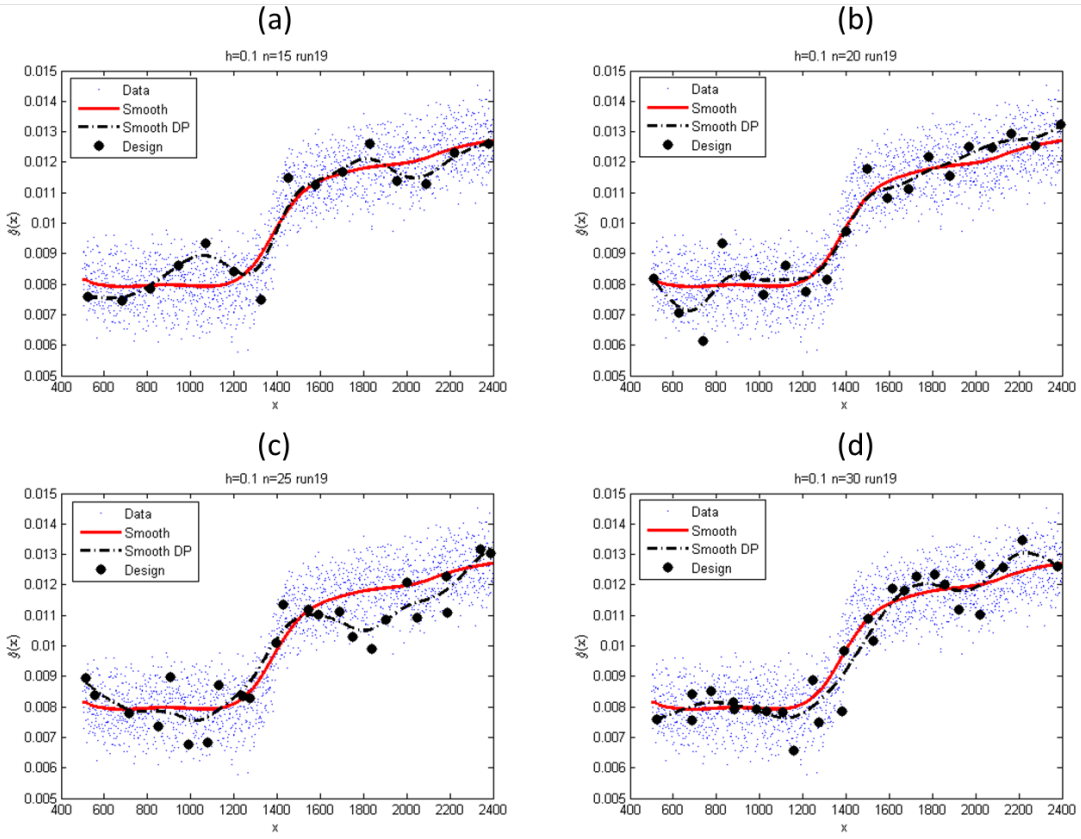

Figure 3.4: Run 19: data (small dot) and design points (large dot), with the smooth fit using whole data (-) and smooth fit using design points (-.)(a) $n = 15$, (b) $n = 20$, (c) $n = 25$ and (d) $n = 30$.

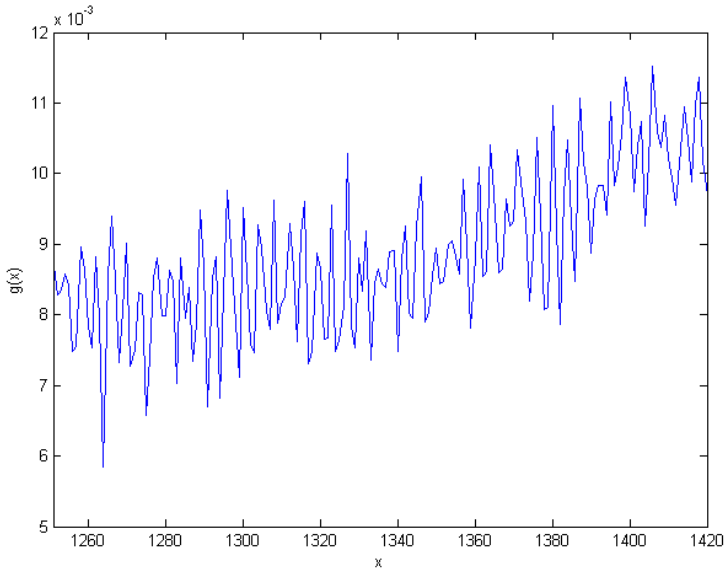

Figure 3.5: Run 19: Data on interval $[1250, 1420]$.

$$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t,$$

see Chatfield (2004). The data exhibit a large degree of seasonality, which does not have a clearly identifiable frequency. This suggests that the errors may not be independent. The optimal designs calculated in Table 3.6 are not directly applicable to experiments where these types of errors may be present because they were found under the assumption of independent errors. Finding appropriate designs for correlated error variables remains an area for future research (Chapter 6).

3.5.1 Application to simulated data

In this section, we demonstrate the design methods and assess the performance of optimal designs from this chapter using a simulated dataset obtained from the tribology data. The simulated data are formed by adding independent errors, from a Normal distribution with zero mean, to the smooth fit $\hat{g}$ from the whole dataset using bandwidths $h = 0.2$ and $h = 0.1$ for run 2 and run 19, respectively. Initially several choices of variance σ^2 were tried and a value of 2.25×10^{-8} was chosen as it was neither too small to eliminate all variability in the fit nor too large to prevent a reasonably accurate prediction.

Figures 3.7 and 3.8 present for run 2 and 19 respectively, two simulated datasets for each run, obtained from optimal designs under Criterion 3.3 with $n = 25$ (run 2) and $n = 30$

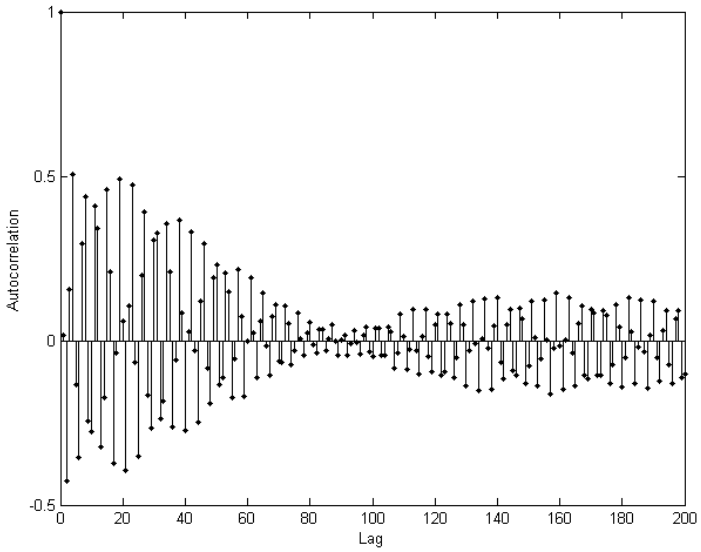

Figure 3.6: Run 19: Residual autocorrelation from fitting $\hat{g}(x)$ as a local linear estimator with $h = 0.1$.

(run 19) and smooth fits to the whole datasets and to data from the optimal designs. The datasets were obtained by adding two different sets of random errors drawn from a $N(0, 2.25 \times 10^{-8})$ distribution. Figure 3.7 shows that for run 2, even with less variable data, the prediction for $\hat{g}(x)$ using $n = 25$ design points over-predicts on $x \in [800, 1200]$. This is because $h = 0.2$ leads to oversmoothing of the data. It may be that a smaller value of h is more appropriate on $[800, 1200]$. On the other hand, a smaller bandwidth could give a 'wiggly' prediction with $n = 25$ design points for $x \in [1200, 2400]$, where the response appears more linear. Figure 3.8 shows that for run 19 the prediction using $n = 30$ design points is more accurate when the data is less variable. However the positioning of a single design point is still having a noticeable effect on prediction, which can be seen by observing the effect of the first design point on the prediction, near the beginning of the interval in the two plots in Figure 3.8.

To assess quantitatively the performance of the optimal designs, we calculated the mean squared error for the fitted model from each design and compared against the mean squared error for the fitted model using the whole dataset. This comparison was made using the standardised difference of the two mean squared errors obtained as follows. We calculated a 'moving window' mean squared error at each point x_i^*, $i = 601, ..., 2300$, as

$$MSE(x_i^*) = \sum_{k=i-100}^{k=i+100} \left[\hat{g}(x_k^*) - y_k\right]^2,$$

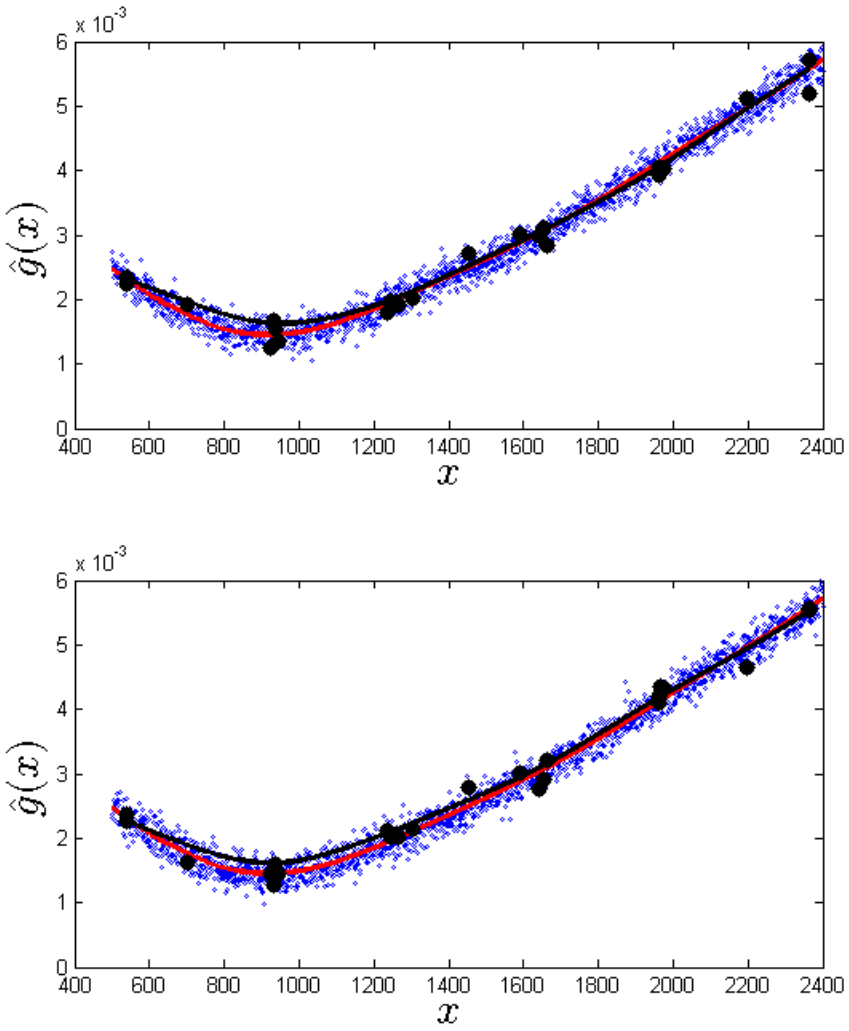

Figure 3.7: Run 2: Simulated data with different errors simulated from $N(0, 2.25 \times 10^{-8})$ for each plot (small dot), $n = 25$ design points (large dot), smooth fit using whole data (red), smooth fit using data from design points (black).

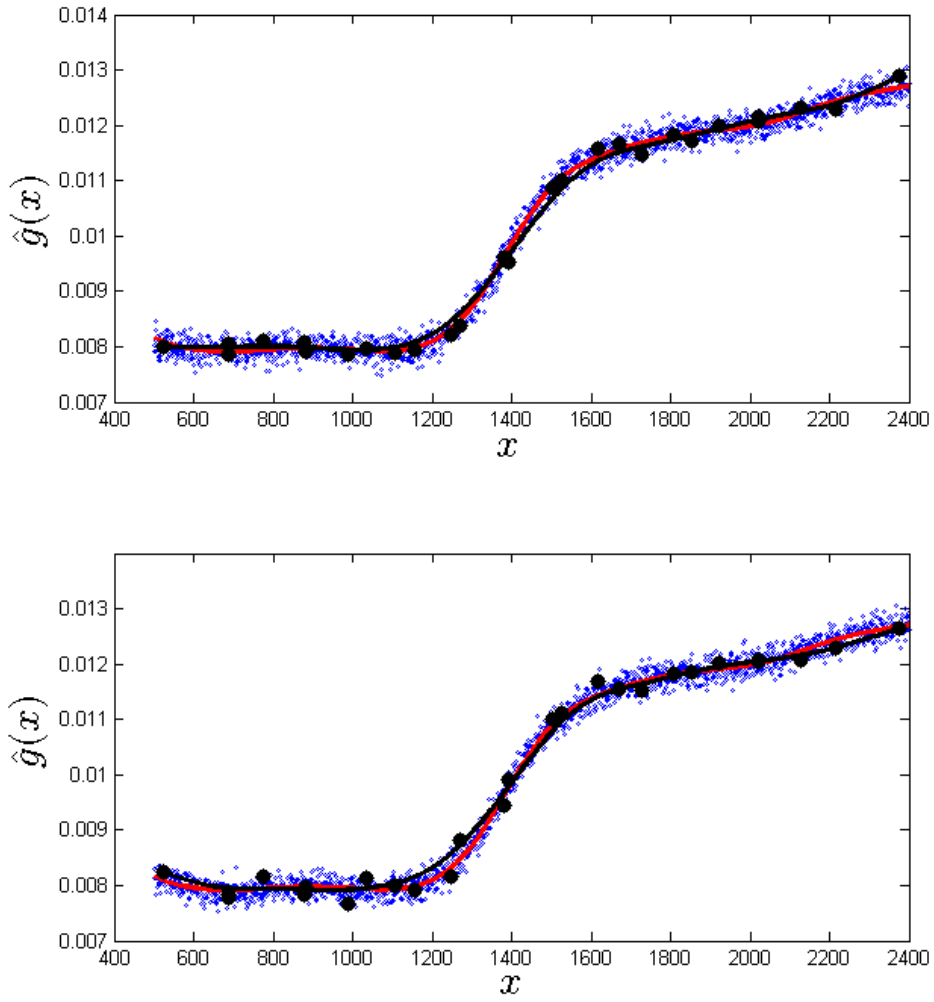

Figure 3.8: Run 19: Simulated data with different errors simulated from $N(0, 2.25 \times 10^{-8})$ for each plot (small dot), $n = 30$ design points (large dot), smooth fit using whole data (red), smooth fit using data from design points (black).

where $\hat{g}(x_k^*)$ is the kth predicted value. This moving window captures the prediction accuracy in different sub-intervals of the design region. It is used in the standardised difference in the mean squared error for a given design

$$SMSE(x_i^*) = \frac{MSE_d(x_i^*) - MSE_w(x_i^*)}{MSE_d(x_i^*)}, \tag{3.29}$$

where MSE_d and MSE_w are the mean squared errors for the fit using data corresponding to the optimal design and the fit using the whole dataset, respectively. If the standardised difference is $\delta > 0$ then there is a $\delta \times 100\%$ reduction in mean squared error when the whole dataset is used (rather than the observations from the design points) to make a prediction. If the difference is less than zero, i.e. $-\delta$, then there is a $\delta \times 100\%$ increase in mean squared error from using the whole dataset. We therefore expect to see a positive standardised difference as the prediction from the whole dataset should always be as good as, if not better than, the prediction from a subset of the data. When comparing and assessing the smooth fits over the whole interval, we use the average standardised difference

$$ASD = \frac{1}{1700} \sum_{i=601}^{2300} SMSE(x_i^*).$$

Throughout these comparisons it should be taken into account that the D_s-optimal design was found to minimise the variance and not the mean squared error.

For run 2, Figure 3.9 shows that the mean squared error is much larger for predictions obtained from either set of data on the sub-interval $[800, 1200]$ than on the rest of the interval, with the exception of $n = 15$ when $x > 1600$ where n is too small for adequate prediction. This supports the suggestion that the bandwidth is too large on $[800, 1200]$ to capture features of the data. As we would expect, the mean squared error for the prediction using the whole dataset is less than when using data from any of the optimal designs. The standardised plot, see Figure 3.10, shows that designs with 15 or 25 points have lower mean squared error values on the interval $[800, 1200]$ than the design with 20 points. Note that this is only for one simulated dataset and, as such, this difference could be due to the particular set of realised data used.

Figure 3.11 has the same plots for run 19, as were presented for run 2 in Figure 3.9. The largest mean squared error occurs around the time the steep change in the data occurs in both Figures 3.11 (a) and (b). We require a smaller bandwidth and more design points in these regions to capture these features accurately.

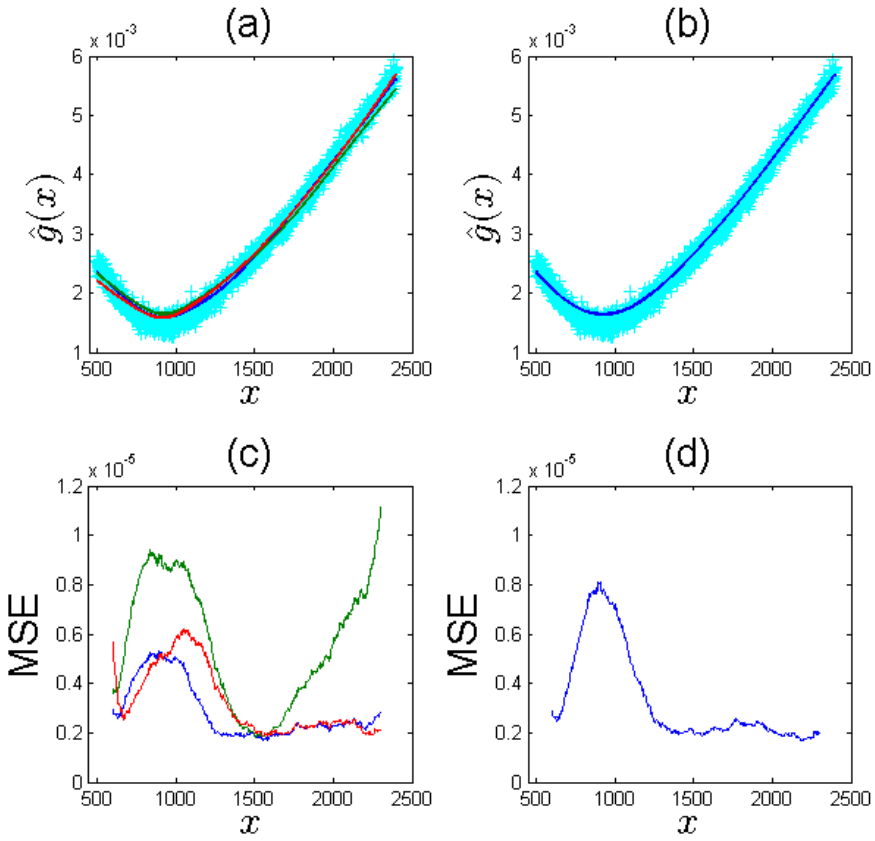

Figure 3.9: Run 2: Smooth fits and MSE plots (a) $\hat{g}(x)$ using data corresponding to optimal designs with 15 (blue), 20 (red) and 25 (green) design points, (b) $\hat{g}(x)$ for the whole dataset, (c) MSE for $\hat{g}(x)$ for 15, 20 and 25 design points and (d) MSE for $\hat{g}(x)$ for the whole dataset.

The average standardised difference (ASD) is smallest for run 2 (0.189 to 3 d.p.) when $n = 15$ and smallest for run 19 (0.166 to 3 d.p.) when $n = 25$. However we would expect the ASD to be smallest when there are 25 or 30 points, for runs 2 and 19 respectively. Again, this could be due to the single set of data.

3.5.2 Comparison with the uniform design

We now compare the optimal designs to a uniform design of equally spaced points on [501, 2400]. We again calculate the mean squared error to compare the fits obtained from using the uniform designs with $n = 15, 20, 25$ for $h = 0.2$ and $n = 15, 20, 25, 30$ for $h = 0.1$ and the whole dataset. We use the same simulated dataset as in the previous section.

We see similar results to those from use of the optimal designs, see Figures 3.13-3.16. For run 2, the mean squared error is, once again, larger on the interval $[800, 1200]$. In order

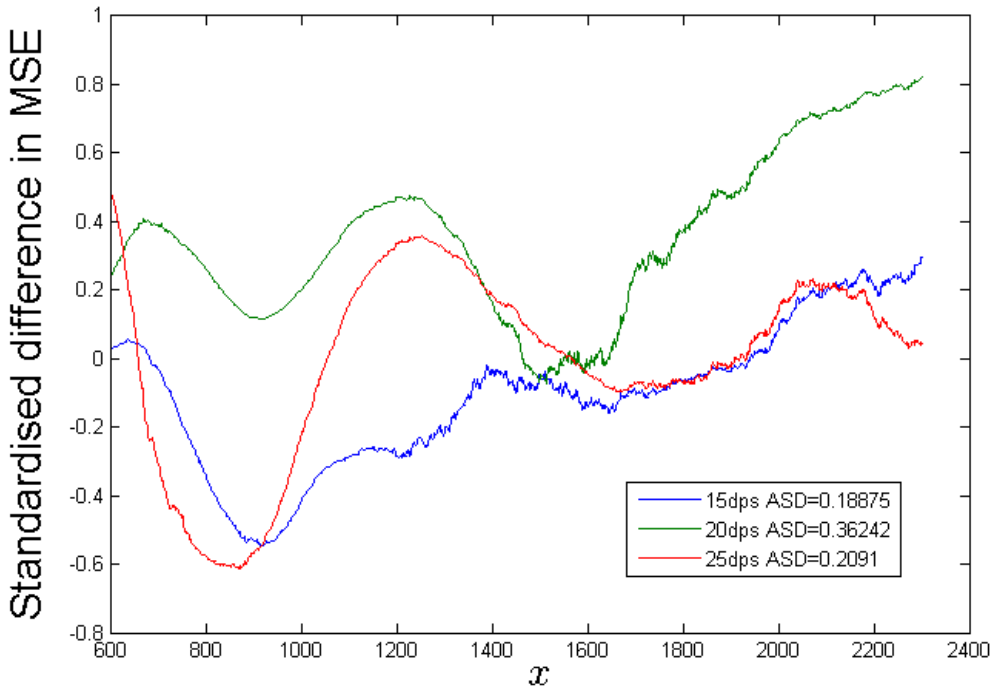

Figure 3.10: Run 2: Standardised difference (3.29) in mean square error between $\hat{g}(x)$ using data from optimal designs with 15 (blue), 20 (red) and 25 (green) design points and $\hat{g}(x)$ from the whole dataset. Values of the average standardised MSE difference (ASD) over x are given in the legend.

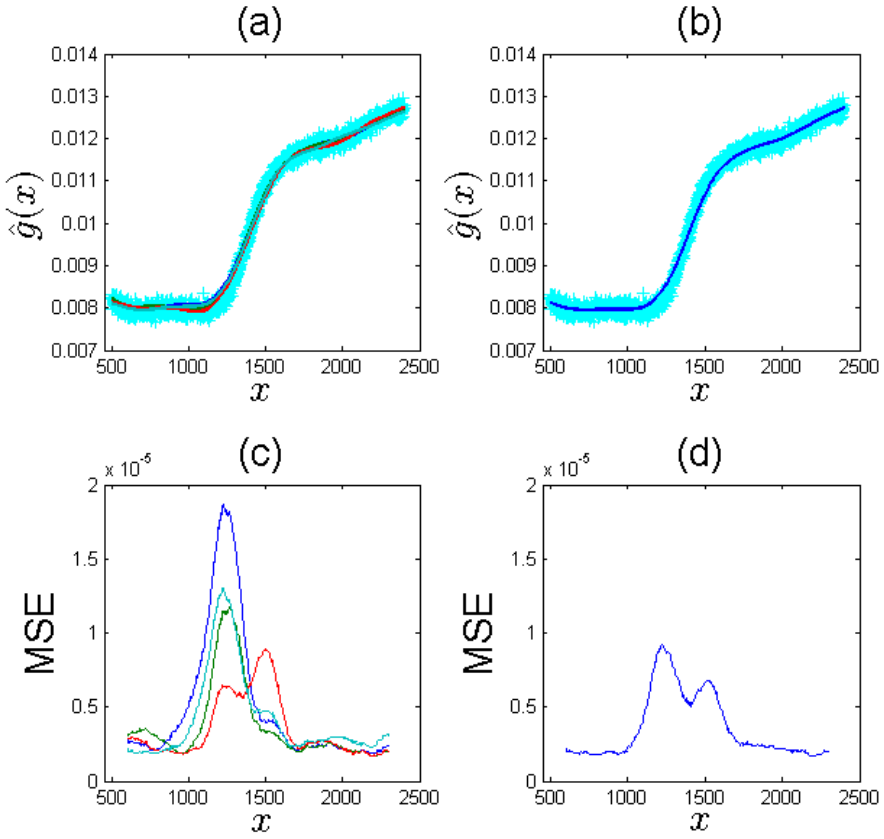

Figure 3.11: Run 19: Smooth fits and MSE plots (a) $\hat{g}(x)$ using data corresponding to optimal designs with 15 (blue), 20 (red), 25 (green) and 30 (light blue) design points, (b) $\hat{g}(x)$ from the whole dataset, (c) MSE for $\hat{g}(x)$ for 15, 20, 25 and 30 design points, and (d) MSE for $\hat{g}(x)$ for the whole dataset.

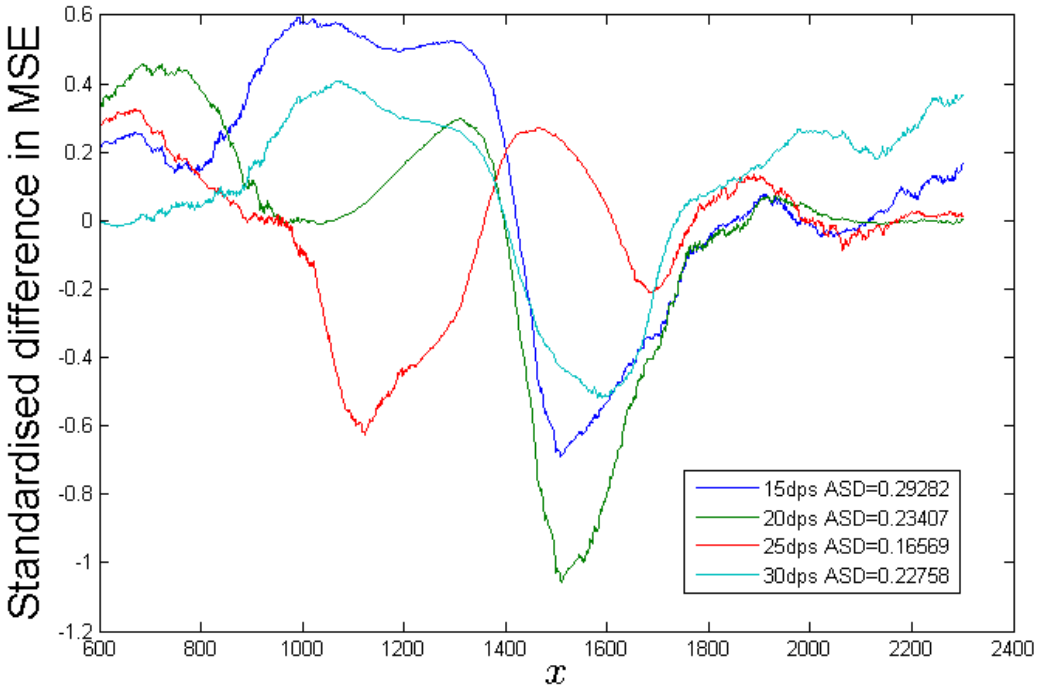

Figure 3.12: Run 19: Standardised difference (3.29) in mean square error between $\hat{g}(x)$ using data from optimal designs with 15 (blue), 20 (red), 25 (green) and 30 (light blue) design points and $\hat{g}(x)$ from the whole dataset. Values of the average standardised MSE difference (ASD) over x are given in the legend.

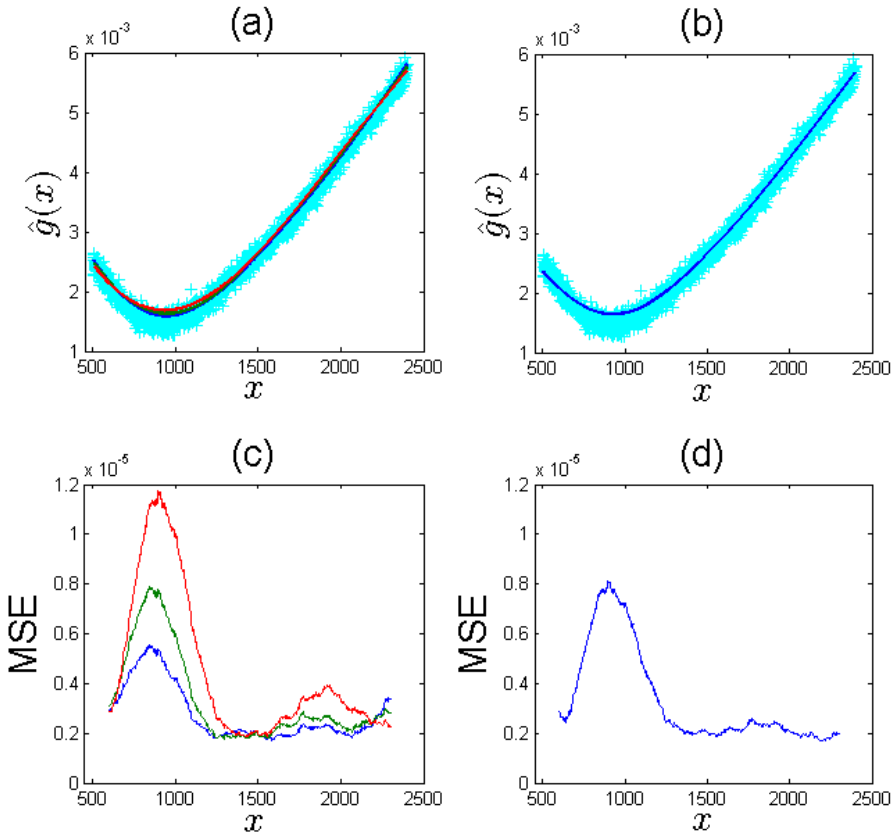

Figure 3.13: Run 2: Smooth fits and MSE plots (a) $\hat{g}(x)$ using data corresponding to uniform designs with 15 (blue), 20 (red) and 25 (green) design points, (b) $\hat{g}(x)$ from the whole dataset, (c) MSE for $\hat{g}(x)$ for 15, 20 and 25 design points and (d) MSE for $\hat{g}(x)$ for the whole dataset.

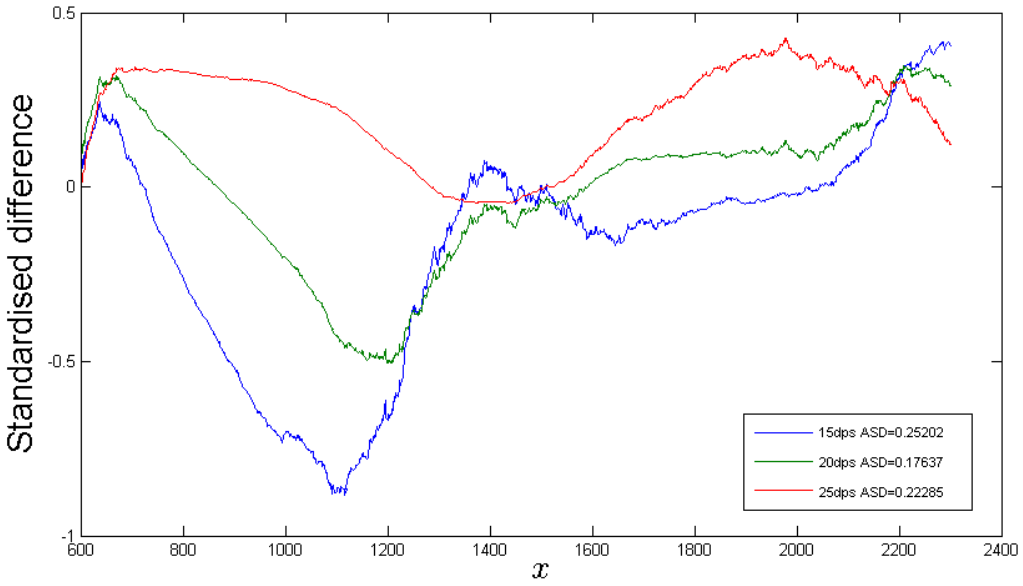

Figure 3.14: Run 2: Standardised difference (3.29) in mean square error between $\hat{g}(x)$ using data from uniform designs with 15 (blue), 20 (red) and 25 (green) design points and $\hat{g}(x)$ from the whole dataset. Values of the average standardised MSE difference (ASD) over x are given in the legend.

to compare quantitatively the optimal designs and the uniform designs, we simulated 500 datasets from each of the uniform and optimal designs for $n = 15, 20$ and 25 (run 2) and $n = 15, 20, 25$ and 30 (run 19). We computed the average standardised difference (ASD) for each design and each of the 500 datasets. Employing the Central Limit Theorem, we calculated a confidence interval for the difference in ASD between the two designs (uniform and optimal design).

n	Run 2	Run 19
15	$[-0.2033, -0.1538]$	$[-0.1292, -0.0903]$
20	$[-0.1188, -0.0825]$	$[-0.0797, -0.0492]$
25	$[-0.0812, -0.0503]$	$[-0.0982, -0.0695]$
30	-	$[-0.0928, -0.0669]$

Table 3.8: Confidence intervals for the difference in average standardised difference (ASD) between the optimal design and uniform design for each value of n.

Table 3.8 shows that, for each value of n, and for each run, the uniform design performs better than the optimal design (as the upper and lower bounds are both negative). For larger n, the uniform design has only very slightly lower SMSE. We would expect that for larger n, say $n = 50$, the optimal design would be better for both run 2 and run 19, as the optimal design concentrates points slightly more centrally than the uniform design. Alternatively, if there is further prior information available about the response, we could

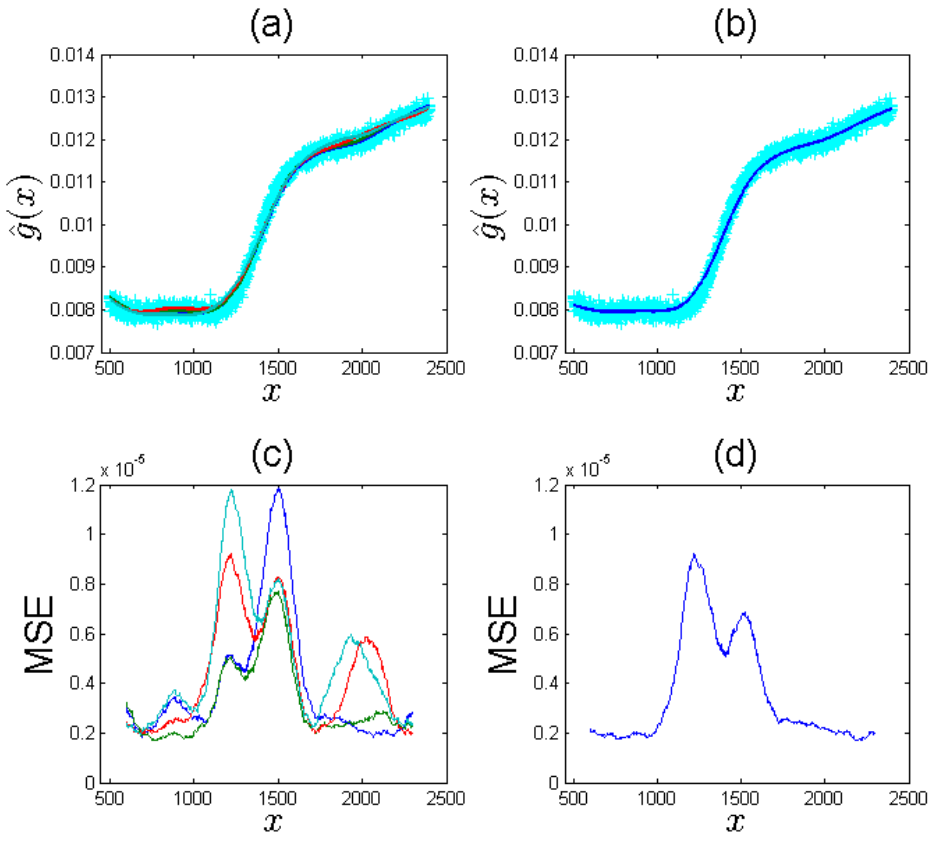

Figure 3.15: Run 19: Smooth fits and MSE plots (a) $\hat{g}(x)$ using data corresponding to uniform designs with 15 (blue), 20 (red), 25 (green) and 30 (light blue) design points, (b) $\hat{g}(x)$ from the whole dataset, (c) MSE for $\hat{g}(x)$ for 15, 20, 25 and 30 design points, and (d) MSE for $\hat{g}(x)$ for the whole dataset.

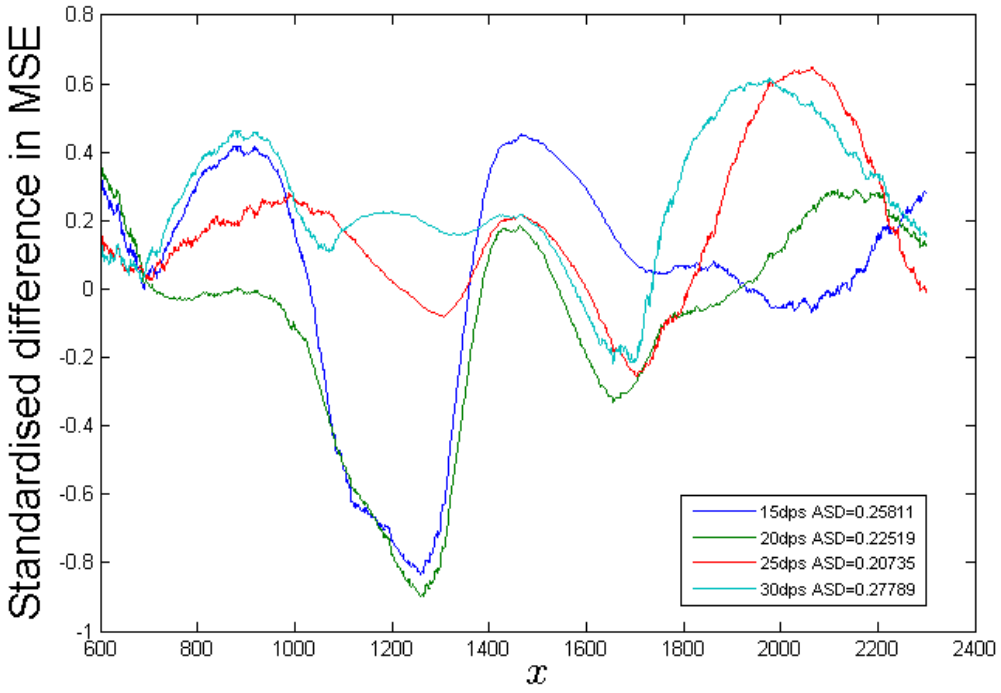

Figure 3.16: Run 19: Standardised difference (3.29) in mean square error between $\hat{g}(x)$ using data from uniform designs with 15 (blue), 20 (red), 25 (green) and 30 (light blue) design points and $\hat{g}(x)$ from the whole dataset. Values of the average standardised MSE difference (ASD) over x are given in the legend.

tailor the optimal designs and obtain an advantage over the uniform design (see Chapter 6).

3.5.3 Robustness of prediction to bandwidth selection

When studying run 2 and run 19, there has been some uncertainty in the correct choice of bandwidth for each of the two datasets. We now assess the robustness of the optimal design to the choice of bandwidth. We do this by assessing the difference in making a prediction using the whole dataset with the 'true' bandwidth and the prediction using optimal designs calculated for other bandwidths. This allows us to compare the 'best' prediction, that is, one using the whole dataset and the 'true' bandwidth, with a prediction made from an optimal design with an alternative bandwidth. By 'true' we mean the bandwidth chosen 'by eye' at the beginning of this section. For run 2, we assume that the bandwidth is $h = 0.2$, see Figure 3.3, and we use this bandwidth to make a prediction using the whole dataset. We then use the optimal designs for $h = 0.1$ and $h = 0.3$ to predict over the interval $[501, 2400]$.

For run 2 the prediction on the interval $[700, 1100]$ is much more accurate using $h = 0.1$ (see Figures 3.17 and 3.18). There is a 200% increase in mean squared error when a prediction is made using the whole dataset with $h = 0.2$ rather than data from the optimal design with $h = 0.1$. However, the prediction using data from the optimal design with $h = 0.1$ was less accurate elsewhere in the interval. This highlights the possible need for a varying bandwidth. Figures 3.19 and 3.20 show that designs with a bandwidth of $h = 0.3$ have a larger mean squared error across the whole interval.

Figures 3.21 and 3.23 show that the prediction for run 19, when the 'true' bandwidth was assumed to be $h = 0.1$ has a much larger mean squared error, especially on the interval $[1000, 1800]$ when $h = 0.2$ or $h = 0.3$. The mean squared error is very similar on the interval $[2000, 2400]$, but larger bandwidths do not perform well for this run.

3.6 Concluding Remarks

This chapter found designs which minimised a compound D_s-optimality criterion for predicting at a finite number of points and over a specified continuous interval. For prediction at a finite number of points, we found the minimum number of points required and conjectured the form of the optimal designs. More generally optimal designs were found numerically for different numbers of runs and choices of bandwidth.

For predicting at a single point, we were able to prove the optimality of the new designs for the uniform kernel. However, for prediction at a finite number of points, it was only possible to establish the optimality of the new designs in particular cases. In other cases, intuitive reasoning suggested the form of an optimal design which was supported by numerical results. The designs found for predicting across an interval were obtained numerically and, as such, some designs presented may only be near-optimal or highly efficient.

The designs for predicting over an interval were applied to the tribology experiment and assessed using a 'moving window' mean squared error. This enabled us to see that predictions made using a subset of the data obtained from the set of point, in an optimal design were very similar to the predictions made using the whole dataset. The optimal designs were also compared to the equally spaced uniform design and we found that the designs performed very similarly. Lastly, we conducted a robustness study to assess how different bandwidths performed using optimal designs for an assumed bandwidth. This study also indicated that the use of different bandwidths on different sections of the interval may have achieved a better fit for predicting the response. This supports further investigation of designs for a varying bandwidth (see Chapter 6 for further discussion).

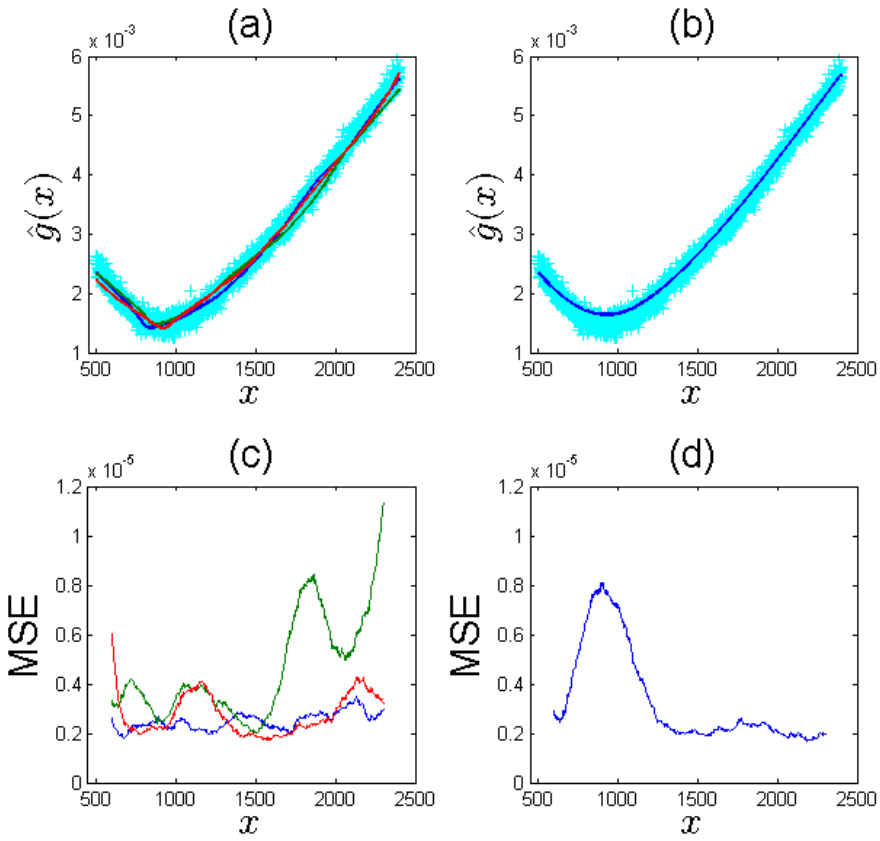

Figure 3.17: Run 2: Smooth fits and MSE plots (a) $\hat{g}(x)$ using data corresponding to optimal designs with 15 (blue), 20 (red) and 25 (green) design points for $h = 0.1$, (b) $\hat{g}(x)$ from the whole dataset and true bandwidth of $h = 0.2$, (c) MSE for $\hat{g}(x)$ for 15, 20 and 25 design points and (d) MSE for $\hat{g}(x)$ for the whole dataset.

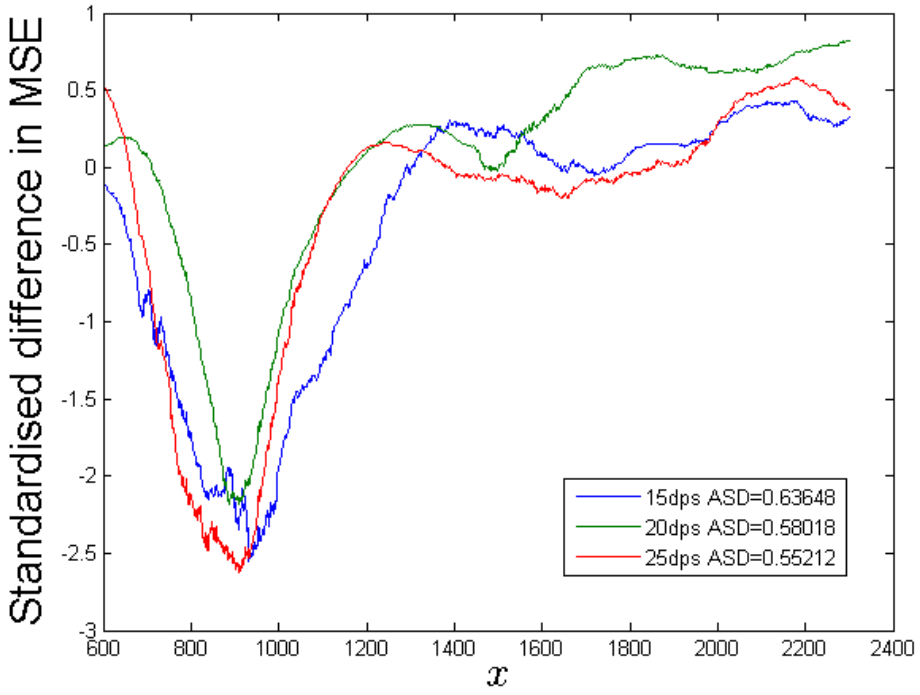

Figure 3.18: Run 2: Standardised difference (3.29) in mean square error between $\hat{g}(x)$ using data from optimal designs with 15 (blue), 20 (red) and 25 (green) design points for $h = 0.1$ and $\hat{g}(x)$ from the whole dataset with true bandwidth $h = 0.2$. Values of the average standardised MSE difference (ASD) over x are given in the legend.

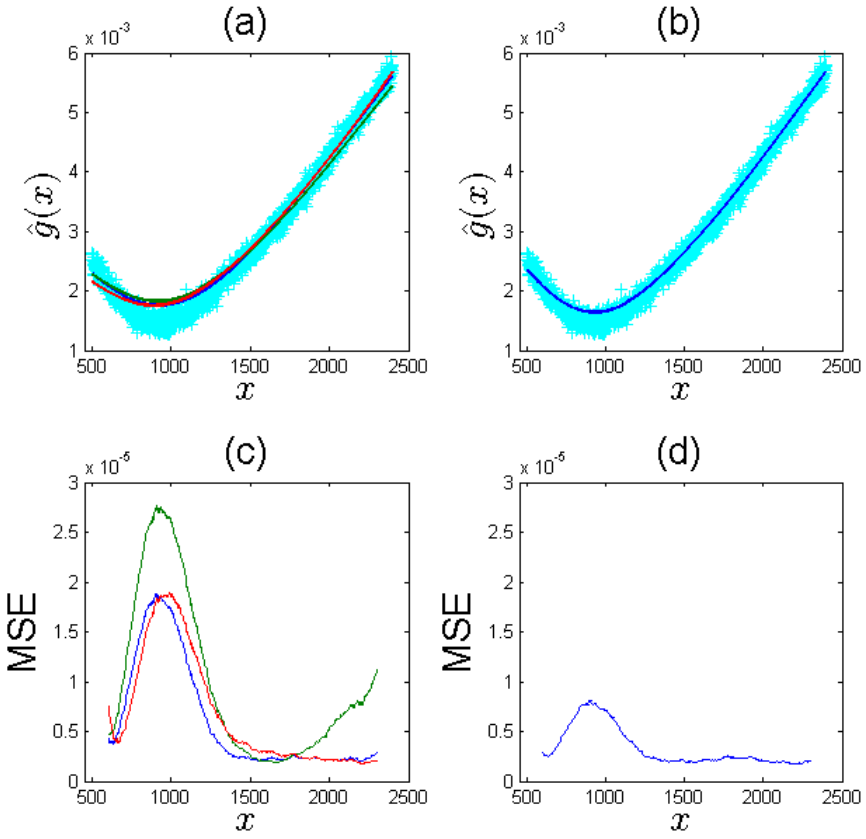

Figure 3.19: Run 2: Smooth fits and MSE plots (a) $\hat{g}(x)$ using data corresponding to optimal designs with 15 (blue), 20 (red) and 25 (green) design points for $h = 0.3$, (b) $\hat{g}(x)$ from the whole dataset with true bandwidth, $h = 0.2$, (c) MSE for $\hat{g}(x)$ for 15, 20 and 25 design points and (d) MSE for $\hat{g}(x)$ for the whole dataset.

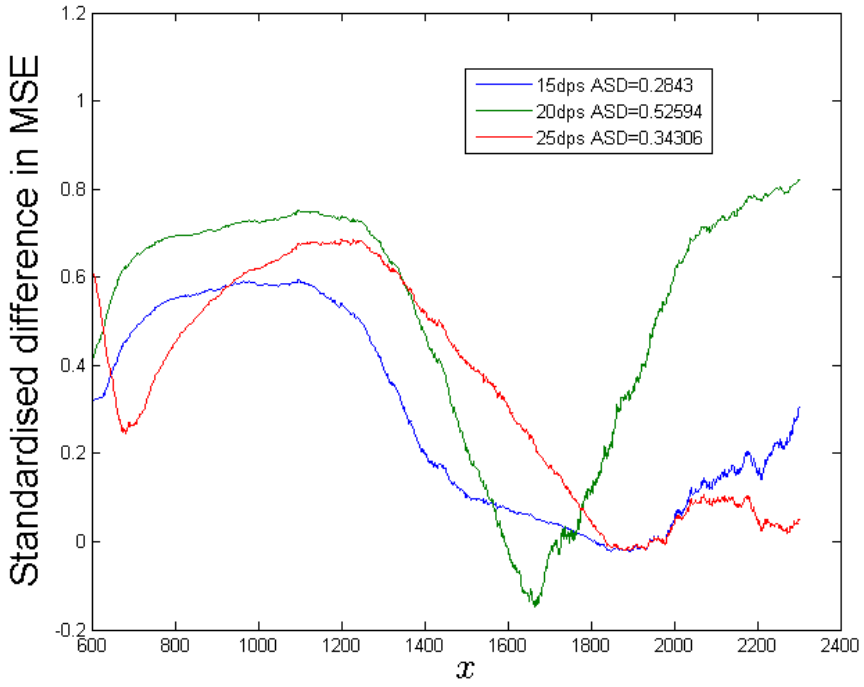

Figure 3.20: Run 2: Standardised difference (3.29) in mean square error between $\hat{g}(x)$ using data from optimal designs with 15 (blue), 20 (red) and 25 (green) design points for $h = 0.3$ and $\hat{g}(x)$ from the whole dataset with true bandwidth $h = 0.2$. Values of the average standardised MSE difference (ASD) over x are given in the legend.

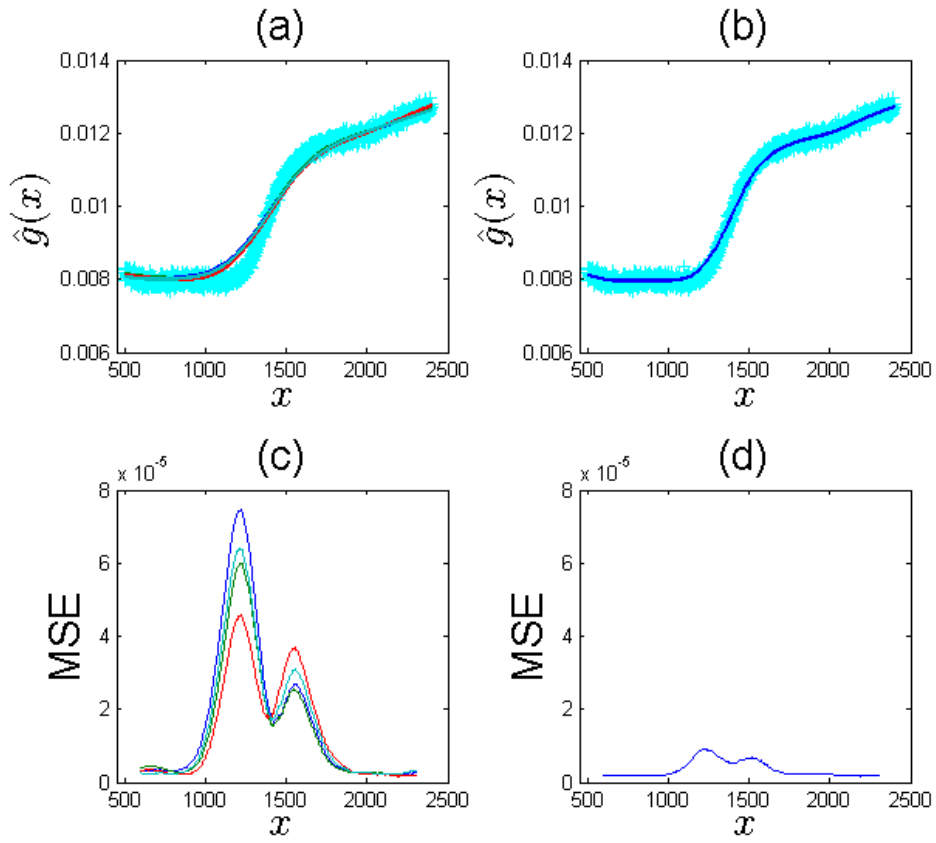

Figure 3.21: Run 19: Smooth fits and MSE plots (a) $\hat{g}(x)$ using data corresponding to optimal designs with 15 (blue), 20 (red), 25 (green) and 30 (light blue) design points for $h = 0.2$, (b) $\hat{g}(x)$ from the whole dataset with true bandwidth, $h = 0.1$, (c) MSE for $\hat{g}(x)$ for 15, 20, 25 and 30 design points, and (d) MSE for $\hat{g}(x)$ for the whole dataset.

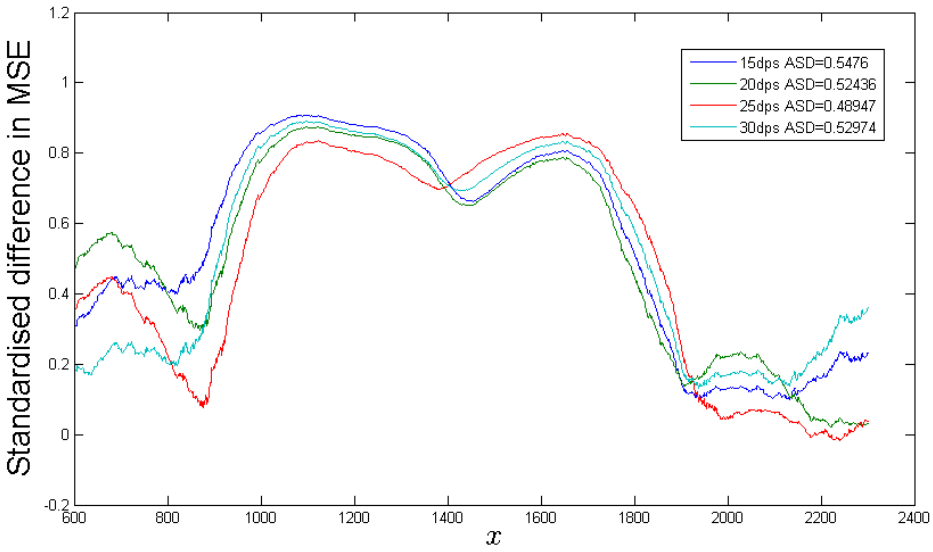

Figure 3.22: Run 19: Standardised difference (3.29) in mean square error between $\hat{g}(x)$ using data from optimal designs with 15 (blue), 20 (red), 25 (green) and 30 (light blue) design points for $h = 0.2$ and $\hat{g}(x)$ from the whole dataset with true bandwidth $h = 0.1$. Values of the average standardised MSE difference (ASD) over x are given in the legend.

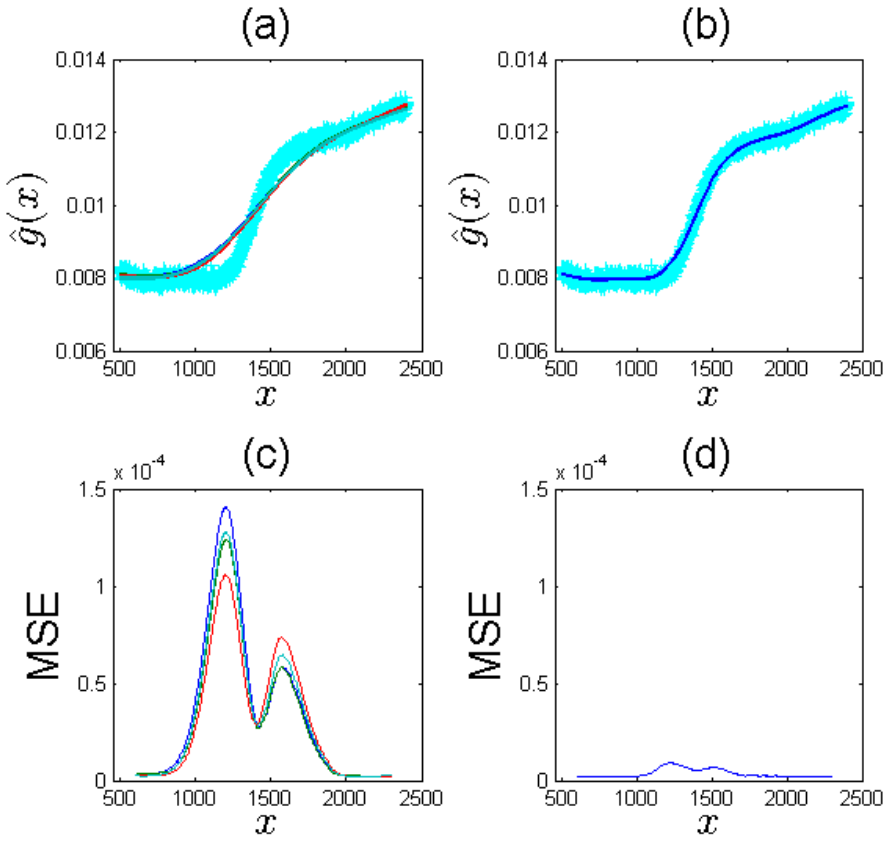

Figure 3.23: Run 19: Smooth fits and MSE plots (a) $\hat{g}(x)$ using data corresponding to optimal designs with 15 (blue), 20 (red), 25 (green) and 30 (light blue) design points for $h = 0.3$, (b) $\hat{g}(x)$ from the whole dataset with true bandwidth, $h = 0.1$, (c) MSE for $\hat{g}(x)$ for 15, 20, 25 and 30 design points, and (d) MSE for $\hat{g}(x)$ for the whole dataset.

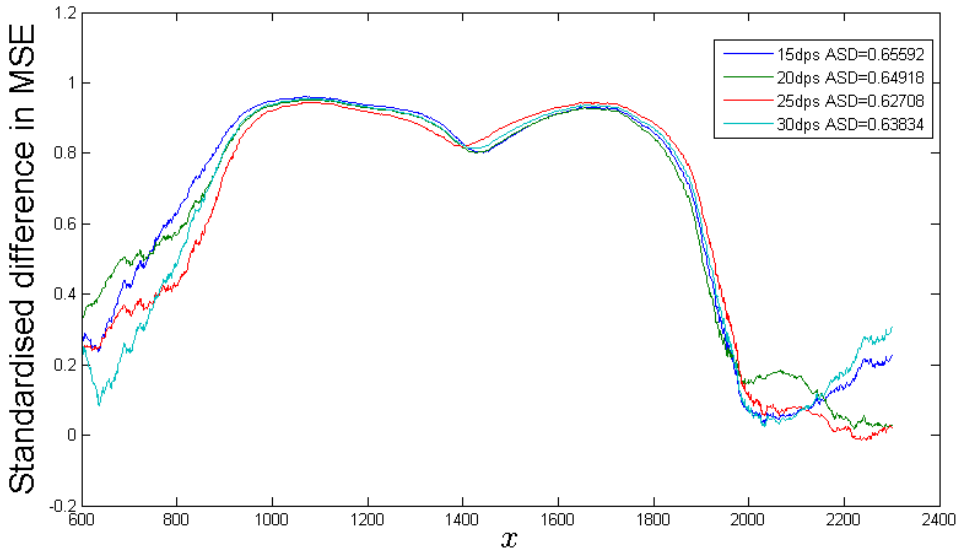

Figure 3.24: Run 19: Standardised difference (3.29) in mean square error between $\hat{g}(x)$ using data from optimal designs with 15 (blue), 20 (red), 25 (green) and 30 (light blue) design points for $h = 0.3$ and $\hat{g}(x)$ from the whole dataset with true bandwidth $h = 0.1$. Values of the average standardised MSE difference (ASD) over x are given in the legend.

Chapter 4

Compound optimal designs for prediction using kernel smoothing

In this chapter we find optimal designs for local prediction by trading-off prediction variance and the complexity of the fitted model. We demonstrate new criteria by finding designs for the Gasser and Müller estimator. Throughout this chapter, we illustrate our new criteria with examples and provide some insights using analytic results for simple cases with the uniform kernel.

We start by providing some background results on linear smoothing generally, and the Gasser and Müller estimator in particular. We then find designs that minimise prediction variance and highlight the disadvantages of this approach. To overcome these issues, we introduce a new criterion that minimises a weighted sum of the integrated prediction variance and a measure of the complexity of the fitted model, given by the inverse of the trace of the smoothing matrix. We discuss some analytic results for a special case and then find designs numerically for the uniform and Gaussian kernels.

4.1 Gasser and Müller kernel smoothing

In this section, we introduce prediction using the Gasser and Müller estimator. Suppose we have design points $x_1, \ldots, x_n$ on a single variable, with associated observations $y_1, \ldots, y_n$ where, as before, we assume that

$$y_j = g(x_j) + \epsilon_j, \tag{4.1}$$

where ϵ_j are independent error variables with constant variance for $j = 1, \ldots, n$. Then, $\text{Var}\{y_j\} = \sigma^2$. We wish to estimate the unknown function g using a linear smoother, $\hat{g}(x)$, which estimates the function through a linear combination of the observations, $y_1, \ldots, y_n$, as

$$\hat{g}(x) = \sum_{j=1}^{n} S_j(x) y_j, \tag{4.2}$$

where $S_j(x)$ are smoothing weights (for example, Ramsay and Silverman, 2005, ch. 4). Smoothing weights are defined for each type of linear smoother, see also Section 2.1.

The prediction at x^*, using the Gasser and Müller estimator (Gasser and Müller, 1979, 1984) is given by

$$\hat{g}(x^*) = \sum_{j=1}^{n} \left[\frac{1}{h} \int_{\bar{x}_{j-1}}^{\bar{x}_j} K\left(\frac{v - x^*}{h} \right) \mathrm{d}v \right] y_j,$$

where $\bar{x}_j = (x_{j+1} + x_j)/2$ for $1 \leq j < n$, $\bar{x}_0 = x_1$ and $\bar{x}_n = x_n$. The kernel function, K, is defined to be symmetric and satisfies $\int K(v)\,\mathrm{d}v = 1$. The bandwidth, h, controls the locality of the prediction; a larger value of h allows more design points to influence the prediction at x^*. The smoothing weights are therefore given by

$$S_j(x^*) = \frac{1}{h} \int_{\bar{x}_{j-1}}^{\bar{x}_j} K\left(\frac{v - x^*}{h} \right) \mathrm{d}v. \tag{4.3}$$

4.1.1 The smoothing matrix

The corresponding smoothing matrix, S, see Ramsay and Silverman (2005, p. 64) is defined as

$$S = \begin{bmatrix} S_1(x_1) & S_2(x_1) & \ldots & S_n(x_1) \\ S_1(x_2) & S_2(x_2) & \ldots & S_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ S_1(x_n) & S_2(x_n) & \ldots & S_n(x_n) \end{bmatrix}. \tag{4.4}$$

Note that for linear models, the trace of S equals the number of parameters; for example

the trace of S for simple linear regression is two as there are two parameters to estimate in the fit.

The trace of the smoothing matrix (4.4) can be used to provide a measure of effective degrees of freedom of a smooth fit (Ramsay and Silverman, 2005, p. 67). It therefore gives a measure of the complexity of the fitted model.

For the Gasser and Müller estimator, the trace of the smoothing matrix is

$$\text{trace}(S) = \frac{1}{h}\sum_{j=1}^{n}\int_{\bar{x}_{j-1}}^{\bar{x}_j} K\left(\frac{v - x_j}{h}\right)\mathrm{d}v. \tag{4.5}$$

4.1.2 The uniform kernel

The uniform kernel is defined as

$$K(u) = \begin{cases} 0.5 & \text{if } |u| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

In obtaining expressions for the smoothing weights (4.3), we use the following indicator function. For a specified interval $A \subset \mathbb{R}$,

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

The smoothing weights for the uniform kernel can then be written as

$$\begin{aligned} S_j(x) &= \frac{1}{h}\int_{\bar{x}_{j-1}}^{\bar{x}_j} K\left(\frac{v - x}{h}\right)\mathrm{d}v \\ &= \frac{1}{h}\int_{\bar{x}_{j-1}}^{\bar{x}_j} \frac{1}{2}1_A(x)\mathrm{d}v, \end{aligned} \tag{4.6}$$

where

$$A = \{v; |x - v| \leq h\}. \tag{4.7}$$

Evaluating (4.6) we find the smoothing weight, $S_j(x)$, as

$$S_j(x) = \frac{1}{2h}[\min(x + h, \bar{x}_j) - \max(x - h, \bar{x}_{j-1})]. \tag{4.8}$$

for $\bar{x}_{j-1} - h \leq x \leq \bar{x}_j + h$ or zero otherwise. To show (4.8) holds, for each $j = 1, \ldots, n$, we consider a set (interval in $\mathbb{R}$) defined by

$$X_j = \{v;\, \bar{x}_{j-1} \leq v \leq \bar{x}_j\},$$

and a subset

$$A_j(x) = \{v \in X_j;\, x - h \leq v \leq x + h\}.$$

Then we can define

$$\begin{aligned} I_j(x) &= \int_{X_j} 1_{A_j}\, \mathrm{d}v \\ &= 2hS_j(x), \end{aligned} \tag{4.9}$$

with

$$1_{A_j}(x) = \begin{cases} 1 & \text{if } x \in A_j \\ 0 & \text{otherwise.} \end{cases}$$

Note that when A_j is the empty set, then $I_j(x) = 0$, by definition of 1_{A_j}.

The set A_j is not empty when the intervals $[x - h, x + h]$ and $[\bar{x}_{j-1}, \bar{x}_j]$ overlap, i.e.

$$\bar{x}_{j-1} - h \leq x \leq \bar{x}_j + h. \tag{4.10}$$

This argument implies $A_j(x)$ is the interval from $\max(x-h, \bar{x}_{j-1})$ to $\min(x+h, \bar{x}_j)$. Therefore

$$\begin{aligned} I_j(x) &= \int_{A_j} 1 \mathrm{d}v + \int_{X_j \setminus A_j} 0 \mathrm{d}v \\ &= \min(x+h, \bar{x}_j) - \max(x-h, \bar{x}_{j-1}). \end{aligned} \tag{4.11}$$

On substituting (4.11) into (4.9), we conclude that (4.8) is satisfied.

In general, the trace is given by

$$\mathrm{trace}(S) = \frac{1}{2h} \sum_{j=1}^{n} [\min(x_j + h, \bar{x}_j) - \max(x_j - h, \bar{x}_{j-1})]. \tag{4.12}$$

From (4.12), it is clear that as the bandwidth, h, increases the trace decreases and therefore the complexity of the model decreases. This is intuitive as increasing h allows more points to influence the prediction at x^*, i.e. a lower level of local smoothing is assumed. Increasing the bandwidth in local linear regression also provides a less complex model, see Section 2.1.1.

Special case:

In order to find a class of analytical designs in Section 4.3, we evaluate the smoothing weight when

$$\bar{x}_{j-1} \leq x \leq \bar{x}_j,$$

and

$$\begin{aligned} \bar{x}_j - \bar{x}_{j-1} &\leq h \\ \Rightarrow x_{j+1} - x_{j-1} &\leq 2h. \end{aligned} \tag{4.13}$$

From (4.10) and (4.13), we see that

$$\begin{aligned} \bar{x}_j - x &\leq h \\ \Leftrightarrow \frac{x_{j+1} + x_j}{2} - x &\leq h \\ \Rightarrow x_j - x &\leq h, \end{aligned}$$

since $\frac{x_{j+1}+x_j}{2} \geq x_j$. Similarly,

$$\begin{aligned} x - \bar{x}_{j-1} &\leq h \\ \Leftrightarrow x - \frac{x_j + x_{j-1}}{2} &\leq h \\ \Rightarrow x - x_j &\leq h, \end{aligned}$$

since $\frac{x_j+x_{j-1}}{2} \leq x_j$. Therefore, $|x - x_j| \leq h$ for all x. Hence, when (4.10) and (4.13) hold

$$\begin{aligned} x + h - \bar{x}_j &= x + h - \frac{x_{j+1} + x_j}{2} \\ &= \frac{2x - x_{j+1} - x_j}{2} + h \\ &= \frac{h + (x - x_{j+1}) + h + (x - x_j)}{2} \\ &\geq 0 \\ \Rightarrow x + h &\geq \bar{x}_j. \end{aligned}$$

Similarly $\bar{x}_{j-1} \geq x - h$. Therefore

$$\begin{aligned} S_j(x^*) &= \frac{1}{h} \int \frac{1}{2} 1_A \, \mathrm{d}v \\ &= \frac{1}{h} \int_{\bar{x}_{j-1}}^{\bar{x}_j} \frac{1}{2} \, \mathrm{d}v \\ &= \frac{\bar{x}_j - \bar{x}_{j-1}}{2h}, \end{aligned}$$

and

$$\begin{aligned}\text{trace}(S) &= \frac{1}{2h}\sum_{j=1}^{n}[\bar{x}_j - \bar{x}_{j-1}] \\ &= \frac{x_n - x_1}{2h}. \end{aligned} \tag{4.14}$$

Clearly, if (4.13) is not satisfied, $S_j(x) \leq (\bar{x}_j - \bar{x}_{j-1})/2h$, and (4.14) is an upper-bound on trace(S).

4.2 Designs to minimise prediction variance

In this section we find a design $\xi_n = \{x_1, \ldots x_n\}$ which minimises the prediction variance at a point x^*. In general, the variance of a linear smoother, $\hat{g}(x^*)$, under model (4.1) is given by

$$\begin{aligned}\text{Var}\left\{\hat{g}(x^*)\right\} &= \text{Var}\left(\sum_{j=1}^{n} S_j(x^*)y_j\right) \\ &= \sum_{j=1}^{n} S_j(x^*)^2\text{Var}(y_j) \\ &= \sigma^2\sum_{j=1}^{n} S_j(x^*)^2. \end{aligned} \tag{4.15}$$

From (4.15), it is clear that the sum of the squared smoothing weights, $\sum_{j=1}^{n} S_j(x^*)^2$, must be minimised in order to minimise Var$\{\hat{g}(x^*)\}$.

Criterion 4.1. *A design ξ_n^* for a linear smoother is optimal if it minimises the prediction variance (4.15) at a single point x^*. That is*

$$\xi^* = \arg\min_{\xi}\sum_{j=1}^{n} S_j(x^*)^2. \tag{4.16}$$

The prediction variance for the Gasser and Müller estimator is given by

$$\begin{aligned}\text{Var}\left\{\hat{g}(x^*)\right\} &= \text{Var}\left(\sum_{j=1}^{n} S_j(x^*)y_j\right) \\ &= \frac{\sigma^2}{h}\sum_{j=1}^{n}\left(\int_{\bar{x}_{j-1}}^{\bar{x}_j} K\left(\frac{v-x^*}{h}\right)\mathrm{d}v\right)^2. \qquad (4.17)\end{aligned}$$

It is straightforward to establish optimal designs under Criterion 4.1 for any kernel function.

Proposition 4.1. *The prediction variance, Var$\{\hat{g}(x^*)\}$, for the Gasser and Müller estimator is minimised for any kernel function by the design that takes all points x_j to be equal, i.e. has just one distinct design point and $x_1 = \ldots = x_n$.*

Proof. Assume that all design points are equal and, without loss of generality, set $x_j = x_1$ for all $2 \leq j \leq n$. Then,

$$\begin{aligned}S_j(x^*) &= \frac{1}{h}\int_{\frac{x_1+x_1}{2}}^{\frac{x_1+x_1}{2}} K\left(\frac{v-x^*}{h}\right)\mathrm{d}v \\ &= \frac{1}{h}\int_{x_1}^{x_1} K\left(\frac{v-x^*}{h}\right)\mathrm{d}v.\end{aligned}$$

As K is a real-valued integrable function defined at x_1, by the Fundamental Theorem of Calculus

$$\begin{aligned}S_j(x^*) &= \left[F\left(\frac{x_1-x^*}{h}\right) - F\left(\frac{x_1-x^*}{h}\right)\right] \\ &= 0,\end{aligned}$$

for all $1 \leq j \leq n$, where $F(x)$ is the anti-derivative of $K(x)$. Thus the variance is simply calculated as

$$\begin{aligned}\text{Var}\left\{\hat{g}(x^*)\right\} &= \sum_{j=1}^{n} S_j(x^*)^2\sigma^2 \\ &= 0.\end{aligned}$$

□

There is a major drawback to using designs found from Criterion 4.1. As the smoothing weights are all set to be zero for $j = 1, \ldots, n$, from (4.2) we see that $\hat{g}(x^*) = \sum_{j=1}^{n} S_j(x^*)y_j = 0$. Hence, a design from Criterion 4.1 minimises the prediction variance by only allowing a zero prediction of $\hat{g}(x^*)$, which does not depend on y_j. This is equivalent to fitting a statistical model with no parameters. This null model has the largest possible bias of $g(\hat{x^*})$.

The mean squared error could be reduced by increasing the variance and reducing the bias. However, in order to reduce the bias through choice of design we are required to make assumptions about the form of the model g. Instead, we add a less restrictive constraint to our objective function to ensure that the design allows a more realistic prediction to be made.

4.3 Constrained and compound designs for the uniform kernel

We now consider designs which minimise the prediction variance with respect to a constraint on the effective degrees of freedom of a smooth fit, ensuring a more complex model can be fitted to the resulting data.

4.3.1 Illustration and results for a simple case

We start by choosing a design to minimise the prediction variance at a single point given a fixed model complexity.

Criterion 4.2. *A design ξ_n^*, is optimal for prediction at a point x^* using a linear smoother if*

$$\xi_n^* = \arg\min_{\xi_n} Var\{\hat{g}(x^*)\} \quad subject\ to \quad trace(S) = d \geq 0.$$

Let χ denote the set of all possible design points, called the design region. For any $x \in \chi$, we obtain the variance of $\hat{g}(x)$ by substitution of the general form of the smoothing weights (4.8) into the variance formula in (4.17) to give

$$\text{Var}\{\hat{g}(x)\} = \frac{\sigma^2}{4h^2} \sum_{j=1}^{n} 1_{B_j}(x) \left[\min(x+h, \bar{x}_j) - \max(x-h, \bar{x}_{j-1})\right]^2, \tag{4.18}$$

where

$$B_j = \{u \in \chi; \bar{x}_{j-1} - h \leq u \leq \bar{x}_j + h\},$$

for $j = 1, \ldots, n$.

We now consider finding designs from Criterion 4.2 under assumption (4.13), that is, $x_{j+1} - x_{j-1} \leq 2h$. This assumption leads via equation (4.14) to the constraint in Criterion 4.2 having the form

$$\text{trace}(S) = \frac{x_n - x_1}{2h} = d. \tag{4.19}$$

Note that x_1 and x_n must be chosen to be distinct, otherwise $\text{trace}(S) = 0$ and the result is again the null model. Hence, this constraint on the trace clearly prevents the design coalescing to a single point.

Under the constraint (4.19), it follows that the length of the interval on which we make a prediction is $x_n - x_1 = 2hd$ and so $|x - x_j| \leq 2hd$ since $x, x_j \in [x_1, x_n]$ for $j = 1, ..., n$. If $d \leq 0.5$ and $x^* \in [x_1, x_n]$, $j = 1, \ldots, n$, then the smoothing weights for observation y_j for prediction at x^* has value $S_j(x^*) = (\bar{x}_j - \bar{x}_{j-1})/2h$, where $\bar{x}_j = (x_{j+1} + x_j)/2$, as $1_{B_j}(x) = 1$ for all $x \in \chi$. This leads to

$$\begin{aligned}
\text{Var}\left\{\hat{g}(x^*)\right\} &= \sum_{j=1}^{n} \left[S_j(x^*)\right]^2 \sigma^2 \\
&= \sum_{j=1}^{n} \frac{\sigma^2}{4h^2} [\bar{x}_j - \bar{x}_{j-1}]^2 \\
&= \frac{\sigma^2}{16h^2} \left[(x_2 - x_1)^2 + \sum_{j=2}^{n-1} (x_{j+1} - x_{j-1})^2 + (x_n - x_{n-1})^2 \right].
\end{aligned}$$

It will be useful to represent $\text{Var}\left\{\hat{g}(x^*)\right\}$ as

$$\text{Var}\left\{\hat{g}(x^*)\right\} = \frac{\sigma^2}{16h^2} \sum_{j=1}^{n_1} I_{1,j}^2 + \frac{\sigma^2}{16h^2} \sum_{j=1}^{n_2} I_{2,j}^2, \tag{4.20}$$

where n_1 and n_2 are integers such that $n_1 + n_2 = n$ and $I_{1,j}$ and $I_{2,j}$ are defined in (4.21) and (4.22).

To establish results for optimal designs under Criterion 4.2, we consider two cases: (a) $n = 2m$ and (b) $n = 2m + 1$ for some integer $m \geq 1$.

Proposition 4.2. *If $n = 2m$ and $d \leq 0.5$, the optimal design under Criterion 4.2 has design points x_1, $x_n = x_1 + 2dh$ and $x_{2j} = x_{2j+1} = x_1 + 4dhj/n$, $j = 1, ..., (n-2)/2$. That is, the design has $n/2 + 1$ distinct points equally spaced over the closed interval $[x_1, x_n]$.*

Proof. Consider expression (4.20). As $d \leq 0.5$, every smoothing weight is non-zero, since $1_{B_j}(x) = 1$ for all $x \in \chi$, with value $S_j(x^*) = (\bar{x}_j - \bar{x}_{j-1})/2h$. Then the elements of each summation in (4.20) may be chosen independently since no two elements contain the same pair of x_j. Then, as n is even, we set $n_1 = n_2 = n/2$ and define

$$I_{1,j} = \begin{cases} x_{2j+1} - x_{2j-1} & \text{for} \quad j = 1, ..., \frac{n-2}{2} \\ x_n - x_{n-1} & \text{for} \quad j = \frac{n}{2}, \end{cases} \tag{4.21}$$

$$I_{2,j} = \begin{cases} x_2 - x_1 & \text{for} \quad j = 1 \\ x_{2j} - x_{2j-2} & \text{for} \quad j = 2, ..., \frac{n}{2}. \end{cases} \tag{4.22}$$

Hence,

$$\sum_{j=1}^{n/2} I_{1,j} = \sum_{j=1}^{n/2} I_{2,j} = x_n - x_1 = 2dh. \tag{4.23}$$

By applying Result 1 from the Appendix on minimising the sum of squared terms to each summation in (4.20), we see that to minimise $\sum_{j=1}^{n/2} I_{1,j}^2$ we set $I_{1,j} = I_{1,k}$ for $j, k = 1, ..., n/2$. Hence to minimise $\text{Var}\{\hat{g}(x^*)\}$ we set

$$x_3 - x_1 = \ldots = x_{2j+1} - x_{2j-1} = \ldots = x_{n-3} - x_{n-1} = x_n - x_{n-1}$$

and $x_1, x_3, x_5, ..., x_{n-3}, x_{n-1}, x_n$ must be equally spaced. Then, using (4.23), we see that $I_{1,j} = 2dh/(n/2) = 4dh/n$ for $j = 1, ..., n/2$ and the design points are

$$\begin{aligned} x_{2j+1} &= x_1 + 4dhj/n, \quad j = 0, ..., (n-2)/2 \\ x_n &= x_1 + 2dh. \end{aligned} \tag{4.24}$$

For the summation of I_{2j}, an analogous set of design points are obtained

$$x_{2j} = x_1 + 4dhj/n, \quad j = 1, ..., (n-2)/2. \tag{4.25}$$

From (4.24) and (4.25), it is clear that $x_{2j} = x_{2j+1}$ for $j = 1, ..., (n-2)/2$. □

The prediction variance from this optimal design is then given by

$$\begin{aligned} \text{Var}\{\hat{g}(x^*)\} &= \frac{\sigma^2}{16h^2} \sum_{j=1}^{n/2} \left[I_{1,j}^2 + I_{2,j}^2 \right] \\ &= \frac{\sigma^2 n d^2}{2n^2} + \frac{\sigma^2 n d^2}{2n^2} \\ &= \frac{\sigma^2 d^2}{n}. \end{aligned} \tag{4.26}$$

Notice that, as expected, the variance is an increasing function of d, the complexity of the fitted model, and a decreasing function of n.

Example 1

We illustrate Proposition 4.2 with a simple example when $n = 4$. Suppose that $n_1 = n_2 = 2$ and

$$\begin{aligned} I_{1,1} &= x_3 - x_1 \\ I_{1,2} &= x_4 - x_3 \\ I_{2,1} &= x_2 - x_1 \\ I_{2,2} &= x_4 - x_2. \end{aligned}$$

Then, by Proposition 4.2, the design points are given by x_1, $x_2 = x_3 = x_1 + dh$ and $x_4 = x_1 + 2dh$. Note that the three points x_1, $x_2 = x_3$ and x_4 are equally spaced. The minimum variance in this case is $\sigma^2 d^2/4$.

Proposition 4.3. *If $n = 2m + 1$ and $d \leq 0.5$ the optimal design under Criterion 4.2 has design points*

$$\begin{aligned} x_{2j+1} &= x_1 + \frac{4dhj}{n-1} \qquad for\, j = 0, ..., \frac{n-1}{2} \\ x_{2j} &= x_1 + \frac{4dhj}{n+1} \qquad for\, j = 1, ..., \frac{n-1}{2}. \end{aligned}$$

Proof. We separate the variance into two independent summations as in (4.20) with $n_1 = (n-1)/2$ and $n_2 = (n+1)/2$:

$$\text{Var}\left\{\hat{g}(x^*)\right\} = \frac{\sigma^2}{16h^2} \sum_{j=1}^{(n-1)/2} I_{1,j}^2 + \frac{\sigma^2}{16h^2} \sum_{j=1}^{(n+1)/2} I_{2,j}^2,$$

where $I_{1,j} = x_{2j+1} - x_{2j-1}$ for $j = 1, \ldots, (n-1)/2$,

$$I_{2,j} = \begin{cases} x_2 - x_1 & \text{for} \quad j = 1 \\ x_{2j} - x_{2j-2} & \text{for} \quad j = 2, \ldots, \frac{(n-1)}{2} \\ x_n - x_{n-1} & \text{for} \quad j = \frac{n+1}{2}, \end{cases}$$

and

$$\sum_{j=1}^{(n-1)/2} I_{1,j} = \sum_{j=1}^{(n+1)/2} I_{2,j} = 2dh. \tag{4.27}$$

To minimise $\sum_{j=1}^{(n-1)/2} I_{1,j}^2$ we again apply Result 1 from the Appendix to each summation in (4.27). This sets $I_{1,j} = I_{1,k}$ for $j, k = 1, ..., (n-1)/2$, and determines that

$$x_3 - x_1 = \ldots = x_{2j+1} - x_{2j-1} = \ldots = x_{n-2} - x_{n-4} = x_n - x_{n-2}, \tag{4.28}$$

and hence $x_1, x_3, x_5 ..., x_{n-4}, x_{n-2}, x_n$ must be equally spaced. Then by (4.27) we see that $I_{1,j} = 4dh/(n-1)$ for $j = 1, ..., (n-1)/2$ and the design points are

$$x_{2j+1} = x_1 + \frac{4dhj}{n-1}, \quad j = 0, ..., \frac{n-1}{2}.$$

Similarly, minimisation of $\sum_{j=1}^{(n+1)/2} I_{2,j}^2$ leads to the design points

$$x_{2j} = x_1 + \frac{4dhj}{n+1}, \quad j = 1, ..., \frac{n-1}{2}.$$

□

The prediction variance for this optimal design is then given by

$$
\begin{aligned}
\mathrm{Var}\left\{\hat{g}(x^*)\right\} &= \frac{\sigma^2}{16h^2}\left[\sum_{j=1}^{(n-1)/2} I_{1,j}^2 + \sum_{j=1}^{(n+1)/2} I_{2,j}^2\right] \\
&= \frac{\sigma^2(n-1)(dh)^2}{2(n-1)^2} + \frac{\sigma^2(n+1)(dh)^2}{2(n+1)^2} \\
&= \sigma^2\left[\frac{d^2}{2(n-1)} + \frac{d^2}{2(n+1)}\right] \\
&= \frac{\sigma^2 d^2 n}{(n-1)(n+1)}.
\end{aligned}
$$

Notice that the variance is a function of the same order in d and n as when $n = 2m$.

Example 2

We illustrate Proposition 4.3 with a simple example when $n = 5$. Suppose that $n_1 = 2$, $n_2 = 3$ and

$$
\begin{aligned}
I_{1,1} &= x_3 - x_1 \\
I_{1,2} &= x_5 - x_3 \\
I_{2,1} &= x_2 - x_1 \\
I_{2,2} &= x_4 - x_2 \\
I_{2,3} &= x_5 - x_4
\end{aligned}
$$

Then by, Proposition 4.2, the optimal design has points $x_1, x_2 = x_1 + 2dh/3, x_3 = x_1 + dh, x_4 = x_1 + 4dh/3$ and $x_5 = x_1 + 2dh$. The minimum variance of $\hat{g}(x)$ is $5\sigma^2 d^2/24$.

In practice, we are unlikely to want to use linear smoothers with effective degrees of freedom as small as $d = 0.5$. We have found that it is not possible to find optimal prediction designs analytically in general for the Gasser and Müller estimator, and hence in the next section we use computational methods.

4.4 Prediction variance for the uniform kernel in general

In this section, we find designs without the restriction that $d \leq 0.5$. Now, it is impossible to give a simple closed form for the smoothing weight $S_j(x)$. Recall that for any $x \in \chi$ the variance of $\hat{g}(x)$ is given in (4.18) by

$$\text{Var}\left\{\hat{g}(x)\right\} = \frac{\sigma^2}{h^2} \sum_{j=1}^{n} 1_{B_j}(x) \left[\min(x+h, \bar{x}_j) - \max(x-h, \bar{x}_{j-1})\right]^2,$$

where

$$B_j = \left\{u \in \chi; \bar{x}_{j-1} - h \leq u \leq \bar{x}_j + h\right\},$$

for $j = 1, \ldots, n$.

Finding designs analytically under Criterion 4.2 for this general case is an intractable problem due to the form of the prediction variance. In the next section, we find designs numerically using a constraint on the trace of the smoothing matrix given in (4.12).

4.4.1 Integrated prediction variance for the uniform kernel in general

Designs that minimise the prediction variance integrated across an interval are more useful for real experiments than designs that simply minimise the prediction variance at a single point. Therefore, we now find designs satisfying Criterion 4.3.

Criterion 4.3. *A design ξ_n^*, is optimal for prediction on the interval [-1,1] using a linear smoother if*

$$\xi_n^* = \arg\min_{\xi_n} \int_{-1}^{1} \mathit{Var}\left\{\hat{g}(x^*)\right\} \mathrm{d}x^* \qquad \textit{subject to } \mathit{trace}(S) \geq d \geq 0,$$

which can be reformulated as

$$\xi_n^* = \arg\min_{\xi_n} \int_{-1}^{1} Var\{\hat{g}(x^*)\}\,\mathrm{d}x^* \qquad subject\ to\ \frac{1}{trace(S)} \leq \frac{1}{d}. \tag{4.29}$$

Clyde and Chaloner (1996) established, in general, the equivalence between a constrained criterion, such as Criterion 4.3, and compound criterion, such as the following

Criterion 4.4. *A design ξ_n^*, is optimal for prediction on the interval $[-1, 1]$ using a linear smoother if*

$$\Psi(\xi_n^*) = \min_{\xi_n} \Psi(\xi_n),$$

where

$$\Psi(\xi_n) = (1-\lambda) \int_{-1}^{1} Var\{\hat{g}(x^*)\}\,\mathrm{d}x^* + \frac{\lambda}{trace(S)}, \tag{4.30}$$

and $0 < \lambda < 1$.

Clearly, (4.30) is similar in structure to a Lagrange function (Arfken et al., 2012) for (4.29). However, minimisation of the objective function (4.30) through choice of ξ_n and the Lagrange multiplier λ results in $\lambda = 0$ and $\int \mathrm{Var}\{\hat{g}(x^*)\}\,\mathrm{d}x^* = 0$ through coalescence of design points. Hence, we treat λ as a tuning constant and find designs under Criterion 4.4 for given values of λ.

Finding designs to minimise (4.30) is an analytically intractable problem. Therefore, we find optimal designs computationally and use Legendre-Gauss quadrature (see Section 3.4) to evaluate the integrated variance with $m = 25$ abscissa values, $x_1^*, \ldots, x_m^*$, and weights $\kappa_1, \ldots, \kappa_m$. Hence we find designs minimising

$$(1-\lambda) \sum_{i=1}^{m} \kappa_i \mathrm{Var}\{\hat{g}(x_i^*)\} + \frac{\lambda}{\mathrm{trace}(S)}, \quad 0 < \lambda < 1. \tag{4.31}$$

4.4.1.1 Results

In this section, we present optimal designs under Criterion 4.4 for a variety of values of h, n and λ and the results are given in Tables 4.1–4.4. Figures 4.1-4.4 give the prediction variance for $\lambda = 0.999$ and $\lambda = 0.3$ for each h over the interval $[-1, 1]$. The values of n were chosen separately for each value of h: use of a small value of h implies that we wish

to fit a more complex model and therefore require more design points than when h is large. The parameter λ controls the weight given to the complexity constraint. Smaller values of λ result in designs that minimise the variance to be favoured. We restrict $0 < \lambda < 1$. In this chapter, the design $-\xi_n^*$, is defined as a design composed of reflections of the points of ξ_n^* in the line $x = 0$.

We assume that $\sigma^2 = 1$. Note that if we choose a different values of σ^2, the designs in Tables 4.1–4.4 are still optimal but for a different value of λ.

For all results with fixed n and h, decreasing λ resulted in the optimal design covering a reduced range of the prediction interval $[-1, 1]$. For example, when $h = 0.2, n = 6$ and $\lambda = 0.999$, an optimal design is

$$\xi_n^* = \{-1.00, -0.60, -0.20, 0.20, 0.60, 1.00\}.$$

However when $\lambda = 0.05$, we obtain

$$\xi_n^* = \{-0.17, -0.07, -0.06, 0.06, 0.07, 0.17\},$$

which has a much smaller range.

The effect of reducing the range of the design points can be seen, for example, in Figure 4.1. The lower plot shows the design points for the design where $h = 0.2$ and $\lambda = 0.3$ in Table 4.1, which has smallest point $x_1 = -0.23$ and largest point $x_6 = 0.5$. (There are two points at 0.24). The plot shows that the prediction variance is zero for $x < -0.23 - h$ and $x > 0.5 + h$. In this example the integrated variance is minimised by making a constant prediction, $\hat{g}(x) = 0$, for points outside $[x_1 - h, x_n + h]$. For most practical experiments, this represents too much weight being given to the variance term in (4.30).

As λ decreases, we note from Tables 4.1–4.4 that both the integrated variance and trace(S) decrease. For example, from Table 4.1 we see that when $n = 6$ and $h = 0.2$, the value of $\lambda = 0.999$ gives $\int \text{Var} = 1.04$ and trace$(S) = 5$, whereas $\lambda = 0.05$ gives $\int \text{Var} = 0.029$ and trace$(S) = 0.86$. These results indicate that smaller values of λ lead to a design with the capacity to estimate a less complex model and hence producing a smaller variance.

Smaller values of h, and hence larger n, result in an optimal design having more clustered design points so that a large part of the interval [-1,1] has constant zero prediction. As an example, we consider four designs for $\lambda = 0.3$ given in Tables 4.1–4.4. The designs are for $n = 6$ and $h = 0.2$, $\xi_n^* = \{-0.23, -0.03, 0.02, 0.24, 0.24, 0.5\}$. For $n = 5$ and $h = 0.3$, $\xi_n^* = \{-0.38, -0.08, 0.03, 0.20, 0.49\}$. For $n = 4$ and $h = 0.5$ $\xi_n^* = \{-0.49, 0.06, 0.06, 0.56\}$. For $n = 3$ and $h = 1$, $\xi_n^* = \{-0.62, 0.08, 0.97\}$. These designs and their prediction

variances can be seen in Figures 4.1–4.4. As h increases and n decreases, we see that the prediction variance increases and trace(S) decreases for fixed λ. The same patten is discovered for other fixed values of λ.

In conclusion, decreasing λ has the effect of reducing the the range of the optimal design and reducing the prediction variance and trace(S). The explanation is that in this case more weight is being given to minimising the integrated variance component of (4.30), resulting in a decreased integrated variance and trace(S). Reducing the value of h and increasing n also reduced the range of the optimal designs. However, in this case the variance decreases and trace(S) increases. We would expect the trace to increase with decreasing h, to allow a more complex model to be fitted.

Unfortunately, in this study we were not able to reduce h below 0.2. This was because smaller h required larger n and the optimisation became computationally expensive when using the uniform kernel function for prediction.

	$n = 6$		
λ	ξ_n^*	$\int$ Var	tr(S)
0.999	±1.00 ±0.60 ±0.20	1.04	5.00
0.9	-0.82 -0.42 -0.09 0.38 0.65	0.77	4.55
0.8	-0.79 ±0.45 ±0.05 0.76	0.54	3.88
0.7	±0.65 ±0.36 ±0.12	0.39	3.26
0.5	±0.50 ±0.19(2)	0.23	2.49
0.3	-0.23 -0.03 0.02 0.24(2) 0.50	0.12	1.78
0.1	-0.06 0.09 0.12 0.24(2) 0.38	0.05	1.11
0.05	±0.17 ±0.06 ±0.07	0.029	0.86

Table 4.1: Optimal designs from Criterion 4.4 for predicting over the interval $[-1, 1]$ using the uniform kernel with $h = 0.2$. Design $-\xi_n^*$ is also optimal. Numbers of repetitions of design points are shown in parenthesis.

Figures 4.1-4.4 show how the prediction variance varies over the interval $[-1, 1]$. We see, from the top plot in all figures, that when $\lambda = 0.999$, the variance is greater than zero on the whole interval. Hence, we are able to predict across the whole interval using the optimal designs. However when $\lambda = 0.3$ (lower plot), we see that, especially for small h, the prediction variance is only non-zero for part of the interval. The remaining section of the interval has zero smoothing weights since there are no design points within h of these prediction points.

We also see from the figures that by reducing λ from 0.999 to 0.3 the prediction variance becomes much smaller. This is due to the fact that we are placing more importance on minimising the variance than the inverse of $trace(S)$ in these cases.

	$n = 5$		
λ	ξ_n^*	$\int \text{Var}$	$\text{tr}(S)$
0.999	$\pm$1.00 -0.40 0.20 0.47	0.86	3.33
0.9	$\pm$1.00 -0.40 0.20 0.47	0.86	3.33
0.8	-0.75 -0.25 0.32 0.49 1	0.64	2.91
0.7	$\pm$0.72 $\pm$0.23 0	0.42	2.40
0.5	$\pm$0.58$\pm$0.20 0.00	0.27	1.92
0.3	-0.38 -0.08 0.03 0.20 0.49	0.15	1.45
0.1	-0.22 -0.03 0.05 0.15 0.33	0.06	0.91
0.05	$\pm$0.21 $\pm$0.06 0.00	0.04	0.70
	$n = 7$		
λ	ξ_n^*	$\int \text{Var}$	$\text{tr}(S)$
0.999	$\pm$1.00 $\pm$0.59 $\pm$0.43 0	0.67	3.33
0.9	$\pm$1.00 $\pm$0.59 $\pm$0.43 0	0.67	3.33
0.8	-0.83 -0.34 -0.17 0.16 0.51 0.63 1.00	0.55	3.05
0.7	-0.71 -0.32 -0.15 0.10 0.42 0.57 0.89	0.41	2.67
0.5	-0.62 -0.33 -0.20 0 0.20 0.33 0.62	0.23	2.06
0.3	-0.47 -0.23 -0.19 0.02 0.18 0.27 0.49	0.14	1.60
0.1	-0.10 0.06 0.11 0.19 0.29 0.34 0.48	0.05	0.98
0.05	-0.07 0.06 0.11 0.18 0.25 0.28 0.40	0.03	0.78

Table 4.2: Optimal designs from Criterion 4.4 for predicting over the interval $[-1, 1]$ using the uniform kernel with $h = 0.3$. Design $-\xi_n^*$ is also optimal.

	$n = 4$		
λ	ξ_n^*	$\int \mathrm{Var}$	$\mathrm{tr}(S)$
0.999	$\pm$1.00 $\pm$0.12	0.64	2.00
0.9	$\pm$1.00 $\pm$0.12	0.64	2.00
0.8	$\pm$1.00 $\pm$0.12	0.64	2.00
0.7	$\pm$1.00 $\pm$0.12	0.64	2.00
0.5	-0.47 0.29 0.29 1.00	0.35	1.47
0.3	-0.49 0.06 0.06 0.56	0.18	1.04
0.1	-0.36 0.00 0.00 0.36	0.08	0.71
0.05	-0.22 0.05 0.05 0.33	0.05	0.55
	$n = 6$		
λ	ξ_n^*	$\int \mathrm{Var}$	$\mathrm{tr}(S)$
0.999	$\pm$1.00 $\pm$0.35(2)	0.45	2.00
0.9	$\pm$1.00 $\pm$0.35(2)	0.45	2.00
0.8	$\pm$1.00 $\pm$0.35(2)	0.45	2.00
0.7	$\pm$1.00 $\pm$0.35(2)	0.45	2.00
0.5	-0.57 -0.06 0.05 0.55(2)	0.28	1.57
0.3	-0.70 -0.36 -0.29 0.08 0.12 0.51	0.17	1.21
0.1	-0.35 -0.09(2) 0.18 0.20	0.07	0.79
0.05	$\pm$0.31 $\pm$0.11 $\pm$0.10	0.01	0.62

Table 4.3: Optimal designs from Criterion 4.4 for predicting over the interval $[-1, 1]$ using the uniform kernel with $h = 0.5$. Design $-\xi_n^*$ is also optimal. Numbers of repetitions of design points are shown in parenthesis.

	$n = 3$		
λ	ξ_n^*	$\int$ Var	tr(S)
0.999	±1.00 0.00	0.46	1
0.9	±1.00 0.00	0.46	1
0.8	±1.00 0.00	0.46	1
0.7	±1.00 0.00	0.46	1
0.5	±1.00 0.00	0.46	1
0.3	-0.62 0.08 0.97	0.31	0.79
0.1	0.06 0.48 1.00	0.11	0.47
0.05	0.25 0.62 1.00	0.06	0.37
	$n = 5$		
λ	ξ_n^*	$\int$ Var	tr(S)
0.999	±1.00 ±0.31 0.00	0.27	1.00
0.9	±1.00 ±0.31 0.00	0.27	1.00
0.8	±1.00 ±0.31 0.00	0.27	1.00
0.7	±1.00 ±0.31 0.00	0.27	1.00
0.5	±1.00 ±0.31 0.00	0.27	1.00
0.3	-0.99 -0.31 0.00 0.29 0.97	0.26	0.98
0.1	-0.14 0.21 0.37 0.57 1.00	0.09	0.57
0.05	0.11 0.39 0.53 0.67 1.00	0.05	0.44

Table 4.4: Optimal designs from Criterion 4.4 for predicting over the interval $[-1, 1]$ using the uniform kernel with $h = 1$. Design $-\xi_n^*$ is also optimal.

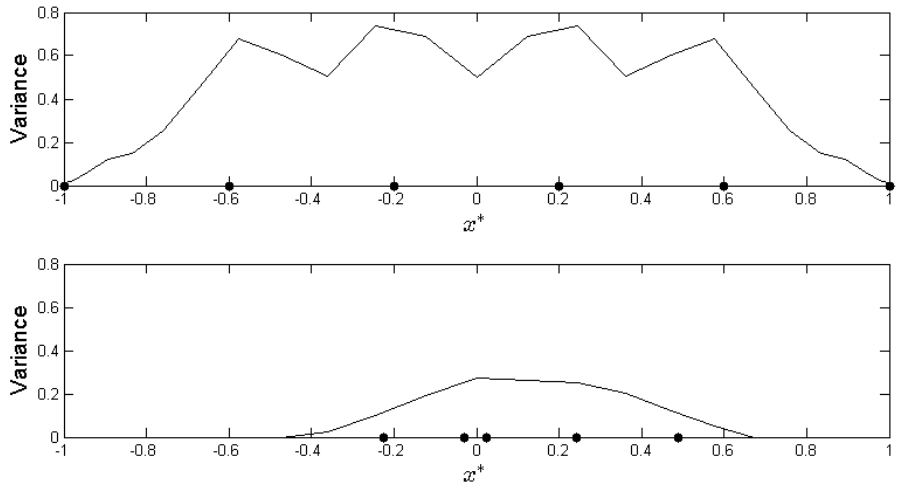

Figure 4.1: Prediction variance using the Criterion 4.4 optimal design (Table 4.1) for the uniform kernel for $h = 0.2$ and $n = 6$: λ=0.999 and trace$(S) = 5$ (top), and $\lambda = 0.3$ and trace$(S) = 1.78$ (lower). Location of the optimal design points are displayed on the x-axis.

As we might expect, the prediction variance is symmetric about 0 when the optimal design is symmetric, see Figure 4.3 for example.

4.5 Designs for the Gaussian kernel

The Gaussian kernel function is given by:

$$K\left(\frac{v-x}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(v-x)^2}{2h^2}\right\},$$

leading to the following smoothing weights using the Gasser and Müller estimator

$$\begin{aligned} S_j(x) &= \frac{1}{h\sqrt{2\pi}} \int_{\bar{x}_{j-1}}^{\bar{x}_j} \exp\left\{-\frac{1}{2}\left(\frac{v-x}{h}\right)^2\right\} \mathrm{d}v \\ &= \Phi\left(\frac{\bar{x}_j - x}{h}\right) - \Phi\left(\frac{\bar{x}_{j-1} - x}{h}\right). \end{aligned} \tag{4.32}$$

Here Φ is the standard normal cumulative distribution function. Unlike the uniform kernel, the Gaussian kernel is not truncated and hence $S_j(x) > 0$ for all $x \in \chi$ and $j = 1, \ldots, n$.

In this section we find optimal designs satisfying Criterion 4.4, which we recall is given by

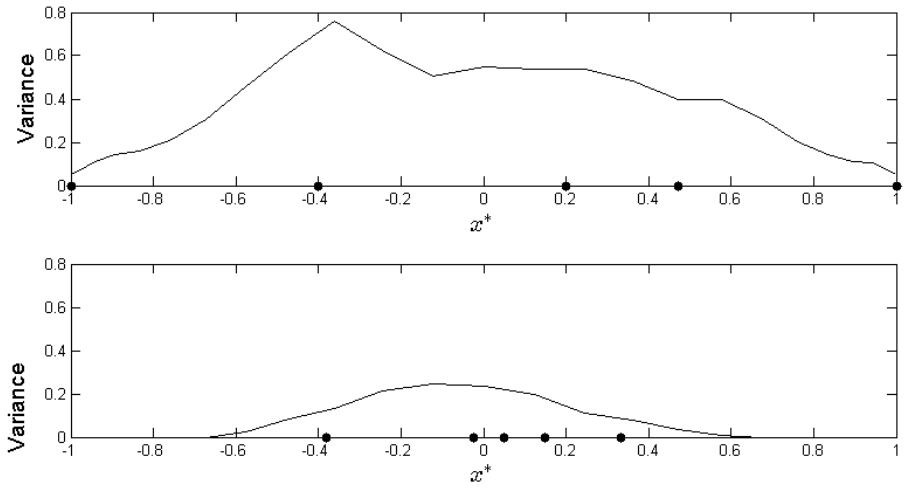

Figure 4.2: Prediction variance using the Criterion 4.4 optimal design (Table 4.2) for the uniform kernel for $h = 0.3$ and $n = 5$: $\lambda = 0.999$ and trace$(S) = 3.33$ (top), and $\lambda = 0.3$ and trace$(S) = 1.45$ (lower). Location of optimal design points are displayed on the x-axis.

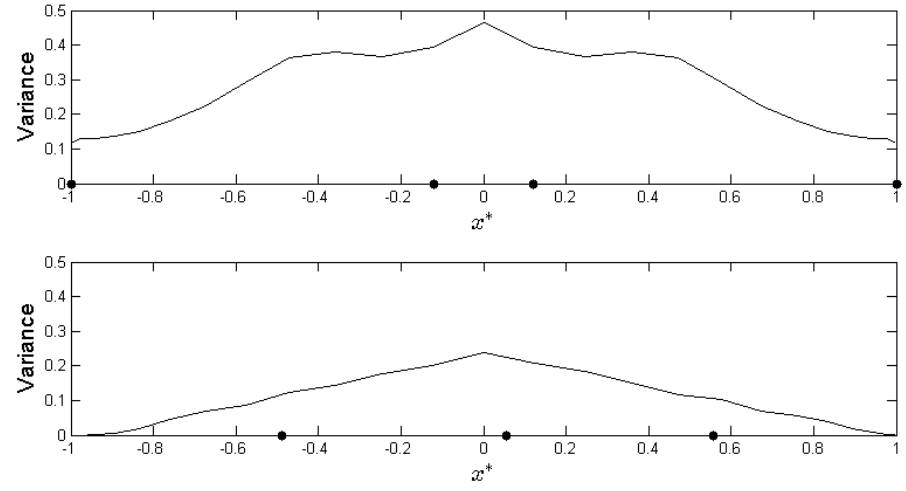

Figure 4.3: Prediction variance using the Criterion 4.4 optimal design (Table 4.3) for the uniform kernel for $h = 0.5$ and $n = 4$: $\lambda = 0.999$ and trace$(S) = 2$ (top), and $\lambda = 0.3$ and trace$(S) = 1.04$ (lower). Location of optimal design points are displayed on the x-axis.

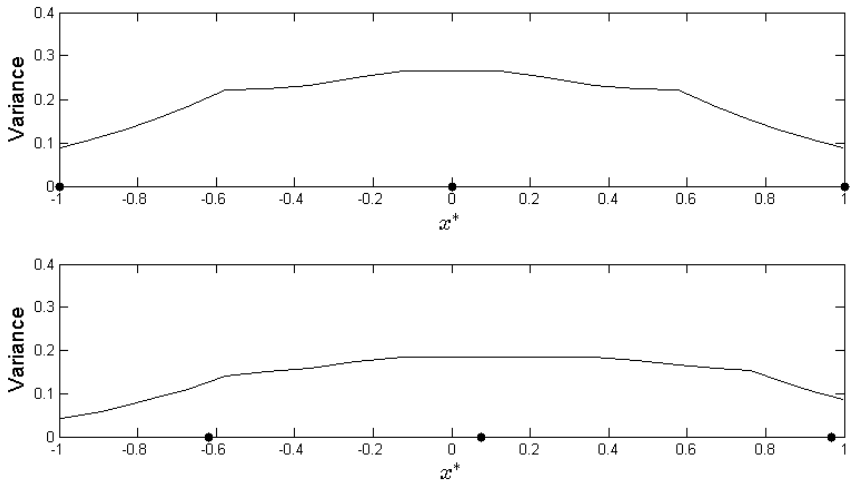

Figure 4.4: Prediction variance using the Criterion 4.4 optimal design (Table 4.4) for the uniform kernel for $h = 1$ and $n = 3$: $\lambda = 0.999$ and trace$(S) = 1$ (top), and $\lambda = 0.3$ and trace$(S) = 0.79$ (lower). Location of optimal design points are displayed on the x-axis.

Criterion 4.4. *A design ξ_n^* is optimal for prediction on the interval $[-1, 1]$ using a linear smoother if*

$$\xi_n^* = \arg\min_{\xi_n} \Psi(\xi_n),$$

with

$$\Psi(\xi_n) = (1 - \lambda) \int_{-1}^{1} Var\{\hat{g}(x^*)\}\, \mathrm{d}x^* + \frac{\lambda}{trace(S)}.$$

The integrated prediction variance can now be written as

$$\begin{aligned} \int \mathrm{Var}\{\hat{g}(x^*)\}\, \mathrm{d}x^* &= \int \left\{ \sum_{j=1}^{n} \sigma^2 S_j(x^*)^2 \right\} \mathrm{d}x^* \\ &= \int \left\{ \sigma^2 \sum_{j=1}^{n} \left[\Phi\left(\frac{\bar{x}_j - x^*}{h} \right) - \Phi\left(\frac{\bar{x}_{j-1} - x^*}{h} \right) \right]^2 \right\} \mathrm{d}x^*, \end{aligned}$$

and the trace of the smoothing matrix given by

$$\mathrm{trace}(S) = \sum_{j=1}^{n} \left[\Phi\left(\frac{\bar{x}_j - x_j}{h} \right) - \Phi\left(\frac{\bar{x}_{j-1} - x_j}{h} \right) \right].$$

We again use a quadrature approximation to the integrated variance using Legendre-Gauss

quadrature, see (4.31). As in the uniform kernel case, $m = 25$ abscissa values are used. The choice of λ is again restricted to $(0, 1)$ and we fix $\sigma^2 = 1$.

4.5.1 Designs for the Gaussian kernel

We were able to find optimal designs for larger numbers of runs, $n = 12$ and $n = 15$ see Table 4.5, when using the Gaussian kernel compared with using the uniform kernel.

For $n = 3$ and 5, the optimal designs in Table 4.9 exhibit similar patterns to those found under the uniform kernel (Table 4.4) for $h = 1$. However, for $h < 1$, we see from Tables 4.5-4.8 that the variance-trace trade off with varying λ is quite different when using the Gaussian kernel. The design points move away from the lower end of the interval [-1,1], that is the design range is $[\alpha, 1]$ where $\alpha > -1$ increases as λ decreases. In other words, the Gaussian kernel optimal designs are more clustered towards the upper end of the interval than the uniform kernel optimal designs for the same value of λ and h. Unlike the designs for the uniform kernel, the only symmetric designs obtained here are those including $x_1 = -1$ and $x_n = 1$. Note that non-symmetric designs have the property that the design $-\xi_n^*$, defined as a design with the points from ξ_n^* reflected in the line $x = 0$, is also optimal.

Tables 4.5–4.9 show that for fixed λ, reducing h results in designs covering a smaller section of the prediction interval. For example when $h = 1$, and $\lambda = 0.5$ (Table 4.9), the design covers the whole of the interval for both $n = 3$ and $n = 5$. Reducing h to 0.1 (Table 4.5) results in a design where points are only within the interval $[0.19, 1.00]$ for $n = 12$ and $[0.12, 1.00]$ for $n = 15$. For $h = 0.1$, it was computationally difficult to find designs for $\lambda < 0.3$. This is due to the clustered nature of the design and the increased number of points in the design. As in the uniform case the use of a small value of h implies that we wish to fit a more complex model and therefore require more design points than when h is large.

Numerical results suggest that the maximum value of the trace using the Gaussian kernel is bounded above by

$$\text{trace}(S) \leq (n-1)\left[2\Phi\left(\frac{2}{(n-1)h}\right) - 1\right],$$

with equality when the design points are equidistant. Therefore, as λ approaches 1, the choice of h does not affect the design for fixed n, as seen in Tables 4.6 and 4.8. Obviously, the choice of h does not affect the design when $\lambda = 0$, when we revert to minimising only

the variance, see Section 4.2 and hence all design points coalesce.

As λ decreases, the trace decreases, as for the uniform kernel. The variance also decreases, as expected. We also note that, as for the uniform kernel, the values of the variance and trace(S) were constant for different values of λ and the same optimal design, for example see Table 4.9 for $n = 3$.

For fixed λ and n, trace(S) decreases as h increases. This can be seen, for example, by comparing Table 4.6, when for $h = 0.2$, $n = 6$ and $\lambda = 0.7$ trace$(S) = 2.50$; and Table 4.8, when $h = 0.5$, $n = 6$ and $\lambda = 0.7$, trace$(S) = 1.55$. For large λ, the variance decreases as h increases. As λ decreases, the variances for different h have similar magnitude and finally for the smallest values of λ, the variance slightly increases with h. Note that when h is also fixed, the prediction variance decreases and the trace increases as n increases, as we would expect.

Figures 4.5–4.9 show how the prediction variance varies over the interval $[-1, 1]$ for examples of Gaussian kernel optimal designs. The variance has similar values for $\lambda = 0.999$ and $\lambda = 0.3$ for $h = 1$, see Figure 4.9, but is much smaller for $\lambda = 0.3$ when $h = 0.1, ..., 0.5$, see Figures 4.5–4.8. This agrees with the values of the integrated variance in Tables 4.5–4.8 and is due to the fact that the variance had more influence on the objective function when $\lambda = 0.3$ and $h = 0.1, ..., 0.5$. We also note that for $h = 0.1, ..., 0.3$, see Figures 4.5–4.7, respectively, when $\lambda = 0.3$, the points from ξ_n^* are at one end of the prediction interval and therefore the variance is only noticeably greater than zero on the section of this interval where the design points have clustered. The smoothing weights are never zero, unlike in the uniform kernel case. However, for these designs they are very small causing the variance to appear close to zero. The smooth Gaussian kernel function results in the prediction variance varying much more smoothly than when using the uniform kernel.

4.6 Robustness of prediction to choice of kernel function

The previous two sections have shown that the choice of kernel function can affect the optimal design. Therefore, we wish to assess how robust the prediction variance is to the choice of the kernel function. Specifically, we calculate the efficiency for prediction of designs found using the uniform kernel, calculated in Section 4.4.1.1, relative to designs from using the Gaussian kernel. We define the efficiency of a design under Criterion 4.4 using the Gaussian kernel as

	$n = 12$		
λ	ξ_n^*	$\int$ Var	tr(S)
0.999	±1.00 ±0.82 ±0.64 ±0.45 ±0.27 ±0.09	0.85	7.00
0.9	-0.71 -0.53 -0.38 -0.23 -0.08 0.07	0.64	6.18
	0.22 0.37 0.52 0.67 0.81 1.00		
0.8	-0.30 -0.15 -0.05 0.07 0.18 0.29	0.38	4.88
	0.40 0.52 0.63 0.75 0.85 1.00		
0.7	-0.08 0.05 0.13 0.23 0.32 0.41	0.26	4.13
	0.50 0.60 0.69 0.79 0.86 1.00		
0.5	0.19 0.29 0.34 0.42 0.48 0.56	0.15	3.16
	0.62 0.69 0.76 0.84 0.89 1.00		
0.3	0.39 0.47 0.50 0.56 0.61 0.66	0.08	2.41
	0.71 0.77 0.81 0.88 0.91 1.00		

	$n = 15$		
λ	ξ_n^*	$\int$ Var	tr(S)
0.999	±1.00 ±0.86 ±0.71 ±0.57 ±0.43 ±0.28 ±0.14 0.00	0.71	7.35
0.9	-0.84 -0.68 -0.56 -0.43 -0.31 -0.18 -0.05 0.07	0.60	6.84
	0.20 0.33 0.46 0.58 0.71 0.83 1.00		
0.8	-0.40 -0.27 -0.18 -0.08 0.01 0.11 0.20	0.36	5.26
	0.30 0.39 0.49 0.59 0.68 0.78 0.86 1.00		
0.7	-0.16 -0.05 0.01 0.10 0.18 0.26 0.33 0.41	0.25	4.50
	0.49 0.57 0.65 0.73 0.82 0.88 1.00		
0.5	0.12 0.21 0.26 0.32 0.38 0.44 0.50 0.56	0.14	3.43
	0.62 0.67 0.74 0.79 0.86 0.90 1.00		
0.3	0.33 0.41 0.43 0.49 0.52 0.57 0.61 0.66	0.08	2.62
	0.70 0.74 0.79 0.83 0.89 0.92 1.00		

Table 4.5: Optimal designs from Criterion 4.4 for predicting over the interval $[-1, 1]$ using the Gaussian kernel with $h = 0.1$. Design $-\xi_n^*$ is also optimal.

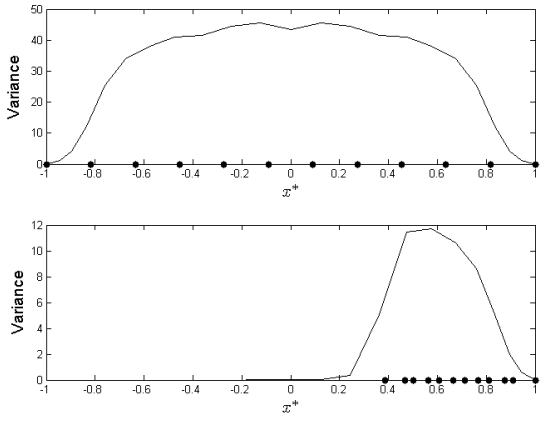

Figure 4.5: Prediction variance using the Criterion 4.4 optimal design (Table 4.6) using the Gaussian kernel for $n = 12$ and $h = 0.1$: $\lambda = 0.999$ and trace(S) = 7.00 (top), and $\lambda = 0.3$ and trace(S) = 2.41 (lower). Location of optimal design points are displayed on the x-axis. Note that the y-axis scales for the two plots are different.

	$n = 6$		
λ	ξ_n^*	$\int \mathrm{Var}$	$\mathrm{tr}(S)$
0.999	±1.00 ±0.60 ±0.20	0.86	3.41
0.9	±1.00 ±0.56 ±0.19	0.84	3.41
0.8	-0.63 -0.26 0.03 0.33 0.62 1.00	0.58	2.91
0.7	-0.36 -0.04 0.19 0.45 0.67 1.00	0.41	2.50
0.5	-0.02 0.23 0.38 0.58 0.73 1.00	0.23	1.94
0.3	0.23 0.43 0.52 0.69 0.78 1.00	0.13	1.49
0.1	0.50 0.64 0.68 0.80 0.84 1.00	0.05	0.98
0.05	0.61 0.72 0.74 0.85 0.87 1.00	0.03	0.77
	$n = 10$		
λ	ξ_n^*	$\int \mathrm{Var}$	$\mathrm{tr}(S)$
0.999	±1.00 ±0.78 ±0.56 ±0.33 ±0.11	0.54	3.79
0.9	±1.00 ±0.73 ±0.53 ±0.32 ±0.11	0.53	3.77
0.8	±1.00 ±0.70 ±0.53 ±0.31 ±0.11	0.53	3.77
0.7	-0.61 -0.38 -0.24 -0.06 0.10 0.27 0.43 0.61 0.75 1.00	0.35	3.10
0.5	-0.22 -0.03 0.06 0.20 0.32 0.45 0.56 0.71 0.79 1.00	0.20	2.37
0.3	0.08 0.23 0.28 0.40 0.47 0.58 0.66 0.78 0.83 1.00	0.11	1.81
0.1	0.41 0.51 0.53 0.62 0.65 0.73 0.76 0.86 0.88 1.00	0.04	1.17
0.05	0.54 0.62 0.63 0.70 0.72 0.79 0.81 0.89 0.90 1.00	0.03	0.92

Table 4.6: Optimal designs from Criterion 4.4 for predicting over the interval $[-1, 1]$ using the Gaussian kernel with $h = 0.2$. Design $-\xi_n^*$ is also optimal.

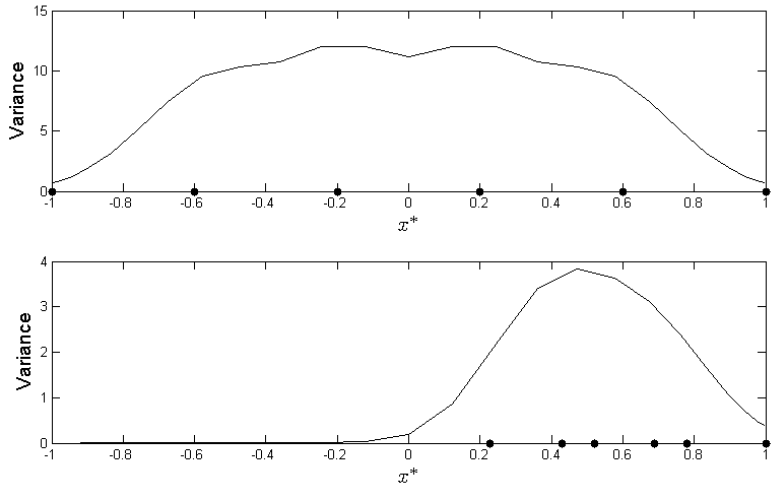

Figure 4.6: Prediction variance using the Criterion 4.4 optimal design (Table 4.6) using the Gaussian kernel for $n = 6$ and $h = 0.2$: $\lambda = 0.999$ and $\mathrm{trace}(S) = 3.41$ (top), and $\lambda = 0.3$ and $\mathrm{trace}(S) = 1.49$ (lower). Location of optimal design points are displayed on the x-axis. Note that the y-axis scales for the two plots are different.

	$n = 5$		
λ	ξ_n^*	$\int$ Var	tr(S)
0.999	±1.00 ±0.50 0.00	0.71	2.38
0.9	±1.00 ±0.46 0.00	0.70	2.38
0.8	±1.00 ±0.43 0.00	0.69	2.37
0.7	-0.68 -0.20 0.15 0.50 1.00	0.50	2.05
0.5	-0.26 0.12 0.36 0.61 1.00	0.28	1.59
0.3	0.05 0.34 0.51 0.69 1.00	0.16	1.22
0.1	0.39 0.58 0.68 0.79 1.00	0.06	0.80
0.05	0.52 0.67 0.75 0.83 1.00	0.04	0.63
	$n = 7$		
λ	ξ_n^*	$\int$ Var	tr(S)
0.999	±1.00 ±0.66 ±0.33 0.00	0.52	2.53
0.9	±1.00 ±0.62 ±0.32 0.00	0.51	2.53
0.8	±1.00 ±0.59 ±0.31 0.00	0.50	2.52
0.7	±1.00 ±0.56 ±0.31 0.00	0.50	2.51
0.5	-0.41 -0.11 0.06 0.28 0.50 0.67 1.00	0.25	1.82
0.3	-0.07 0.17 0.28 0.45 0.62 0.73 1.00	0.14	1.39
0.1	0.31 0.47 0.53 0.64 0.75 0.82 1.00	0.06	0.91
0.05	0.46 0.58 0.63 0.71 0.80 0.85 1.00	0.03	0.71

Table 4.7: Optimal designs from Criterion 4.4 for predicting over the interval $[-1, 1]$ using the Gaussian kernel with $h = 0.3$. Design $-\xi_n^*$ is also optimal.

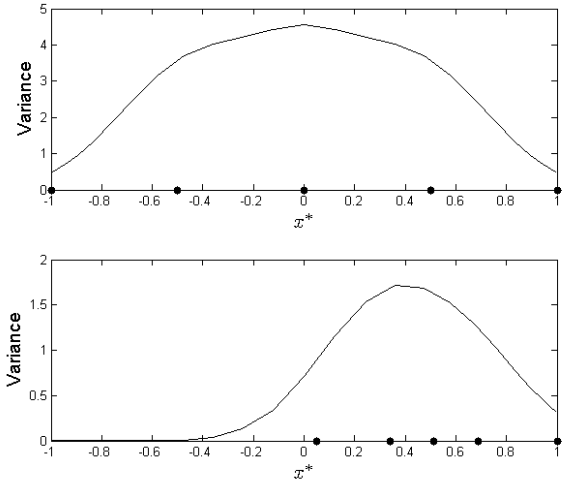

Figure 4.7: Prediction variance using the Criterion 4.4 optimal design (Table 4.7) using the Gaussian kernel for $n = 5$ and $h = 0.3$: $\lambda = 0.999$ and trace(S) = 2.38 (top), and $\lambda = 0.3$ and trace(S) = 1.22 (lower). Location of optimal design points are displayed on the x-axis. Note that the y-axis scales for the two plots are different.

	$n = 4$		
λ	ξ_n^*	$\int$ Var	tr(S)
0.999	±1.00 ±0.33	0.53	1.49
0.9	±1.00 ±0.30	0.52	1.48
0.8	±1.00 ±0.27	0.51	1.48
0.7	±1.00 ±0.24	0.51	1.47
0.5	-0.69 -0.01 0.29 1.00	0.37	1.26
0.3	-0.25 0.27 0.43 1.00	0.20	0.96
0.1	0.19 0.55 0.59 1.00	0.08	0.63
0.05	0.36 0.65 0.67 1.00	0.05	0.50
	$n = 6$		
λ	ξ_n^*	$\int$ Var	tr(S)
0.999	±1.00±0.60 ±0.20	0.35	1.55
0.9	±1.00 ±0.55 ±0.19	0.34	1.55
0.8	±1.00 ±0.52 ±0.19	0.34	1.55
0.7	±1.00 ±0.49 ±0.19	0.33	1.55
0.5	±1.00 ±0.43 ±0.22	0.33	1.54
0.3	-0.45 -0.04 0.07 0.43 0.55 1.00	0.18	1.13
0.1	0.07 0.34 0.38 0.64 0.68 1.00	0.07	0.73
0.05	0.27 0.48 0.50 0.72 0.74 1.00	0.04	0.58

Table 4.8: Optimal designs from Criterion 4.4 for predicting over the interval $[-1, 1]$ using the Gaussian kernel with $h = 0.5$. Design $-\xi_n^*$ is also optimal.

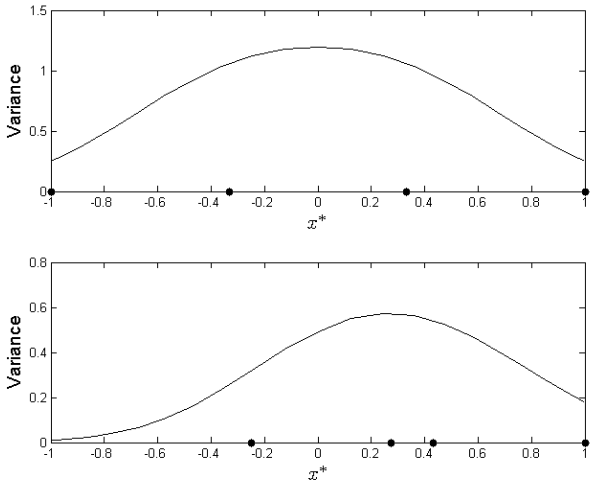

Figure 4.8: Prediction variance using the Criterion 4.4 optimal design (Table 4.6) using the Gaussian kernel for $n = 4$ and $h = 0.5$: $\lambda = 0.999$ and trace(S) = 1.49 (top), and $\lambda = 0.3$ and trace(S) = 0.96 (lower). Location of optimal design points are displayed on the x-axis. Note that the y-axis scales for the two plots are different.

	$n = 3$		
λ	ξ_n^*	$\int$ Var	tr(S)
0.999	±1.00 0.00	0.31	0.77
0.95	±1.00 0.00	0.31	0.77
0.9	±1.00 0.00	0.31	0.77
0.8	±1.00 0.00	0.31	0.77
0.7	±1.00 0.00	0.31	0.77
0.5	±1.00 0.00	0.31	0.77
0.3	-0.95 0.02 1.00	0.30	0.75
0.1	-0.18 0.40 1.00	0.11	0.46
0.05	0.07 0.53 1.00	0.06	0.37
	$n = 5$		
λ	ξ_n^*	$\int$ Var	tr(S)
0.999	±1.00 ±0.50 0.00	0.18	0.79
0.9	±1.00 ±0.46 0.00	0.18	0.79
0.8	±1.00 ±0.44 0.00	0.18	0.79
0.7	±1.00 ±0.41 0.00	0.17	0.79
0.5	±1.00 ±0.39 0.00	0.17	0.79
0.3	±1.00 ±0.34 0.00	0.17	0.79
0.1	-0.46 0.01 0.24 0.49 1.00	0.09	0.58
0.05	-0.12 0.23 0.41 0.60 1.00	0.05	0.45

Table 4.9: Optimal designs from Criterion 4.4 for predicting over the interval $[-1, 1]$ using the Gaussian kernel with $h = 1$. Design $-\xi_n^*$ is also optimal.

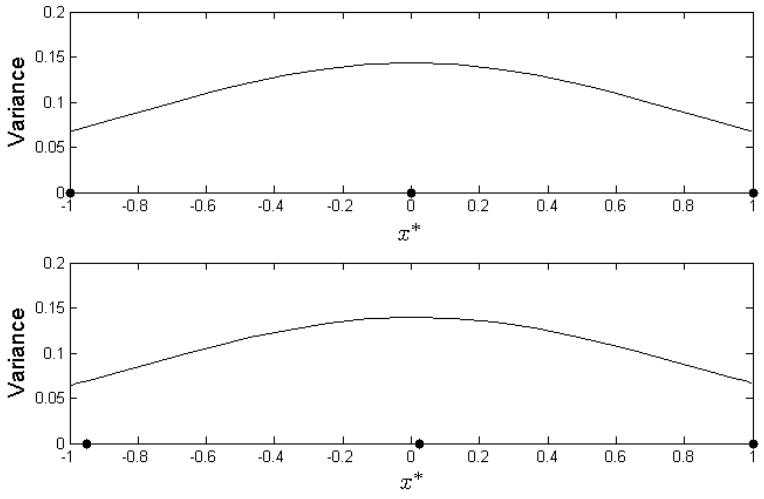

Figure 4.9: Prediction variance using the Criterion 4.4 optimal design (Table 4.6) using the Gaussian kernel for $n = 3$ and $h = 1$: $\lambda = 0.999$ and trace(S) = 0.77 (top), and $\lambda = 0.3$ and trace(S) = 0.75 (lower). Location of optimal design points are displayed on the x-axis.

$$\text{Eff} = \frac{|\Psi_G(\xi^G)|}{|\Psi_G(\xi^u)|}, \tag{4.33}$$

where $\Psi_G(\xi^u)$ and $\Psi_G(\xi^G)$ are the values of the objective function, calculated with the Gaussian kernel, using (i) ξ^u, the optimal design under Criterion 4.4 using the uniform kernel and (ii) ξ^G, the optimal design under Criterion 4.4 using the Gaussian kernel.

Tables 4.10 and 4.11 show the efficiencies for two different scenarios. Table 4.10 gives efficiencies for $h = 1$ and $n = 3$ for optimal designs for both the uniform and Gaussian kernels found in Tables 4.4 and 4.9 respectively. In this case we would expect to fit a relatively simple model. However, the second investigation, see Table 4.11, includes efficiencies for $h = 0.3$ and $n = 7$, where the optimal designs for uniform and Gaussian kernels are found in Tables 4.2 and 4.7, respectively. In the latter case, the fitted model is expected to be more complex.

λ	ξ^uor ξ^G	Optimal Design	Eff	trace$_u$	trace$_G$
0.999	ξ^u	$\pm$1.00 0.00	1	1	0.766
	ξ^G	$\pm$1.00 0.00			
0.9	ξ^u	$\pm$1.00 0.00	1	1	0.766
	ξ^G	$\pm$1.00 0.00			
0.8	ξ^u	$\pm$1.00 0.00	1	1	0.766
	ξ^G	$\pm$1.00 0.00			
0.7	ξ^u	$\pm$1.00 0.00	1	1	0.766
	ξ^G	$\pm$1.00 0.00			
0.5	ξ^u	$\pm$1.00 0.00	1	1	0.766
	ξ^G	$\pm$1.00 0.00			
0.3	ξ^u	-0.62 0.08 0.97	0.968	0.79	0.749
	ξ^G	-0.95 0.02 1.00			
0.1	ξ^u	0.06 0.48 1.00	0.952	0.47	0.463
	ξ^G	-0.18 0.40 1.00			
0.05	ξ^u	0.25 0.62 1.00	0.956	0.37	0.366
	ξ^G	0.07 0.53 1.00			

Table 4.10: Efficiencies of uniform kernel optimal designs for prediction with the Gaussian kernel for $h = 1$ and $n = 3$. The trace for the uniform kernel optimal design and the Gaussian kernel optimal design are given by trace$_u$ and trace$_G$ respectively.

Table 4.10 shows that the uniform kernel optimal design performs very well when evaluated using the Gaussian kernel, with efficiencies all greater than 0.95. For $\lambda \geq 0.3$ the efficiency is 1 since the design $\xi_n = \{-1, 0, 1\}$ is optimal using both the uniform and Gaussian kernels.

Table 4.11 also shows that for $h = 0.3$ the uniform kernel optimal design performs very

λ	ξ^u or ξ^G	Optimal Design	Eff	trace_u	trace_G
0.999	ξ^u	±1.00 ±0.59 ±0.43 0.00	0.983	3.33	2.53
	ξ^G	±1.00 ±0.66 ±0.33 0.00			
0.9	ξ^u	±1.00 ±0.59 ±0.43 0.00	0.985	3.33	2.53
	ξ^G	±1.00 ±0.62 ±0.32 0.00			
0.8	ξ^u	-0.83 -0.34 -0.17 0.16 0.51 0.63 1.00	0.918	3.05	2.52
	ξ^G	±1.00 ±0.59 ±0.31 0.00			
0.7	ξ^u	-0.71 -0.32 -0.15 0.10 0.42 0.57 0.89	0.962	2.67	2.51
	ξ^G	±1.00 ±0.56 ±0.31 0.00			
0.5	ξ^u	-0.62 -0.33 -0.20 0 0.20 0.33 0.62	0.969	2.06	1.82
	ξ^G	-0.41 -0.11 0.06 0.28 0.50 0.67 1.00			
0.3	ξ^u	-0.47 -0.23 -0.19 0.02 0.18 0.27 0.49	0.998	1.60	1.39
	ξ^G	-0.07 0.17 0.28 0.45 0.62 0.73 1.00			
0.1	ξ^u	-0.10 0.06 0.11 0.19 0.29 0.34 0.48	0.939	0.98	0.91
	ξ^G	0.31 0.47 0.53 0.64 0.75 0.82 1.00			
0.05	ξ^u	-0.07 0.06 0.11 0.18 0.25 0.28 0.40	0.932	0.78	0.71
	ξ^G	0.46 0.58 0.63 0.71 0.80 0.85 1.00			

Table 4.11: Efficiencies of uniform kernel optimal designs for prediction with the Gaussian kernel for $h = 0.3$ and $n = 7$. The trace for the uniform kernel optimal design and the Gaussian kernel optimal design are given by trace_u and trace_G respectively.

well when evaluated using the Gaussian kernel. Interestingly for many values of λ, the two designs cover different sections of the interval $[-1, 1]$. For example, when $\lambda = 0.3$, the uniform kernel optimal design covers the centre of the interval whereas the Gaussian kernel optimal design clusters at one end of the interval, yet performs almost as well.

4.7 Concluding remarks

In this chapter we have developed a new compound criterion: minimising a weighted sum of the integrated prediction variance and the inverse trace of the smoothing matrix. This enabled designs to be tailored to different complexities of fitted models. Optimal designs were found for both the uniform and Gaussian kernel functions using both analytic and numerical methods. These designs were critically assessed by investigating the robustness of the prediction to the choice of kernel function.

We investigated designs for different compromises, via the parameter λ, between prediction variance and model complexity. Larger values of λ, placing more emphasis on model complexity, resulted in designs with points spread more evenly across the design region for both the uniform and Gaussian kernels. However, decreasing λ had different effects for the two kernels. For the uniform kernel, design points concentrated around the centre of

the design region; for the Gaussian kernel, points concentrated at one end of the region.

By investigating designs for the uniform and Gaussian kernels, which have very different forms, we were able to assess the sensitivity of design performance to choice of kernel. For both bandwidths investigated, the performance of designs was robust to the choice of kernel function.

Appendix

Result 1:

For $a_i > 0$ and $\sum_{i=1}^{n} a_i = c$, for some constant c, $\sum_{i=1}^{n} a_i^2$ is minimised when $a_i = c/n$ for all $1 \leq i, j \leq n$.

Proof. We can write

$$a_n = c - \sum_{i=1}^{n-1} a_i$$

and hence

$$\sum_{i=1}^{n} a_i^2 = a_1^2 + ... + a_{n-1}^2 + \left(c - \sum_{i=1}^{n-1} a_i \right)^2 .$$

Differentiating with respect to a_k $(k \leq n-1)$ gives

$$\frac{\partial(\sum_{i=1}^{n} a_i^2)}{\partial a_k} = 2a_k - 2 \left[c - \sum_{i=1}^{n-1} a_i \right].$$

Equating to zero gives

$$\begin{aligned} a_k &= c - \sum_{i=1}^{n-1} a_i \\ &= a_n. \end{aligned}$$

This holds for all $k = 1, \ldots, n-1$. To establish this solution is a minimum, we check the

second derivative

$$\frac{\partial^2(\sum_{i=1}^n a_i^2)}{\partial a_k^2} = 2 + 2 > 0.$$

Hence, $a_1, \ldots, a_n = c/n$ minimises $\sum_{i=1}^n a_i^2$ subject to $\sum_{i=1}^n a_i = c$

□

Chapter 5

Designed experiments and functional linear models

5.1 Introduction

Functional data arise from experiments for which multiple observations, assumed to come from a smooth function, are measured on each unit to which a treatment has been applied. These functions potentially vary between treatments and are often too complex to be modelled using any obvious parametric form; see, for example, Faraway (1997), Shen and Faraway (2004) and Shen and Xu (2006). Examples of experiments that produce functional data can be found in chemistry, biology, tribology and engineering.

For some experiments, longitudinal data analysis methods (Diggle et al., 2002) may seem sensible for analysing functional data. However, longitudinal datasets usually have fewer observations per run than functional data and often a parametric model can be assumed for the responses from each run. Functional data analysis has its place providing methods which may work when longitudinal methods are not appropriate (Faraway, 1997). Typically, functional data has larger numbers of measurements per run and the functional response is estimated using nonparametric methods.

There are several different types of functional data: the response variable depends on indexing variable, t; one or more of the covariates depends on t; or both the response and covariates depend on t (Ramsay and Silverman, 2005, p. 218). Throughout this work, we consider the first case. Functional data with functional covariates and scalar response has been considered by authors such as Ramsay and Silverman (2005, ch. 12), Cardot et al. (1999) and Cardot et al. (2004). Examples where both the response and covariates

vary with time were considered by West, Harrison and Migon (1985) who assumed the regression coefficients had autoregressive structure, referring to the model as a dynamic generalised linear model.

The purpose of this chapter is to develop methodology for designing experiments for functional data whose aim is to discriminate between two functional linear models. In the first half of the chapter, we review the modelling and inferential work of Faraway (1997), Shen and Faraway (2004) and Shen and Xu (2006) on tests for choosing between two nested linear models. The methods are illustrated on two examples. In the second half of this chapter, we develop a T-criterion and find T-optimal designs for discriminating between two functional linear models. The designs are then critically assessed through two simulation studies.

5.2 Examples of functional data

In the first half of this chapter we consider two examples: (i) a simple simulated experiment with only one factor; (ii) the tribology example introduced in Chapter 1 which has six factors.

5.2.1 Simulated experiment

The first example is an n-run experiment to investigate two treatments, A and B, whose functional responses are described by different models. We assume that a runs of the experiment have treatment A and that the response follows

$$y_A(t) = \alpha_{0A} + \alpha_{1A}t + \epsilon(t), \tag{5.1}$$

where α_{0A} and α_{1A} are model parameters, $\epsilon(t) \sim N(0, \sigma^2)$ for all $t \in \mathcal{I} \subset \mathbb{R}$ and $\epsilon(t), \epsilon(s)$ are independent for all $t, s \in \mathcal{I}$ such that $t \neq s$.

Each of the remaining $b = n - a$ runs of the experiment has treatment B and produces a response following the model

$$y_B(t) = \alpha_{0B} + \epsilon(t), \tag{5.2}$$

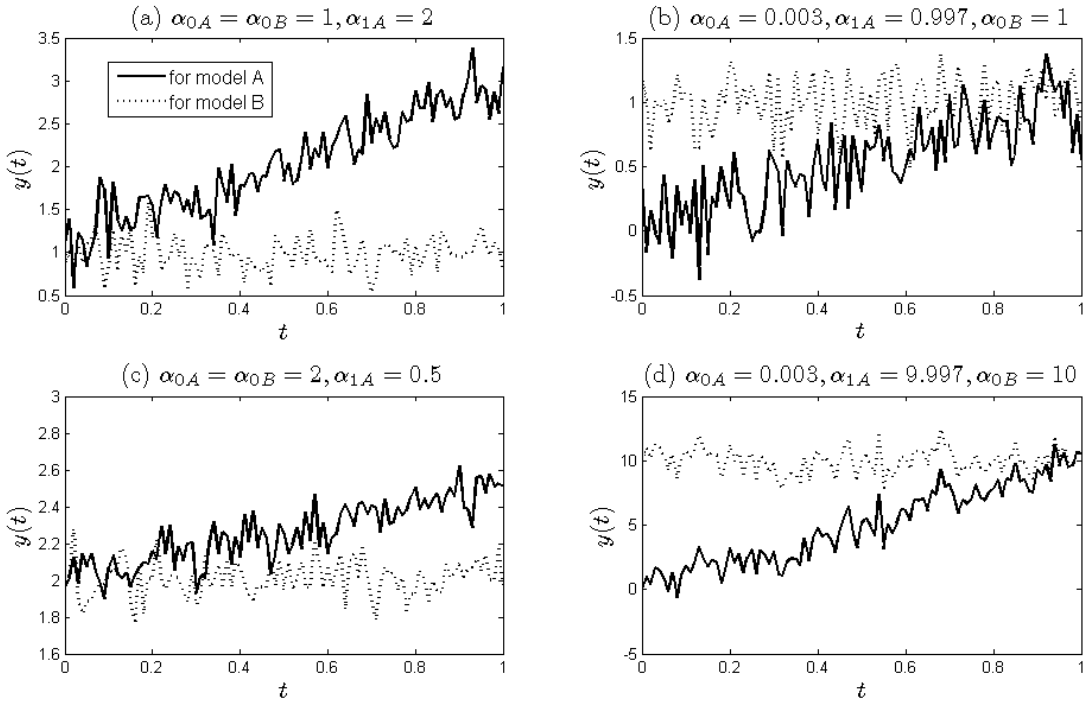

Figure 5.1: Example 1: Simulated data from the application of treatments A and B (see (5.1) and (5.2)) with various parameter values

where α_{0B} is a model parameter and $\epsilon(t)$ is defined as before.

Figure 5.1 shows simulated responses from each treatment for a selection of different parameter values, all with $\sigma^2 = 0.04$. We see that in Figure 5.1 (a), (b) and (d), the responses from treatments A and B differ much more than in Figure 5.1 (c); hence model discrimination would be much harder in the latter case.

5.2.2 Tribology experiment

We use the tribology experiment, introduced in Chapter 1, as a motivating and illustrative example. Recall that data were recorded from a wear study conducted to assess which factors affected the rate of wear of a pin and disc assembly for a given lubricant. The experiment involved 16 runs, each with a different combination of values of the six factors: disc material, pin material, addition of soot, level of oxidation, addition of H_2SO_4 and level of moisture.

The experiment used an unreplicated 2^{6-2} fractional factorial design with 16 runs and defining relation I=ACEF=ABDE=BCDF. Hence pairs of two factor interactions were aliased together. For each of the 16 runs, the functional response 'wear' was measured over

a time index. Unfortunately, for two functional runs there was no data available resulting in a 14 run experiment. Therefore, the realised design has a partial aliasing scheme and only 14 effects, at most, can be estimated. As stated in Chapter 1, observations during a burn in and a period at the end of each run were removed after consultation with the engineers.

The aim of the experiment is to predict the value of the response, that is, the profile over the interval [500, 2400] of the combined wear on the pin and disc for a given combination of values of the six factors. In this experiment the measured response was the the combined wear on the pin and the disc, measured by a Linear Variable Displacement Transformer at a large number of equally spaced discrete time points (referred to as the time index).

Figure 5.2 displays responses from two runs of the tribology experiment. Notice that there is much more variation between runs than within runs because the form of the functional response changes with treatment.

5.3 Definitions and notation

We formally introduce functional data following the notation of Faraway (1997). Note that, throughout this chapter, p denotes the total number of terms in the model, including the intercept. Suppose that the ith run of the experiment involves taking observations on a smooth function

$$y_i(t) = g_i(t) + \epsilon_i(t), \tag{5.3}$$

for $i = 1, \ldots, n$ where $\epsilon_i(t)$ is a realisation of a stochastic process with mean zero and covariance function $\gamma(s, t)$ with s, t belonging to an interval, $\mathcal{I} \subset \mathbb{R}$. When the experiment is performed, we take n_i observations on each function at values of an index variable, such as time, and express the jth observation from the ith run, taken at t_{ij}, as

$$y_i(t_{ij}) = g_i(t_{ij}) + \epsilon_{ij}, \tag{5.4}$$

for $j = 1, \ldots, n_i$. Note that the dataset may satisfy $n_i = m$ for all i, i.e. the same number m of observations is recorded for each run of the experiment. This often occurs in practice and is now assumed in the rest of the chapter. Even if $n_i \neq m$ for all i, it is often the

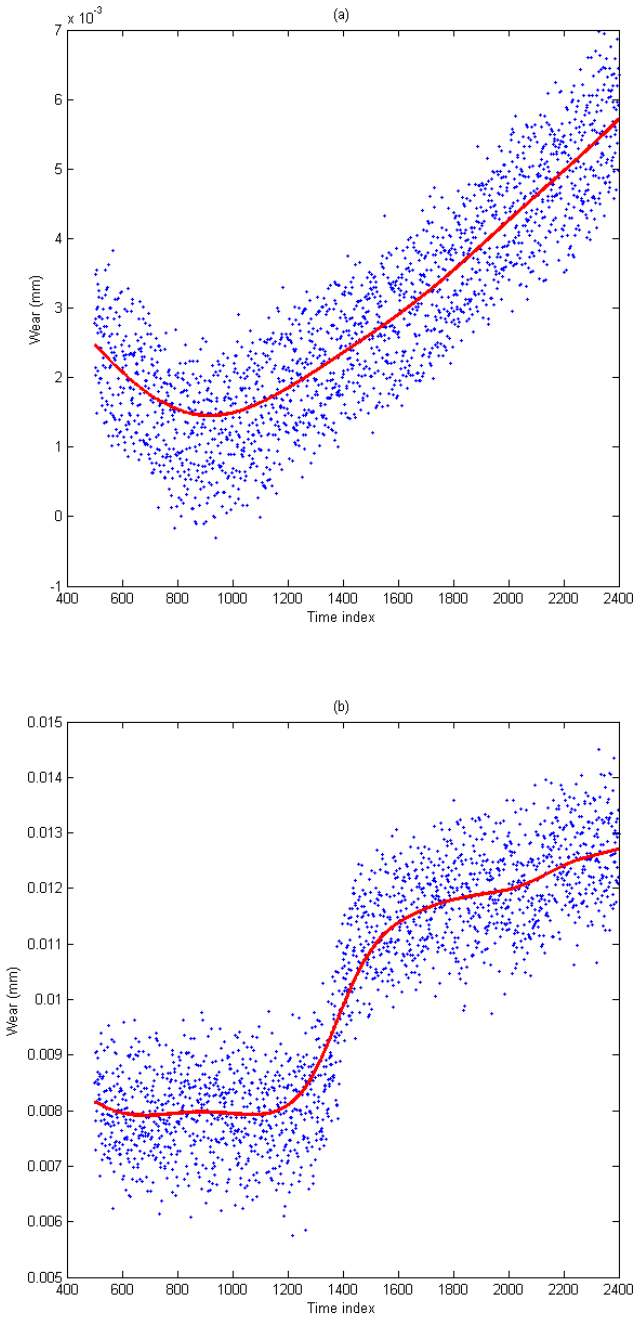

Figure 5.2: Data from run 2, plot (a), and run 19, plot (b), of the tribology experiment

case that the observations from each run can be smoothed or interpolated to obtain m 'observations'.

In the remainder of this section, we consider methods of approximating functional responses and then fit functional linear models to explain the variation due to the different treatments.

5.3.1 Approximating the functional response

We now wish to reconstruct $g_i(t)$ using the discrete measurements from (5.4). In order to do so, we require some form of smoothing. If there is little or no within-run error, then we could simply interpolate to approximate $g_i(t)$ (Faraway, 1997). However this would not be advisable when measurements are made with a degree of error.

There are two clear smoothing methods pointed out by Faraway (1997): (i) smooth each run, y_i, separately without reference to the type of model fitted, using methods such as those described in Chapters 3 and 4, (ii) use cross-validation to determine the amount of smoothing to be done, through a roughness penalty; see Ramsay and Dalzell (1991). Note that cross-validation performs poorly when the errors are correlated (Faraway, 1997). Other methods should be used in this case.

Smoothing methods used to reconstruct $g_i(t)$ from m observations were discussed in Chapters 3 and 4. In those chapters the method of local smoothing to reconstruct the function $g_i(t)$ from a finite number of design points which were chosen to minimise the average variance of $\hat{g}_i(t)$ over the prediction interval.

5.3.2 Functional linear model

The functional linear model with constant covariates is defined by Faraway (1997) and Ramsay and Silverman (2005, p. 235) as

$$\boldsymbol{Y}(t) = X\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t), \tag{5.5}$$

where $\boldsymbol{\beta}(t) = (\beta_0(t), \ldots, \beta_{p-1}(t))^T$ is a vector of functions depending on t, X is an $n \times p$ model matrix, $\boldsymbol{Y}(t) = (y_1(t), \ldots, y_n(t))^T$ is a vector of functional responses and $\boldsymbol{\epsilon} =$

$(\epsilon_1(t), \ldots, \epsilon_n(t))^T$, is a vector of error functions with $\epsilon_i(t)$ defined as in (5.3). That is, we assume the functional response depends linearly on p unknown functional parameters.

5.4 Fitting functional linear models

In this section, we consider fitting a functional linear model (5.5). We consider using (i) pointwise methods and (ii) regularised basis expansions.

We assume the data from the ith run follows the model

$$y_i(t) = \boldsymbol{x}_i^T \boldsymbol{\beta}(t) + \epsilon_i(t),$$

where $\boldsymbol{x}_i^T = (x_{i0}, \ldots, x_{i,p-1})$ is the ith row of the matrix X, and $\boldsymbol{\beta}(t)$ and $\epsilon_i(t)$ are defined in Section 5.3.2.

The extension of the least squares principle to the functional case is described by Ramsay and Silverman (2005, p. 236). We wish to find estimators of $\boldsymbol{\beta}(t)$ to minimise

$$\sum_{i=1}^{n} \int \left[y_i(t) - \left(\boldsymbol{x}_i^T \boldsymbol{\beta}(t) \right) \right]^2 \mathrm{d}t, \tag{5.6}$$

where in this expression, and throughout this chapter, the integral is evaluated over the interval $\mathcal{I}$. This integral is mathematically intractable and therefore must be approximated.

5.4.1 Pointwise Methods

The least squares estimator for $\boldsymbol{\beta}(t)$ for each t can be calculated pointwise as

$$\hat{\boldsymbol{\beta}}(t) = \left(X^T X \right)^{-1} X^T \boldsymbol{Y}(t). \tag{5.7}$$

Values of (5.7) can then be interpolated across t to provide an approximate solution to (5.6). Therefore if we can evaluate or approximate $\boldsymbol{Y}(t)$ for given t, then $\hat{\boldsymbol{\beta}}$ can be calculated. In addition, we can make predictions $\hat{\boldsymbol{Y}}(t) = H\boldsymbol{Y}(t)$, where $H = X(X^T X)^{-1} X^T$

directly from the approximation of $\boldsymbol{Y}(t)$.

However, $\hat{\boldsymbol{\beta}}(t)$ may be a very wiggly function of t due to the noise in the data. Hence, we may desire a fitting method that places smoothness constraints on $\hat{\boldsymbol{\beta}}(t)$.

5.4.2 Fitting functional linear models with regularised basis expansions

The use of regularised basis expansions allows control of the smoothness of $\hat{\boldsymbol{\beta}}(t)$ whilst incorporating the level of detail the data requires (Ramsay and Silverman, 2005, p. 236). In comparison to the pointwise method, which places no constraint on the parameter $\boldsymbol{\beta}(t)$, this method uses a roughness penalty to control the smoothness of $\hat{\boldsymbol{\beta}}(t)$.

The data, $\boldsymbol{Y}(t)$, can be represented by the product of a basis expansion and coefficient matrix, for instance, using a B-spline basis (see, for example Eubank (1999, ch. 6)):

$$\boldsymbol{Y}(t) = C\boldsymbol{\phi}(t), \tag{5.8}$$

where $\boldsymbol{Y}(t)$ contains n observed response functions and C is a $n \times K_y$ matrix of coefficients of the expansion of $\boldsymbol{Y}(t)$ in its ith row for $i = 1, \ldots, n$ runs. Here K_y is the number of basis functions chosen to represent the response, and $\boldsymbol{\phi}(t)$ is the K_y-vector containing the linearly independent basis functions.

The parameter vector $\boldsymbol{\beta}(t)$, of length p, can also be expressed in terms of a basis vector $\boldsymbol{\theta}(t)$, of length K_β, and a $p \times K_\beta$ coefficient matrix M, giving $\boldsymbol{\beta}(t) = M\boldsymbol{\theta}(t)$. We define K_β as the number of basis functions chosen to represent $\boldsymbol{\beta}$.

The roughness penalty is defined for $\boldsymbol{\beta}$ as

$$PEN_L(\boldsymbol{\beta}) = \int \left[L\boldsymbol{\beta}(t)\right]^T \left[L\boldsymbol{\beta}(t)\right] \mathrm{d}t, \tag{5.9}$$

where L is a linear differential operator, that is $L\boldsymbol{\beta}$ is a vector containing derivatives of $\boldsymbol{\beta}(t)$ of a given order. Note that a common penalty is the integrated squared second derivative given by

$$PEN_2(\boldsymbol{\beta}) = \int \left\{D^2[\boldsymbol{\beta}(t)]\right\}^T \left\{D^2[\boldsymbol{\beta}(t)]\right\} \mathrm{d}t, \tag{5.10}$$

where $D^2[x(t)]$ denotes the second derivative of $x(t)$ with respect to t. The sum of squared errors, ignoring the roughness penalty, in the functional case is

$$SSE(y|\boldsymbol{\beta}) = \int [\boldsymbol{Y}(t) - X\boldsymbol{\beta}(t)]^T [\boldsymbol{Y}(t) - X\boldsymbol{\beta}(t)] \mathrm{d}t.$$

The basis expansion for the sum of squared errors can be written as

$$SSE_B(y|\boldsymbol{\beta}) = \int \left[C\boldsymbol{\phi}(t) - XM\boldsymbol{\theta}(t)\right]^T \left[C\boldsymbol{\phi}(t) - XM\boldsymbol{\theta}(t)\right] \mathrm{d}t.$$

The penalised least squares criterion is to minimise

$$PENSSE(y|\boldsymbol{\beta}) = SSE_B(y|\boldsymbol{\beta}) + \mu \int [LM\boldsymbol{\theta}(t)]^T [LM\boldsymbol{\theta}(t)] \mathrm{d}t. \tag{5.11}$$

The scalar μ controls the degree of smoothing applied via the penalty. It is possible to re-write (5.11) in terms of Kronecker products (Ramsay and Silverman, 2005, p. 238-239), giving the exact solution for $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}} = \hat{M}\boldsymbol{\theta}(t), \tag{5.12}$$

where

$$\mathrm{vec}(M) = \left[J_{\boldsymbol{\theta}\boldsymbol{\theta}} \otimes (X^T X) + R \otimes \mu I\right]^{-1} \mathrm{vec}(X^T C J_{\boldsymbol{\phi}\boldsymbol{\theta}}),$$

with $J_{\boldsymbol{\theta}\boldsymbol{\theta}} = \int \boldsymbol{\theta}(t)[\boldsymbol{\theta}(t)]^T \mathrm{d}t$, $J_{\boldsymbol{\phi}\boldsymbol{\theta}} = \int \boldsymbol{\phi}(t)[\boldsymbol{\theta}(t)]^T \mathrm{d}t$ and $R = \int L\boldsymbol{\theta}(t)[L\boldsymbol{\theta}(t)]^T \mathrm{d}t$, and recall-

ing that $\boldsymbol{Y}(t) = C\boldsymbol{\phi}(t)$ from (5.8).

An optimal value of μ, the smoothing parameter, can be calculated using the cross-validated integrated squared error

$$CVISE = \sum_{i=1}^{n} \int \left[y_i(t) - \hat{y}^{(-i)}(t)\right]^2 \mathrm{d}t, \tag{5.13}$$

where $\hat{y}^{(-i)}(t)$ is the predicted value for $y_i(t)$ when $y_i(t)$ is excluded from the estimation of $\boldsymbol{\beta}$, see Ramsay, Hooker and Graves (2009, p. 153-154).

5.5 Inferential methods for model comparison

In this section, we begin by discussing three methods for comparing two rival functional linear models. We then demonstrate two of the methods by applying them to a simulated example, in Section 5.5.2, and to data from the tribology experiment, in Section 5.5.3.

5.5.1 Methods of comparing two models

(i) **Pointwise methods**

When we have fitted a functional linear model, for example using the methods in Section 5.4.2, we can use the optimal values for $\hat{\boldsymbol{\beta}}(t)$ to test the null hypothesis H_0: $\boldsymbol{Y}(t) = \beta_0(t)\mathbf{1}_n + \boldsymbol{\varepsilon}(t)$ against the alternative hypothesis H_1: $\boldsymbol{Y}(t) = X\boldsymbol{\beta}(t) + \boldsymbol{\varepsilon}(t)$ by calculating the corresponding pointwise F-ratio for every t. The pointwise F-ratio can be calculated pointwise over t as

$$\text{Fratio}(t) = \frac{MSR(t)}{MSE(t)}, \tag{5.14}$$

with

$$\begin{aligned} MSE(t) &= \frac{SSE(t)}{n-p} \\ &= \frac{\sum_{i=1}^{n}\left[y_i(t)-\hat{y}_i(t)\right]^2}{n-p}, \end{aligned}$$

and

$$\begin{aligned} MSR(t) &= \frac{SSY(t)-SSE(t)}{p-1} \\ &= \frac{\sum_{i=1}^{n}\left[y_i(t)-\hat{\beta}_0(t)\right]^2-\sum_{i=1}^{n}\left[y_i(t)-\hat{y}_i(t)\right]^2}{p-1}, \end{aligned}$$

with $\hat{\beta}_0(t)$ the intercept in the functional linear model. The number of degrees of freedom for error, $df(\text{error}) = n-p$, is the total number of runs less the number of independent variables in the model. The number of degrees of freedom for regression, $df(\text{error}) = p-1$, is the difference in the numbers of degrees of freedom for error for the two models being compared.

Analogous to usual linear model theory (see, for example Draper and Smith (1998, ch. 6)), this testing procedure can be generalised for pointwise comparison of any two nested functional linear models.

There are some caveats to only using pointwise tests. Firstly there is the problem with making multiple comparisons. Applying Bonferroni corrections to the significance level to account for this would compromise the power due to the within run correlation (Shen and Xu, 2006). Secondly, Fan and Lin (1998) remarked that the correlation between two neighbouring observations for a given run should not be ignored in the analysis. In order to remove correlation from stationary data, these authors applied Fourier or wavelet transformations. A final disadvantage of pointwise hypothesis testing is that it does not give an overall assessment of significance for the difference between the functional linear models. There are instances where two models may be falsely shown to differ significantly at a point, when they do not differ significantly over the whole interval (Ramsay and Silverman, 2005, p. 228).

(ii) **Multivariate-based methods**

Faraway (1997) and Shen and Faraway (2004) argued that if $y_i(t)$ is measured on an equally

spaced grid of m values of t, then methods of multivariate analysis may be used provided predictions are only required at the same values of t for each function. To compare two nested multivariate linear models, the likelihood ratio test could be used, where the test statistic depends on the log ratio of eigenvalues of the estimated covariance matrix from each model. However, this sequence of ratios need not tend to 0 as $m \to \infty$. Hence, the test statistic may be dominated by terms (ratios) which only nominally contribute to the variation in the data. Faraway (1997) and Shen and Faraway (2004) concluded that the likelihood ratio statistic is only suitable for small m, which is unusual for functional datasets. For this reason, we will not use this test in the first half of this chapter.

(iii) **Functional F-test**

Faraway (1997), Cuevas, Febrero and Fraiman (2004) and Shen and Faraway (2004) introduced the idea of a functional F-test. In contrast to the pointwise methods of Section 5.5.1 (i), a functional F-test considers differences between models across the whole interval for t. This type of test can be used to compare two nested functional linear models, the smaller model 1 and model 2, having p and q parameters respectively, with $q > p$. The null hypothesis is that labelled model 1 is true, and the alternative hypothesis is that model 2 is true.

The test statistic for the functional F-test is given by

$$\mathrm{F}_{fun} = \frac{(rss_1 - rss_2)/(q-p)}{rss_2/(n-q)}, \tag{5.15}$$

where rss_1 and rss_2 are the residual sum of squares for the smaller and larger model respectively, given by

$$rss_l = \sum_{i=1}^{n} \int (y_i(t) - \hat{y}_{li}(t))^2 \mathrm{d}t \quad \text{for } l = 1, 2,$$

with $\hat{y}_{li}(t)$ the fitted value for run i and model l.

The distribution of (5.15) is too complicated to derive analytically. This led Cuevas et al. (2004) to propose an asymptotic test based on the numerator of (5.15). Shen and Faraway (2004) considered a more intuitive approximation, used later in this chapter, where

$$rss_l \approx \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_i(t_j) - \hat{y}_{li}(t_j))^2,$$

for $l = 1, 2$. The test statistic (5.15) is then compared to an F distribution with $\lambda(q-p)$ and $\lambda(n-q)$ degrees of freedom where λ is the degrees-of-freedom-adjustment-factor defined as

$$\lambda = \frac{(\sum_{k=1}^{\infty} \lambda_k)^2}{\sum_{k=1}^{\infty} \lambda_k^2}. \tag{5.16}$$

The values λ_k are eigenvalues of the covariance function $\gamma(s,t)$ from (5.3). We can estimate λ by

$$\hat{\lambda} = \frac{[\text{trace}(\hat{\Sigma})]^2}{\text{trace}(\hat{\Sigma}^2)}, \tag{5.17}$$

where $\hat{\Sigma}_{j,k} = \sum_{i=1}^{n} \hat{\epsilon_i}(t_j)\hat{\epsilon_i}(t_k)/(n-q)$ with $\epsilon_i(t_j) = y_i(t_j) - \hat{y}_{2i}(t_j)$. Large degrees of freedom are required for an accurate estimate of λ, see Shen and Faraway (2004).

If $\epsilon_i(t_j), \epsilon_i(t_k)$ were identically and independently distributed then $\hat{\lambda} = m$. Correlation between $\epsilon_i(t_j)$ and $\epsilon_i(t_k)$ would reduce the value of $\hat{\lambda}$, leading to lower degrees of freedom for the reference F distribution and hence a higher critical value. Therefore, data that is more highly correlated with t will need a larger value of the test statistic in order to reject the null hypothesis.

Suppose a model is being considered which has p terms. We may wish to examine the importance of each term individually. For the rth term; $r = 0, \ldots, p-1$ we test $H_0 : \beta_r(t) \equiv 0$ against $H_1 : \beta_r(t) \not\equiv 0$ by using the test statistic given by

$$F_r = \frac{rss_{0r} - rss_1}{rss_1/(n-p)}, \tag{5.18}$$

where rss_{0r} is the residual sum of squares under $\beta_r(t) \equiv 0$. Shen and Faraway (2004) provide straightforward methods for calculating this ratio as

$$F_r = \frac{\int \hat{\beta}_r^2(t)\mathrm{d}t}{(rss_1/(n-p))(X'X)^{-1}_{rr}}, \tag{5.19}$$

where $(X'X)^{-1}_{rr}$ is the rth diagonal element of $(X'X)^{-1}$ and rss_1 is the residual sum of squares for the model under the alternative hypothesis. As described by Shen and Faraway (2004), it is possible to approximate the null distribution of F_r by an F distribution with λ and $\lambda(n-p)$ degrees of freedom, where λ is defined in (5.16) and approximated by (5.17).

5.5.2 Application to a simulated example

We return to the example of Section 5.2.1 and consider a simulation of $n = 20$ runs of data: a runs have treatment A applied to them, resulting in responses from model (5.1); $b = n - a$ runs have treatment B and responses from model (5.2). We wish to investigate how inbalance between the number of times each treatment occurs in a simple design affects the performance of tests (i) and (iii) described in the last section.

The functional model for the ith run can be expressed as:

$$y_i(t) = \beta_0(t) + \beta_1(t)x_i + \epsilon_i, \tag{5.20}$$

where $t \in [0, 1]$ and

$$x_i = \begin{cases} 1 & \text{if treatment A is applied to the } i\text{th unit} \\ 0 & \text{if treatment B is applied to the } i\text{th unit .} \end{cases}$$

For each run, observations were simulated at $m = 100$ values of t, equally spaced between 0.01 and 1. We set $\alpha_{0A} = 0.003, \alpha_{1A} = 9.997$ in (5.1) and $\alpha_{0B} = 10$ in (5.2), and obtain values of ϵ_{ij} by random draws from a $N(0, 1)$ distribution, where $i = 1, \ldots, 20, j = 1, \ldots, 100$.

(i) **Pointwise F-ratio**

We used the pointwise method outlined in Section 5.4.1 to fit a functional linear model (5.20) by finding $\hat{\boldsymbol{\beta}}(t)$ to minimise (5.7), that is $\hat{\boldsymbol{\beta}}(t) = (X^T X)^{-1} X^T \boldsymbol{Y}(t)$ for each t.

We simulated five different allocations of numbers of runs having treatment A and B, chosen as different proportions and shown in Table 5.1. A single set of errors was simulated and used for every combination of a and b in order to eliminate Monte Carlo error. The values of the Fratio(t) statistic (5.14) for testing the null hypothesis: $\beta_1(t) \equiv 0$ are shown in Figure 5.3. The 95% percentile of the F distribution with 1 and 18 degrees of freedom

respectively is also shown.

a	b
2	18
5	15
10	10
15	5
18	2

Table 5.1: Values of a and b used in the simulated example

Figure 5.3 indicates how the Fratio(t) changes with t, i.e. we can see the changing evidence of a difference between the treatments. In general, the Fratio(t) is largest when $a = b = 10$. Also note that as the difference between a and b increase, this Fratio decreases as expected.

Examining how the pointwise F-ratio varies over time is useful as we can see whether the significance of the treatment effects changes over the interval. In this example, Figure 5.3 shows that the pointwise F-ratio value decreases as t decreases, subject to random error. This is because the effects of model A and B differ more towards the beginning of the interval, see Figure 5.1 (d). Figure 5.3 shows that for all treatment combinations, except $a = 10, b = 10$, there is a significant difference between models A and B for $t \in [0, 0.80]$. The pointwise F-ratio was larger than the critical value, 4.41. For plots (a), (b), (d) and (e) it is noticeable that for some $t > 0.80$, there is significant difference between models A and B. This is due to the random error in models A and B. When $a = 10, b = 10$, there was a significant difference between the two models over the whole interval, see Figure 5.3 (c).

Figure 5.3 (f) shows the maximum value of the pointwise F-ratio and the functional F-test statistic (5.19) for $a = 1, \ldots, 19$. The maximum pointwise F-ratio and the functional F-test statistic both increase for $a = 1, \ldots, 10$ and decreases from $a = 10, \ldots, 19$. The largest maximum pointwise F-ratio and functional F-test values occurring when $a = 10$ and $b = 10$ is not unexpected as the balanced number of runs for each treatment should give us the most information to discriminate between models A and B, through minimising the variance of $\hat{\beta}_1(t)$.

Functional F-test

We next investigate the functional F-test using the simulated example. Recall, we are comparing a functional linear model to a 'constant' model, therefore there are 18 degrees of freedom for error and 1 for regression.

Table 5.2 contains the values of the functional F-test statistic, F_1, values calculated from

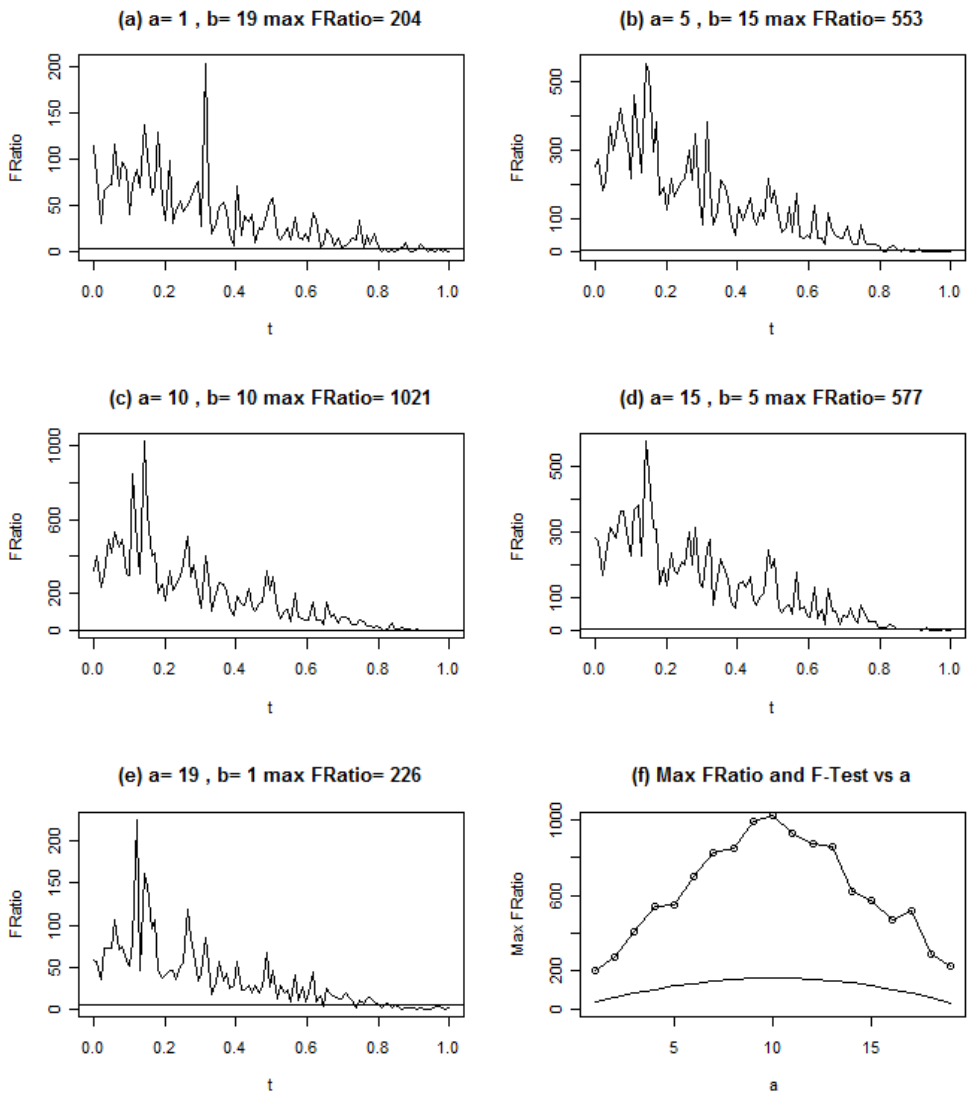

Figure 5.3: Pointwise Fratio(t) against t for testing $\beta_1(t) \equiv 0$ for each combination of a and b together with the 95th percentile of $F_{1,18}$ [plots (a)-(e)] and maximum values of the pointwise F-ratio and the functional F-test statistics for $a = 1, \ldots, 19$ over the interval [0,1] (f)

(5.19) for $r = 1$ to test the null hypothesis that the constant model, $\boldsymbol{Y} = \beta_0(t)\mathbf{1}_n + \boldsymbol{\epsilon(t)}$, where $\mathbf{1}_n$ is an $(n \times 1)$ vector of ones, is true against the alternative that the linear model, (5.20), with the additional parameter $\beta_1(t)$, is true. To calculate critical values, values of $\hat{\lambda}$, the adjusted degrees of freedom were calculated using (5.17) and were also included in the table. Figure 5.3 (f) displays the values of the test statistic graphically. The critical value for this test at the 5% significance level, is shown in Table 5.2 for each combination of a, b values.

Table 5.2 shows that the value of F_1 is largest for $a = 10$ and $b = 10$ and decreases as the difference between a and b increases. In all cases, the F_1 values are greater than the corresponding critical value from the F-distribution. This agrees with the findings from Figure 5.3 for the functional F-test that, for the majority of the interval, the pointwise F-ratio shows a significant difference between the two models.

a	**b**	$\hat{\lambda}$	F_1	**Critical value**
2	18	14.92	34.29	1.71
5	15	15.01	121.63	1.70
10	10	15.01	163.72	1.70
15	5	14.94	123.13	1.70
18	2	14.94	31.82	1.71

Table 5.2: Test statistic and functional F tests for each split of 20 runs between treatment A and B (simulated data).

5.5.3 Application to a tribology experiment

We now investigate selecting a functional linear model to describe the data from the 2^{6-2} tribology experiment where each linear term corresponds to the main effect of a factor. We model the observations on the ith run of the experiment by

$$y_i(t) \quad = \quad \boldsymbol{x}_i^T \beta(t) + \epsilon_i(t), \tag{5.21}$$

for $i = 1, \ldots, 14$, where $\boldsymbol{x}_i^T$ is the ith row of the model matrix X having first entry 1, and rth entry 1 if factor r is at the high level or -1 if it is at the low level ($r = 1, \ldots, 6$).

We wish to investigate whether each main effect should be included in the fitted model using both the functional F-test and the pointwise F-ratio test.

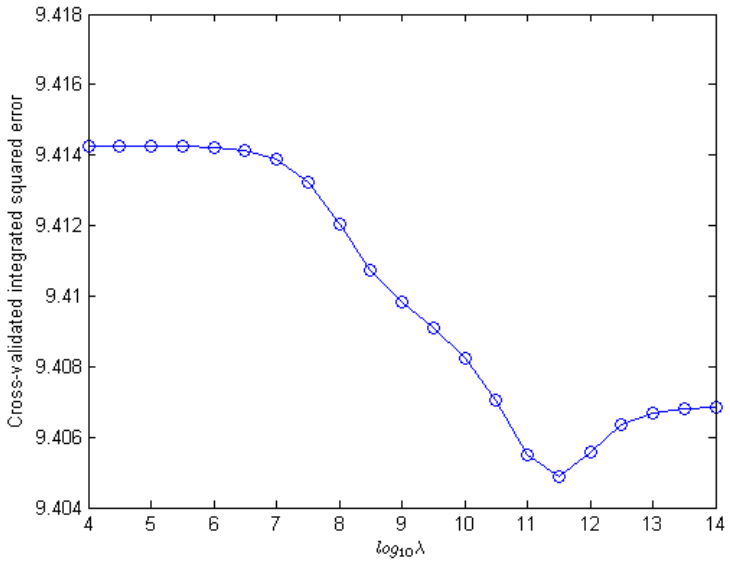

Figure 5.4: Tribology experiment: Cross-validated integrated squared error scores for $\log_{10}\mu = 4, 4.5, \ldots, 14$.

The model (5.21) was fitted using a regularised basis expansion as in Section 5.4.2. We used 8 basis functions for both Y and $\boldsymbol{\beta}$ and (5.10) for the roughness penalty. The smoothing parameter $\mu = 10^{11.5}$ was calculated by minimising the cross-validation integrated squared error, see Figure 5.4.

(i) Pointwise F-ratio

Following Section 5.5.1 (i), pointwise F-ratios were calculated for each of the 1900 points labelled $t = 1, \ldots, 1900$. Figure 5.5 shows that only the intercept was significant over the whole interval at the 5% level. We also found, evidence at the 5% significance level that oxidisation, β_4, had influence on wear but only for the first section of the t interval. These results provide little evidence that any of the factors have an important influence on wear.

We now consider adding two-factor interactions to model (5.21). However, we came across a problem in using cross-validation to calculate the optimal value of μ. When we estimate 14 parameters (intercept, 6 main effects and 7 two-factor interactions), the information matrix $X'X$ is very close to being singular. To overcome this problem we removed one two-factor interaction (between moisture and pin material). The resulting information matrix had full rank. However, the cross-validation method used to calculate optimal μ removes the ith run from the data when fitting a model to predict the ith wear measurement. This means that the corresponding design matrix has the ith row removed. For some i (e.g. $i = 2$) this results in $(X^{(-i)})^T(X^{(-i)})$ becoming singular. This is the result of the lack of orthogonality of the realised design due to the two missing runs.

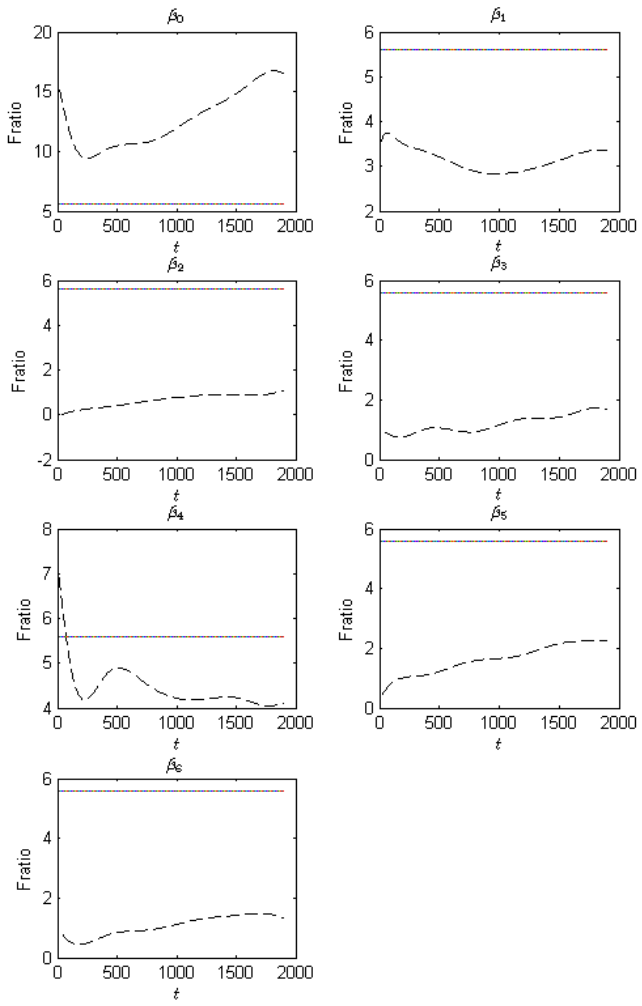

Figure 5.5: Smoothed values of the pointwise F-ratios for $\beta_0, \ldots, \beta_6$, against t, together with the 95th percentile of $F_{1,7}$

Therefore, we now fit a model with all main effects and including the two two-factor interactions between disc material and pin material, and disc material and soot. This model has nine parameters to be estimated. The roughness parameter was again calculated by cross-validation to be $\mu = 10^{11.5}$.

Figure 5.6 shows that only the intercept was significant over the whole interval at the 5% significance level. We also found that oxidisation, β_4, was significant at the 5% level for most of the interval $t \in [1, 1900]$, which differs from the findings when the model fitted had only main effects. As the realised design with 14 runs is not orthogonal it is, of course, possible for the addition of interactions to change the estimates of main effect terms.

(iii) Functional F-test

Analogous to our strategy for the pointwise F-ratio, we first consider the model which only includes main effects and an intercept term. We fitted (5.21) and minimised (5.11) to find the seven estimators $\hat{\beta}_0, \ldots, \hat{\beta}_6$. The functional F-test (5.19) was then performed to individually test which effects should be in the model. The critical value of the F distribution with $\hat{\lambda} = 1.035$ and $\hat{\lambda}(n-p) = 1.035(14-7) = 7.25$ degrees of freedom (see (5.17)) at the 5% significance level is 5.47. The values of the test statistic for each effect can be found in Table 5.3. We see that the intercept was found to be the only significant term.

In order to compare the above results with those of the pointwise F-ratio, we constructed a functional F-test for the model with main effects and the two interactions: between disc material, and pin material, and disc material and soot. For a 5% significance level test, the critical value of the F distribution with $\hat{\lambda} = 1.029$ and $\hat{\lambda}(n-p) = 1.029(14-9) = 5.15$ degrees of freedom is 6.46. The test statistic value for each term is given in Table 5.3. The intercept was found to be the only significant term, as it was when a model with only main effects was fitted. Note that for oxidation, the β_4 term, the test statistic is close to the critical value. We may expect this since this term was significant over most of the interval; see Figure 5.6.

5.6 Conclusions from the examples

The example constructed using the simulated data in Section 5.5.2 provides evidence that unequal numbers of runs for each treatment lead to smaller values of the test statistics in both the pointwise and functional F-tests. This is due to the larger variance of $\hat{\beta}_1(t)$ than would occur for a balanced design and is analogous to what happens in a 'scalar' linear

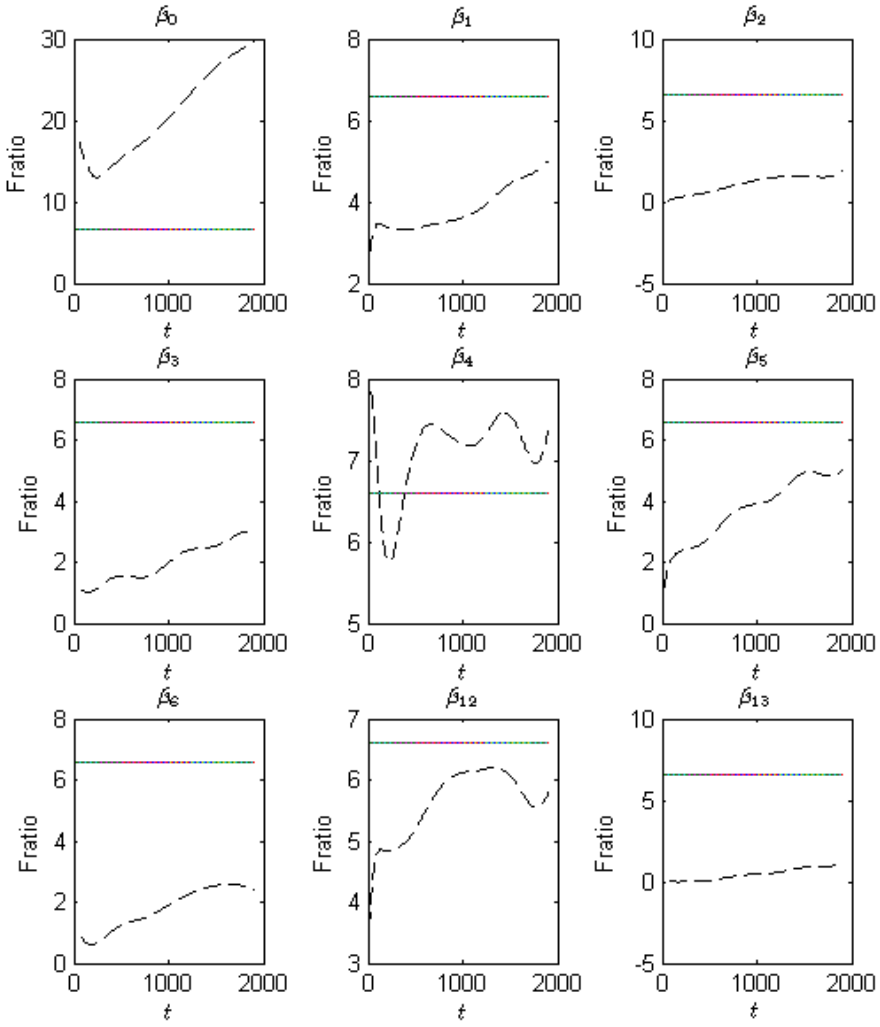

Figure 5.6: Pointwise F-ratio plots for $\beta_0, \ldots, \beta_{13}$, representing all main effects and two interactions, together with the 95th percentile of $F_{1,5}$.

Model term	Model 1	Model 2
β_0	10.75	17.19
β_1	3.09	3.45
β_2	0.54	0.87
β_3	1.02	1.63
β_4	3.80	6.08
β_5	1.57	3.28
β_6	0.90	1.45
β_{12}	-	5.09
β_{13}	-	0.38

Table 5.3: Example 2: Functional F-test statistics for model 1 (all main effects) and model 2 (all main effects and the disc material–pin material and disc material–soot interactions).

model.

The performance of the pointwise and functional F-tests explored in the simulated example can only be indicative because only one set of data was simulated.

As highlighted in Section 5.5.1, the functional F-test does not have the multi-testing disadvantages of the pointwise F-ratio and takes some account of the possible correlation in the observations. Hence the functional test is preferred. In the second half of this chapter, we find optimal designs using a criterion which is similar to the functional F-test statistic.

5.7 Optimal designs for model discrimination

The remainder of this chapter is focussed on optimal design for model discrimination. In this section, we review goodness-of-fit testing and the criterion of T-optimality for a univariate response (Atkinson and Fedorov, 1975), describe an adaptation of T-optimality for multivariate response models (cf Uciński and Bogacka, 2005) and develop a T-optimality criterion for functional linear models. In the following section we assess the performance of designs found from this criterion to approximate the power for discriminating between two functional linear models using simulation studies.

We begin by considering the notation for the univariate case to illustrate the design problem. Suppose that we wish to compare two linear models:

$$\text{model 1:} \quad \boldsymbol{Y_1} = F\boldsymbol{\beta} + \boldsymbol{\epsilon}, , \tag{5.22}$$

and

$$\text{model 2:} \quad \boldsymbol{Y_2} = G\boldsymbol{\theta} + \boldsymbol{\eta}, \tag{5.23}$$

where $\boldsymbol{Y}_i$ $(i = 1, 2)$ are the $n \times 1$ vectors of observations, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are, respectively a $p \times 1$ and $q \times 1$ vectors of parameters, $\boldsymbol{\epsilon}, \boldsymbol{\eta} \sim N(\mathbf{0}, \sigma^2 I)$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ and σ^2 is known. Here F and G are model matrices given by

$$F = \begin{pmatrix} f_0(x_1) & \dots & f_{p-1}(x_1) \\ \vdots & & \vdots \\ f_0(x_n) & \dots & f_{p-1}(x_n) \end{pmatrix}, \tag{5.24}$$

and

$$G = \begin{pmatrix} g_0(x_1) & \dots & g_{q-1}(x_1) \\ \vdots & & \vdots \\ g_0(x_n) & \dots & g_{q-1}(x_n) \end{pmatrix}. \tag{5.25}$$

The points x_i for $i = 1, \dots, n$ are design points, that is values taken by the single variable $-1 \leq x \leq 1$, and the functions f and g are known functions of x.

We may regard an optimal design for model discrimination, as a design which enables a 'best' test of the hypothesis that model 1 is correct given data arising from model 2. Therefore the theory of optimal design for model discrimination can be motivated by goodness-of-fit testing used in classical linear modelling.

5.7.1 Likelihood-based goodness-of-fit testing

A goodness-of-fit test compares a given model with $p < n$ parameters to the full saturated model, assumed to have n parameters. A likelihood-based goodness-of-fit test uses the deviance as a statistic to measure discrepancy between models (McCullagh and Nelder, 1989, p33). The deviance for model 1 is given by

$$D(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = 2l(\boldsymbol{Y}, \boldsymbol{Y}) - 2l(\boldsymbol{Y}, \hat{\boldsymbol{Y}}), \tag{5.26}$$

where

$$\hat{\boldsymbol{Y}} = F\hat{\boldsymbol{\beta}},$$

$$l(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(\boldsymbol{Y} - F\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - F\hat{\boldsymbol{\beta}}),$$

and

$$l(\boldsymbol{Y}, \boldsymbol{Y}) = -\frac{n}{2} \log\left(2\pi\sigma^2\right).$$

Here $\hat{\boldsymbol{\beta}} = (F^T F)^{-1} F^T \boldsymbol{Y}$ is the least squares estimator under model 1. The deviance for model 1 is therefore defined as

$$D(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{1}}) = \frac{(\boldsymbol{Y} - F\hat{\boldsymbol{\beta}})^T (\boldsymbol{Y} - F\hat{\boldsymbol{\beta}})}{\sigma^2}. \tag{5.27}$$

If the deviance is larger than an appropriate critical value from a χ^2 distribution with $n - p$ degrees of freedom, then there is evidence to reject the null hypothesis that model 1 is an adequate description of the data.

5.7.2 T-optimality

In this section we introduce the ideas behind optimal design for model discrimination. We review the work by Atkinson and Fedorov (1975) for a univariate response, apply these ideas to the case of multivariate response and develop a T-optimality criterion for functional linear models. We illustrate these methods with a simple example.

Atkinson and Fedorov (1975) introduced T-optimality for univariate models, with an objective function based on the sum of squares for lack of fit of a first model given that the data comes from a second model.

5.7.2.1 Univariate response

We review the methodology for designing experiments to discriminate between two univariate linear models.

The deviance, or sum of squares, for testing the assumption that model 1 is correct given we expect data to come from model 2 is given by substituting $E(\boldsymbol{Y_2}) = G\boldsymbol{\theta}$ into (5.27):

$$
\begin{aligned}
D(E(\boldsymbol{Y_2}), \hat{\boldsymbol{Y}}_{\boldsymbol{1}}) &= \frac{(E(\boldsymbol{Y_2}) - F\hat{\boldsymbol{\beta}})^T (E(\boldsymbol{Y_2}) - F\hat{\boldsymbol{\beta}})}{\sigma^2} \\
&\propto \left[E(\boldsymbol{Y_2}) - F(F^TF)^{-1}F^TE(\boldsymbol{Y_2})\right]^T \left[E(\boldsymbol{Y_2}) - F(F^TF)^{-1}F^TE(\boldsymbol{Y_2})\right], \\
&\propto \left[G\boldsymbol{\theta} - F(F^TF)^{-1}F^TG\boldsymbol{\theta}\right]^T \left[G\boldsymbol{\theta} - F(F^TF)^{-1}F^TG\boldsymbol{\theta}\right] \\
&\propto \boldsymbol{\theta}^T G^T \left[I - H\right]^2 G\boldsymbol{\theta} \\
&\propto \boldsymbol{\theta}^T G^T \left[I - H\right] G\boldsymbol{\theta}, \qquad (5.28)
\end{aligned}
$$

where $H = F(F^TF)^{-1}F^T$ is the hat matrix and $(I - H)$ is an idempotent matrix. An exact T-optimal design maximises $\Psi(\xi) = \boldsymbol{\theta}^T G^T \left[I - H\right] G\boldsymbol{\theta}$ with respect to the design points, which feature in both G and H.

Note that in the univariate case, $\Psi(\xi)$ is proportional to the non-centrality parameter for the distribution of the test statistic (5.27) assuming an alternative hypothesis of model 2 being true; see McCulloch, Searle and Neuhaus (2008, p. 126).

In the examples in this chapter, we find approximate optimal designs rather than exact designs. Recall from Section 1.4, approximate designs are represented by a measure ξ on the design region χ. An approximate design with observations at s distinct design points, often referred to as support points, in χ is written:

$$
\xi = \left\{ \begin{array}{cccc} x_1 & x_2 & ... & x_s \\ w_1 & w_2 & ... & w_s \end{array} \right\}. \qquad (5.29)
$$

For simplicity, and without loss of generality, in this chapter we assume that points $x_1, \ldots, x_s$ in an n-point ($n \geq s$) exact design are distinct. The first line of (5.29) gives the s support points and the second line gives the associated design weights, $0 < w_i \leq 1$; $\sum_{i=1}^{s} w_i = 1$. More details can be found in Section 1.4.

Approximate designs are useful for the simulation studies later in this chapter, as they allow the straightforward construction of designs with large numbers of points through scaling of s. Instead of finding an exact design for large n, for example $n > 50$, we can find an approximate design and use the weights, rounding nw_i, to construct an exact design.

Criterion 5.1. *A T-optimal design ξ^* for discriminating between models 1 and 2 maximises*

$$\Psi(\xi) = \boldsymbol{\theta}^T G^T (I - H_w)^T W (I - H_w) G\boldsymbol{\theta}, \tag{5.30}$$

where $W = diag(w_i)$ for $i, \ldots, s$ and $H_w = F(F^T W F)^{-1} F^T W$.

Proposition 5.1. *The optimal design under Criterion 5.1 is independent of $\boldsymbol{\theta}$ if models 1 and 2, defined in equations* (5.22) *and* (5.23)*, are nested and only differ by one term $(q = p + 1)$.*

Proof. The proof follows that of Atkinson et al. (2007, p. 360). Assume that model 2 is the larger model and partition the model matrix G into $[F : \tilde{F}]$ where $\tilde{F}$ is an $s \times 1$ vector. The vector of parameters $\boldsymbol{\theta}$ can also be partitioned as $\boldsymbol{\theta}^T = [\boldsymbol{\theta_1^T} : \theta_2]$, with $\boldsymbol{\theta}_1$ of size $p \times 1$ and θ_2 a scalar.

The T-optimality objective function can then be written as follows:

$$\begin{aligned}
\Psi(\xi) &= \boldsymbol{\theta}^T G^T (I - H)^T W (I - H) G\boldsymbol{\theta} \\
&= [\boldsymbol{\theta_1^T} : \theta_2] \begin{bmatrix} F^T \\ \tilde{F}^T \end{bmatrix} [I - H]^T W [I - H] [F : \tilde{F}] \begin{bmatrix} \boldsymbol{\theta_1} \\ \theta_2 \end{bmatrix} \\
&= [\boldsymbol{\theta_1^T} : \theta_2] A \begin{bmatrix} \boldsymbol{\theta_1} \\ \theta_2 \end{bmatrix}.
\end{aligned}$$

Here,

$$\begin{aligned}
A &= \begin{bmatrix} F^T \\ \tilde{F}^T \end{bmatrix} [I - H_w]^T W [I - H_w] [F : \tilde{F}] \\
&= \begin{bmatrix} F^T [I - H_w]^T W [I - H_w] F & F^T [I - H_w]^T W [I - H_w] \tilde{F} \\ \tilde{F}^T [I - H_w]^T W [I - H_w] F & \tilde{F} [I - H_w]^T W [I - H_w] \tilde{F} \end{bmatrix} \\
&= \begin{bmatrix} 0 & 0 \\ 0 & \tilde{F}^T (I - H_w)^T W (I - H_w) \tilde{F} \end{bmatrix},
\end{aligned} \tag{5.31}$$

as $H_w F = (F^T W F)^{-1} F^T W F = I$. Therefore,

$$\begin{aligned}\Psi(\xi) &\propto [\boldsymbol{\theta_1^T} : \theta_2] \begin{bmatrix} 0 & 0 \\ 0 & \tilde{F}^T(I-H_w)^T W(I-H_w)\tilde{F} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta_1} \\ \theta_2 \end{bmatrix} \\ &= \theta_2 \tilde{F}^T(I-H_w)^T W(I-H_w)\tilde{F}\theta_2 \\ &\propto \tilde{F}^T(I-H_w)^T W(I-H_w)\tilde{F}.\end{aligned}$$

The last step follows from θ_2 being scalar. Hence the result is shown. □

Example

We wish to find an approximate T-optimal design to enable a test of whether a simple linear model (model 1):

$$y_{1i} = \beta_0 + \beta_1 x_i + \epsilon_i,$$

is a suitable fit given we expect data from a quadratic model (model 2):

$$y_{2i} = \beta_0 + \theta_1 x_i + \theta_2 x_i^2 + \eta_i.$$

for $i = 1, \ldots, s$. Hence F has the form

$$F = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_s \end{pmatrix}, \tag{5.32}$$

where $x_1, ..., x_s$ are support points. The expected data are assumed to follow model 2, i.e. $E(\boldsymbol{Y_2}) = G\boldsymbol{\theta}$, where: $\boldsymbol{\theta}$, a 3×1 vector of unknown model parameters and

$$G = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_s & x_s^2 \end{pmatrix}. \tag{5.33}$$

From Proposition 5.1, the two models differ by only one term and the value of $\boldsymbol{\theta}$ does not affect the optimal design.

In order to find approximate T-optimal designs numerically, we used a grid of $n_g = 21$ equally spaced points on the interval $[-1, 1]$ and found the optimal weights $w_1, w_2, ..., w_{n_g}$ to maximise $\Psi(\xi)$ in Criterion 5.1 using the Nelder-Mead algorithm. We would expect some of the weights to be zero, with the non-zero weights indicating which s of the n_g grid points were in the support of the design.

The optimal design was found to be

$$\xi = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 0.25 & 0.5 & 0.25 \end{array} \right\}. \tag{5.34}$$

5.7.2.2 Multivariate response

We describe T-optimal designs for the multivariate case as a stepping stone to designs for the functional linear model, although we also derive some results which are of interest in themselves.

Uciński and Bogacka (2005) considered T-optimality for multivariate non-linear models. Specifically they found T-optimal designs for discriminating between two specific multiresponse models, assuming that observations on an individual response variable were correlated. Their work was applied to dynamic systems and chemical kinetic models. Here we adapt their ideas to linear models.

In multivariate regression, we record m observations, one for each response variable, for each of the n runs of an experiment. For example, the m responses, of the first run are of the form

$$\begin{aligned} y_{11} &= \beta_{01} f_0(x_1) + \beta_{11} f_1(x_1) + \ldots + \beta_{p-1,1}(x_1) f_{p-1}(x_1) + \epsilon_{11} \\ &\vdots \\ y_{1m} &= \beta_{0m} f_0(x_1) + \beta_{1m} f_1(x_1) + \ldots + \beta_{p-1,m}(x_1) f_{p-1}(x_1) + \epsilon_{1m}, \end{aligned}$$

where the error variable $\boldsymbol{\epsilon_1} = (\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{1m})^T$ has $E(\boldsymbol{\epsilon_1}) = 0$ and $Var(\boldsymbol{\epsilon_1}) = \Sigma$ (Johnson and Wichern, 1998).

Suppose we wish to discriminate between the following two multivariate linear models:

$$\text{model 1:} \quad Y_1 = FB + R_1,$$

and

$$\text{model 2:} \quad Y_2 = GT + R_2. \tag{5.35}$$

Here the matrices Y_1 and Y_2 hold the $n \times m$ responses from models 1 and 2, respectively:

$$Y_i = \begin{pmatrix} Y_{i11} & \dots & Y_{i1m} \\ \vdots & & \vdots \\ Y_{in1} & \dots & Y_{inm} \end{pmatrix} \quad (i = 1, 2).$$

The model matrices F and G are defined in (5.24) and (5.25). The matrices, B and T, of parameters in models 1 and 2, respectively are

$$B = \begin{pmatrix} \beta_{01} & \dots & \beta_{0m} \\ \vdots & & \vdots \\ \beta_{p_1-1,1} & \dots & \beta_{p_1-1,m} \end{pmatrix},$$

and

$$T = \begin{pmatrix} \theta_{01} & \dots & \theta_{0m} \\ \vdots & & \vdots \\ \theta_{q_1-1,1} & \dots & \theta_{q_1-1,m} \end{pmatrix}.$$

The matrices of errors, R_1 and R_2, are given by

$$R_1 = \begin{pmatrix} \epsilon_{11} & \dots & \epsilon_{1m} \\ \vdots & & \vdots \\ \epsilon_{n1} & \dots & \epsilon_{nm} \end{pmatrix},$$

and

$$R_2 = \begin{pmatrix} \eta_{11} & \dots & \eta_{1m} \\ \vdots & & \vdots \\ \eta_{n1} & \dots & \eta_{nm} \end{pmatrix},$$

where $\text{vec}(R_i^T) \sim N(\mathbf{0}, I \otimes \Sigma)$, with Σ the within run covariance and the vector operator for a matrix A, written in terms of columns $A = [\boldsymbol{a}_1, \dots \boldsymbol{a}_n]$, is defined as

$$\text{vec}(A) = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_n \end{pmatrix}. \tag{5.36}$$

For each of the i runs, $i = 1, \dots, n$, the (j,k)th element of the covariance matrix is given by $\Sigma_{jk} = cov(\epsilon_{ij}, \epsilon_{ik})$. Note that we assume observations from different runs are not correlated.

The multivariate maximised log-likelihood is given by

$$l\left(Y, \hat{Y}_1\right) = -\frac{mn}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{1}{2}tr\left[\Sigma^{-1}\left\{Y - F\hat{B}\right\}^T\left\{Y - F\hat{B}\right\}\right],$$

see Johnson and Wichern (1998). The deviance for testing model 1 is correct, given the expected data from model 2, is given by

$$\begin{aligned} D\left(E(Y_2), \hat{Y}_1\right) &= 2l\left(E(Y_2), E(Y_2)\right) - 2l\left(E(Y_2), \hat{Y}_1\right) \\ &= -mn\log(2\pi) - n\log(|\Sigma|) + mn\log(2\pi) + n\log(|\Sigma|) \\ &\quad + tr\left[\Sigma^{-1}\left\{E(Y_2) - F\hat{\boldsymbol{B}}\right\}^T\left\{E(Y_2) - F\hat{\boldsymbol{B}}\right\}\right] \\ &= tr\left[\Sigma^{-1}\left\{E(Y_2) - F\hat{\boldsymbol{B}}\right\}^T\left\{E(Y_2) - F\hat{\boldsymbol{B}}\right\}\right] + C \\ &= tr\left[\Sigma^{-1}T^TG^T(I-H)GT\right] + C, \end{aligned} \tag{5.37}$$

where m is the number of within run observations and C does not depend on the design or the data. Details on the derivation of the multivariate normal deviance can be found

in Johnson and Wichern (1998).

Criterion 5.2. *A T-optimal design ξ^* for discriminating between two multivariate response models maximises*

$$\Psi(\xi) \;=\; tr\left[\Sigma^{-1}T^TG^T(I-H_w)^TW(I-H_w)GT\right], \tag{5.38}$$

where W is the matrix of design weights in (5.29) and H_w is defined after (5.30).

Proposition 5.2. *The optimal design from Criterion 5.2, is independent of the choice of T and Σ if models 1 and 2 are nested and only differ by one term ($q = p + 1$).*

Proof. The proof is similar to that of Proposition 5.1. Once again, we assume model 2 is the larger model and partition the model matrix G into $[F : \tilde{F}]$. The matrix of parameters T can also be partitioned with $T^T = [B^T : \tilde{T}^T]$, with B of size $p \times m$ and $\tilde{T}$ a $1 \times m$ vector.

The deviance for lack of fit of model 1 can be then written as follows:

$$\begin{aligned}
\Psi(\xi) &\propto tr\left\{\Sigma^{-1}T^TG^T(I-H_w)^TW(I-H_w)GT\right\} \\
&= tr\left\{\Sigma^{-1}[B^T:\tilde{T}^T]\begin{bmatrix}F^T\\ \tilde{F}^T\end{bmatrix}[I-H_w]^TW[I-H_w][F:\tilde{F}]\begin{bmatrix}B\\ \tilde{T}\end{bmatrix}\right\} \\
&= tr\left\{\Sigma^{-1}\tilde{T}^T\tilde{F}^T(I-H_w)^TW(I-H_w)\tilde{F}\tilde{T}\right\} \\
&= tr\left\{\tilde{F}^T(I-H_w)^TW(I-H_w)\tilde{F}\tilde{T}\Sigma^{-1}\tilde{T}^T\right\} \\
&= \tilde{T}\Sigma^{-1}\tilde{T}^T tr\left\{\tilde{F}^T(I-H_w)^TW(I-H_w)\tilde{F}\right\} \\
&\propto tr\left\{\tilde{F}^T(I-H_w)^TW(I-H_w)\tilde{F}\right\},
\end{aligned}$$

using the result in (5.31) and the fact that $\tilde{T}\Sigma^{-1}\tilde{T}^T$ is a scalar and does not depend on the design.

Hence we have shown that the optimal design is independent of the choice of T and Σ when models 1 and 2 only differ in one term. □

Example

We wish to compare two multivariate models: the single linear multivariate model (model 1)

$$
\begin{aligned}
Y &= FB + R_1 \\
&= \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_{01} & \dots & \beta_{0m} \\ \beta_{1,1} & \dots & \beta_{1,m} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} & \dots & \epsilon_{1m} \\ \vdots & & \vdots \\ \epsilon_{n1} & \dots & \epsilon_{nm} \end{pmatrix},
\end{aligned}
$$

and the quadratic model (model 2)

$$
\begin{aligned}
Y &= GT + R_2 \\
&= \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \\ 1 & x_n & x_n^2 \end{pmatrix} \begin{pmatrix} \theta_{01} & \dots & \theta_{0m} \\ \theta_{11} & \dots & \theta_{1m} \\ \theta_{21} & \dots & \theta_{2m} \end{pmatrix} + \begin{pmatrix} \eta_{11} & \dots & \eta_{1m} \\ \vdots & & \vdots \\ \eta_{n1} & \dots & \eta_{nm} \end{pmatrix}.
\end{aligned}
$$

Once again we assume that model 1, the linear model, is true. We wish to find an approximate design which allows us to discern whether a linear model is appropriate given that expect data from the quadratic model.

As the two models differ by only one term, the model parameters, T, and covariance matrix, Σ, do not influence the optimal design by Proposition 5.2.

An optimal design was found using the simplex Nelder Mead numerical search algorithm for different values of m.

$$
\xi = \begin{Bmatrix} -1 & 0 & 1 \\ 0.25 & 0.5 & 0.25 \end{Bmatrix},
$$

to be T-optimal.

5.7.2.3 Functional response

We adapt the methodology for a multivariate response to the case of functional response and propose a T-optimality criterion for design selection.

Recall that for a functional response, a linear model is written in the form

$$y_i(t) = \boldsymbol{f}^T(x)\boldsymbol{\beta}(t) + \epsilon_i(t),$$

for the i^{th} run of the experiment, $i = 1, \ldots, n$ and $\boldsymbol{f}(x) = (f_0(x), \ldots, f_{p-1}(x))$. We assume that errors $\epsilon_i(t)$ and $\epsilon_i(s)$ are realisations from a Gaussian stochastic process with mean zero and covariance function $\gamma(s,t)$ with s, t belonging to a real interval.

As for the multivariate and univariate cases, we wish to choose design points which enable a test of whether model 1 is true, given that we expect data to come from model 2. For a functional response, we define

$$\text{model 1:} \quad y_{1i}(t) = \boldsymbol{f}^T(x)\boldsymbol{\beta}(t) + \epsilon_i(t), \tag{5.39}$$

and

$$\text{model 2:} \quad y_{2i}(t) = \boldsymbol{g}^T(x)\boldsymbol{\theta}(t) + \eta_i(t), \tag{5.40}$$

with $\eta_i(t)$ following the same definition as $\epsilon_i(t)$.

For each run we assume that the functional response can be evaluated at m points. These points may be the actual measurements or predictions from the reconstructed functional response (Shen and Faraway, 2004). We can approximate $y_{ji}(t)$, for $j = 1, 2$, by the vector $(y_{ji}(t_1), \ldots, y_{ji}(t_m))$, and the realised dataset may be placed in a matrix

$$Y_j = \begin{pmatrix} y_{j1}(t_1) & \ldots & y_{j1}(t_m) \\ \vdots & & \vdots \\ y_{jn}(t_1) & \ldots & y_{jn}(t_m) \end{pmatrix}.$$

Note that in general there may be different numbers of observations per run. In the work presented here, we assume there are m observations per run.

Now we can define model 1 and model 2 in terms of realised data

$$\text{model 1:} \quad Y_1 = FB + R_1, \tag{5.41}$$

and

$$\text{model 2:} \quad Y_2 = GT + R_2. \tag{5.42}$$

Here B is the matrix of parameters

$$B = \begin{pmatrix} \beta_1(t_1) & \dots & \beta_1(t_m) \\ \vdots & & \vdots \\ \beta_{p_1-1}(t_1) & \dots & \beta_{p_1-1}(t_m) \end{pmatrix},$$

and R_1 is the matrix of errors

$$R_1 = \begin{pmatrix} \epsilon_1(t_1) & \dots & \epsilon_1(t_m) \\ \vdots & & \vdots \\ \epsilon_n(t_1) & \dots & \epsilon_n(t_m) \end{pmatrix},$$

Also, T is a $q \times m$ parameter matrix and R_2 is an $n \times m$ error matrix defined similarly to B and R_1, respectively.

The log-likelihood is given by

$$l\left(Y, \hat{Y}_1\right) = -\frac{mn}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{1}{2}tr\left[\Sigma^{-1}\left\{Y - F\hat{B}\right\}^T\left\{Y - F\hat{B}\right\}\right],$$

where Σ is the variance-covariance matrix for model 1. Analogous to (5.37), the deviance is given by

$$D\left(E(Y_2), \hat{Y}_1\right) \;\propto\; tr\left[\Sigma^{-1}T^TG^T(I-H)^TGT\right],$$

Hence, following Section 5.7.2.2, we can find approximate T-optimal designs using the following criterion

Criterion 5.3. *A functional T-optimal approximate design ξ^* maximises*

$$\Psi(\xi) = tr\left\{\Sigma^{-1}T^T(I - H_w)^T W(I - H_w)T\right\}, \tag{5.43}$$

where W is the matrix of design weights in (5.29) and $H_w = (F^T W F)^{-1} F^T W$.

Proposition 5.3. *The choice of design points satisfying Criterion 5.3 is independent of the choice of T and Σ if models 1 and 2 are nested and only differ by one term.*

Proof. Follows directly from Proposition 5.2 □

Proposition 5.4. *Univariate, multivariate and hence functional T-optimal designs are identical when model 1 and model 2 are nested and only differ by one term.*

Proof. The objective functions for the univariate, multivariate and hence functional T-optimality criteria are identical for nested models that only differ by one term. Therefore the optimal designs are the same. □

Corollary 5.1. *If models 1 and 2 are nested and differ by more than one term, the T-optimal design depends only on Σ and the additional parameters in model 2.*

Proof. Proof follows from the proof of Proposition 5.2, where $\tilde{T}$ is a $(q-p) \times m$ matrix and $\tilde{F}$ is a $m \times (q-p)$ matrix. □

Note that in the case where $\Sigma = I$ we find that

$$\begin{aligned} D\left(E(Y_2), \hat{Y}_1\right) &\propto tr\left[T^T G^T (I - H)^T G T\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left(E\left[y_{2i}(t_j)\right] - \hat{y}_{1i}(t_j)\right)^2, \end{aligned} \tag{5.44}$$

where $y_{2i}(t_j)$ is a realisation of (5.39) and $\hat{y}_{1i}(t_j)$ is a fitted value from (5.40). The function (5.44) is equivalent to the test statistic proposed in Shen and Faraway (2004) and Shen and Xu (2006) when the data are observed without error. These authors accounted for correlation by adapting the degrees of freedom for the test rather than explicitly through inclusion of a the matrix Σ, see Section 5.5.1.

Example

Suppose we wish to discriminate between two functional models: the linear model

$$y_{1i}(t) = \beta_0(t) + \beta_1(t)x + \epsilon_i(t),$$

and the quadratic model

$$y_{2i}(t) = \theta_0(t) + \theta_1(t)x + \theta_2(t)x^2 + \eta_i(t).$$

We again assume that the linear model (model 1) is true. We wish to find an approximate design which allows us to discern whether a linear model is appropriate given we expect data from the quadratic model. In order to do this we maximise the objective function (5.43). As the models differ by only one term, neither T nor Σ influence the optimal design. Using the Nelder-Mead algorithm, we find the optimal design to be

$$\xi = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 0.25 & 0.5 & 0.25 \end{array} \right\}, \tag{5.45}$$

agreeing with Proposition 5.4.

We verify that the design (5.45) is indeed T-optimal by showing that it satisfies a sufficient condition for optimality obtained from a General Equivalence Theorem (Atkinson and Fedorov, 1975). A sufficient condition for the design ξ^* to be T-optimal is that

$$\max_{x \in \chi} \psi(x, \xi^*) \leq \Psi(\xi^*),$$

where

$$\psi(x, \xi^*) = (E(y_2(t)) - \hat{y}_1(t))^2, \tag{5.46}$$

and, in our example,

$$\begin{array}{ll} \text{model 1:} & E(y_1(t)) = \beta_0(t) + \beta_1(t)x \\ \text{model 2:} & E(y_2(t)) = \theta_0(t) + \theta_1(t)x + \theta_2(t)x^2. \end{array}$$

The function $\psi(x, \xi)$ is the derivative of $\Psi(\xi)$ in the direction of the point x. Here

$$\hat{y}_1(t) = \hat{\beta}_0(t) + \hat{\beta}_1(t)x. \tag{5.47}$$

Assuming expected data from model 2,

$$\begin{aligned} \hat{\boldsymbol{\beta}}(t) &= (F^T W F)^{-1} F^T W E(Y_2(t)) \\ &= \begin{pmatrix} \theta_0(t) + \frac{1}{2}\theta_2(t) \\ \theta_1(t) \end{pmatrix}, \end{aligned} \tag{5.48}$$

where $W = \text{diag}([0.25, 0.5, 0.5])$,

$$F = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix},$$

and $E(Y_2(t))$ is the vector of expected response under model 2 for each of $x = -1, 0, 1$,

$$E(Y_2(t)) = \begin{pmatrix} \theta_0(t) - \theta_1(t) + \theta_2(t) \\ \theta_0(t) \\ \theta_0(t) + \theta_1(t) + \theta_2(t) \end{pmatrix}.$$

Substituting the values for $\hat{\boldsymbol{\beta}}$ from (5.48) into (5.47) and (5.46), we find

$$\begin{aligned} \psi(x, \xi_n^*) &= \left(\theta_0(t) + \theta_1(t)x + \theta_2(t)x^2 - \theta_0(t) - \frac{1}{2}\theta_2(t) - \theta_1(t)x \right)^2 \\ &= \left[\theta_2(t) \left(x^2 - \frac{1}{2} \right) \right]^2. \end{aligned} \tag{5.49}$$

Now $\max_{x \in [-1,1]} \psi(x, \xi_n^*) = \frac{1}{4}\theta_2^2(t)$ when $x = -1, 0, 1$. For this example,

$$\begin{aligned}\Psi(\xi_n^*) &= \frac{1}{4}\left(\theta_0(t) - \theta_1(t) + \theta_2(t) - \theta_0(t) - \frac{1}{2}\theta_2(t) + \theta_1(t)x\right)^2 \\ &+ \frac{1}{2}\left(\theta_0 - \theta_0(t) - \frac{1}{2}\theta_2(t)\right)^2 \\ &+ \frac{1}{4}\left(\theta_0(t) + \theta_1(t) + \theta_2(t) - \theta_0(t) - \frac{1}{2}\theta_2(t) - \theta_1(t)x\right)^2 \\ &= \frac{1}{4}\theta_2^2(t)\end{aligned}$$

Therefore $\psi(x, \xi_n^*) \leq \Psi(\xi_n^*)$ with equality at the support points $x = -1, 0, 1$. Therefore by the General Equivalence Theorem for T-optimality, the design in (5.45) is indeed T-optimal.

5.8 Simulation studies to assess power

In this section we conduct simulation studies to assess the power of the functional F-test for data obtained using the T-optimal design found in Section 5.7.2.3. Recall that the power of a test is the probability that the test will reject the null hypothesis when the null hypothesis is false, that is, Power $= P(H_0$ is rejected$|H_0$ is false). Simulation studies were conducted for two examples: the first testing the goodness-of-fit of a linear model given the data came from a quadratic model, and the second one testing the goodness-of-fit of a first order model with two factors and their interaction when the data came from a two-factor model with their interaction and both quadratic terms.

5.8.1 Example 1

In this example we test the hypothesis that a first order model describes the data, when the alternative hypothesis states that a quadratic model is correct using the optimal design (5.45). Specifically,

$$\begin{aligned}H_0 : \boldsymbol{Y}(t) &= \boldsymbol{Y}_1(t) = \beta_0(t) + \beta_1(t)x + \boldsymbol{\epsilon}(t) \\ H_1 : \boldsymbol{Y}(t) &= \boldsymbol{Y}_2(t) = \theta_0(t) + \theta_1(t)x + \theta_2(t)x^2 + \boldsymbol{\eta}(t).\end{aligned}$$

We assume a linear model for the observation on each run i

$$y_{1i}(t) = \beta_0(t) + \beta_1(t)x_i + \beta_2(t)x_i^2 + \epsilon_i(t), \tag{5.50}$$

but simulated a response from the model

$$y_{2i}(t) = \theta_0(t) + \theta_1(t)x_i + \theta_2(t)x_i^2 + \eta_i(t), \tag{5.51}$$

in order to test the assumption that a linear model was true, given the data was simulated from a quadratic model.

We assume an AR1 auto-regressive covariance function where $\epsilon_i(t), \eta_i(t) \sim N(0, \sigma^2)$ and $Cov(\epsilon_i(t_j), \epsilon_i(t_k)) = Cov(\eta_i(t_j), \eta_i(t_k)) = \rho^{|j-k|}$ for t_j and t_k on some real interval and $|\rho| \leq 1$. Throughout the study, we employ $t \in [-1, 1]$. Note that we will consider a different covariance structure later in the section. Now to generate data, the parameter functions $\theta_0, \theta_1, \theta_2$ are defined in this example to be

$$\begin{aligned}
\theta_0(t) &= \alpha_{00} + \alpha_{01}t + \alpha_{02}t^2 \\
\theta_1(t) &= \alpha_{10} + \alpha_{11}t + \alpha_{12}t^2 \\
\theta_2(t) &= \alpha_{20} + \alpha_{21}t + \alpha_{22}t^2,
\end{aligned}$$

The values of parameters α_{20}, α_{21} and α_{22} are most important in the simulation study as they determine, through $\theta_2(t)$, the difference between the linear and the quadratic models, (5.50) and (5.51). Therefore the parameters $\alpha_{00}, \alpha_{01}, \alpha_{02}, \alpha_{10}, \alpha_{11}$ and α_{12} were fixed while α_{20}, α_{21} and α_{22} were investigated.

The response function is observed at m points on each run and therefore the data generating model can be written as

$$Y_2 = GT + R_2, \tag{5.52}$$

where T is calculated from

$$T = \begin{pmatrix} \alpha_{00} & \alpha_{01} & \alpha_{02} \\ \alpha_{10} & \alpha_{11} & \alpha_{12} \\ \alpha_{20} & \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} 1 & \dots & 1 \\ t_1 & \dots & t_m \\ t_1^2 & \dots & t_m^2 \end{pmatrix}.$$

and $vec(R_2) \sim N(0, I \otimes \Sigma + J \otimes I\sigma_b^2)$ where J is the $n \times n$ matrix of ones. The variance-covariance matrix Σ has an autoregressive (AR1) autocorrelation structure defined as

$$\Sigma = \frac{\sigma_a^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \dots & \rho^{m-1} \\ \vdots & & & \vdots \\ \rho^{m-1} & \dots & \rho & 1 \end{pmatrix} \tag{5.53}$$

where σ_a^2 is the within run error variance. The between run error variance is denoted by σ_b^2.

The approximate optimal design in (5.45) was used to calculate an exact design, with n design points. The proportion of design points placed at each of the three support points, -1, 0 and 1 was determined by the weigths $w_1 = 0.25, w_2 = 0.5$ and $w_3 = 0.25$. Table 5.4 gives the values of n used in the simulation study, with the corresponding exact design for each number of runs.

n	-1	0	1
12	3	6	3
24	6	12	6
72	18	36	18

Table 5.4: Exact designs for various n

In this example, we use the functional F-statistic (5.18) to compare the residual sum of squares for the assumed first order model (5.50) to the residual sum of squares for the ANOVA model containing three parameters (one for each distinct value of x). Note that for this design with 3 distinct design points, the residual sum of squares for the ANOVA model is the same as that for the quadratic model.

The residual sum of squares for the linear model is calculated as

$$RSS = \sum_{i=1}^{n} \int \left(y_i(t) - f_i^T (F^T F)^{-1} F^T \boldsymbol{Y}(t)\right)^2 \mathrm{d}t, \tag{5.54}$$

with f_i^T defined as the ith row of the design matrix F. The integral is analytically intractable and hence approximated using Legendre-Gauss quadrature, with κ_j the roots of the Legendre polynomials and t_j the assumed abscissa values:

$$RSS \approx \sum_{i=1}^{n} \sum_{j=1}^{m_2} \kappa_j \left(y_i(t_j) - f_i^T (F^T F)^{-1} F^T \boldsymbol{Y}(t_j) \right)^2 . \tag{5.55}$$

Here we assume $m_2 = m$ and that we take observations at the quadrature points. The residual sum of squares for the ANOVA model is given by

$$RSS = \sum_{i=1}^{n} \int \left(y_i(t) - \bar{y}_i(t) \right)^2 \mathrm{d}t, \tag{5.56}$$

approximated by

$$RSS \approx \sum_{i=1}^{n} \sum_{j=1}^{m} \kappa_j \left(y_i(t_j) - \bar{y}_i(t_j) \right)^2 ,$$

where

$$\bar{y}_i(t) = \begin{cases} \frac{4}{n} \sum_{j=1}^{n/4} y_j(t) \text{ for } j = 1 \ldots \frac{n}{4} \\ \\ \frac{2}{n} \sum_{j=n/4+1}^{3n/4} y_j(t) \text{ for } j = \frac{n}{4} + 1 \ldots \frac{3n}{4} \\ \\ \frac{4}{n} \sum_{j=3n/4+1}^{n} y_i(t) \text{ for } j = \frac{3n}{4} + 1 \ldots n, \end{cases}$$

i.e. $\bar{y}_i(t)$ is the average of the replicated response from the corresponding support point.

We now describe four simulation studies to investigate how various aspects of these designs affect the power for discriminating between (5.50) and (5.51). We first outline the simulation algorithm, which was followed in the studies.

Algorithm

For each combination of values of the parameters investigated:

1. Generate a dataset of $n \times m$ responses from the quadratic model (5.52) according to the selected experimental design.

2. Fit the smaller model (5.50) and the ANOVA model to the generated data.

3. Calculate the residual sum of squares for each of the first order and ANOVA models.

4. Compute the test statistic (5.15).

5. Compare the value of the test statistic to the 95th percentile of the F distribution with $\hat{\lambda}(q-p)$ and $\hat{\lambda}(n-q)$ degrees of freedom, where $p = 2$ and $q = 3$ for the T-optimal design and $\hat{\lambda}$ is the degrees-of-freedom-adjustment-factor in (5.17).

6. Repeat steps 1-5 1000 times and calculate the power as the proportion of times that the null hypothesis, i.e. the smaller model is the 'true' model, is rejected.

We now describe the four studies and present the results.

Study 1: To investigate the effect of the parameters α_{20}, α_{21} and α_{22}, which determine $\theta_2(t)$, on the power. The parameters α_{20} and α_{21} were set to each of 0.5, 1 and 1.5 and α_{22} varied over the interval $[0, 2]$.

In the simulation study we assumed $\sigma_a^2 = 0.1$ and $\sigma_b^2 = 2$ to make the between run error variance much larger than the within run error variance, as usually occurs in practice,

The results of this study are shown in Figure 5.7. We see that as the number n, of design points, or runs, increases the power increases for 21 equally spaced values of α_{22} such that $0 \leq \alpha_{22} \leq 2$. Further, the larger the values of α_{20} and α_{21}, the larger the power over the whole range of α_{22}. This is to be expected as θ_2 increases for larger α_{22} and therefore it is easier to discriminate between the linear and quadratic models. Also note that for $\alpha_{21} = 1.5$, the power is close to 1 for $n = 72$. Fixing α_{20} and increasing α_{21} (across rows in Figure 5.7) has little or no effect on the power. This is explained by the form of the parameter function $\theta_2(t) = \alpha_{20} + \alpha_{21}t + \alpha_{22}t^2$ and that $-1 \leq t \leq 1$, resulting in α_{21} having no effect overall on the size of θ_2. For larger α_{20}, there is a smaller difference in power between the different numbers of runs because the discrimination problem is then easier.

Study 2: Influence of size of between run error variance, σ_b^2, on power

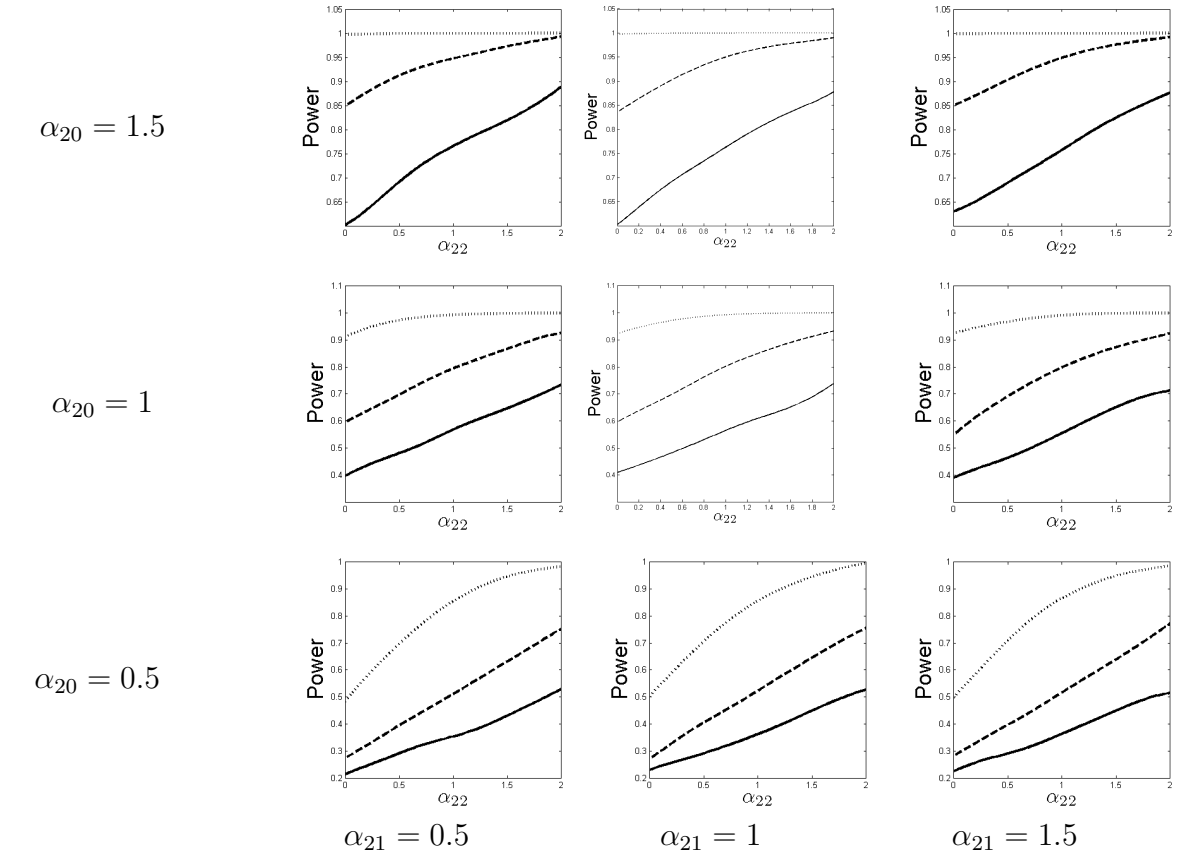

Figure 5.7: Example 1: Power calculated from 1000 simulations using the functional T-optimal design for 9 combinations of α_{20} and α_{21} values with $0 \leq \alpha_{22} \leq 2$, and number of runs $n = 12$ (–), $n = 24$ (- -) and $n = 72(\cdots)$..

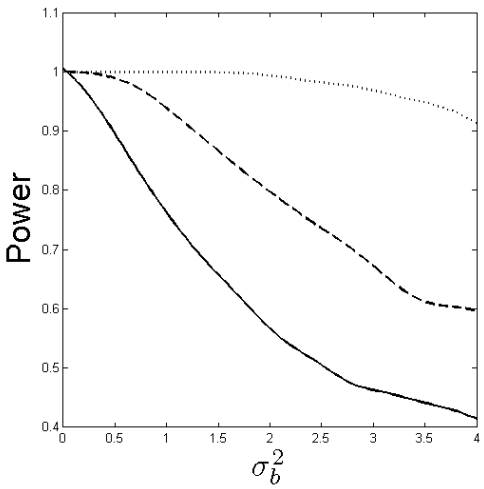

Figure 5.8: Example 1: Power calculated from 1000 simulations using the functional T-optimal design against between run error $0 \leq \sigma_b^2 \leq 4$ for $n = 12$ (–) , $n = 24$ (- -) and $n = 72$ $(\cdots)$

We carried out a further study to investigate how varying the parameter σ_b^2, the between run error variance, influences power. The other parameters were fixed at $\alpha_{20} = \alpha_{21} = \alpha_{22} = 1$, $\sigma_a^2 = 0.1$ and $\rho = 0.75$. For each of 17 equally spaced values of $0 \leq \sigma_b^2 \leq 4$, the algorithm was used to obtain 1000 responses.

A comparison of the three curves in Figure 5.8 shows that the power decreases as σ_b^2 increases. Also, the power decreases more quickly for smaller values of n. This is because a larger amount run-to-run error makes it harder to discern departures from the first order model.

Study 3: Power comparisons to alternative designs

We compared the power curve for the functional T-optimal design (5.45) to the curves for three possible alternative designs for discrimination between the first order and quadratic models, each having different numbers of support points:

(a) the D-optimal design for the quadratic model

$$\xi^a = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} \right\},$$

(b) the discrete uniform 4-point design on [-1,1] with equal weights

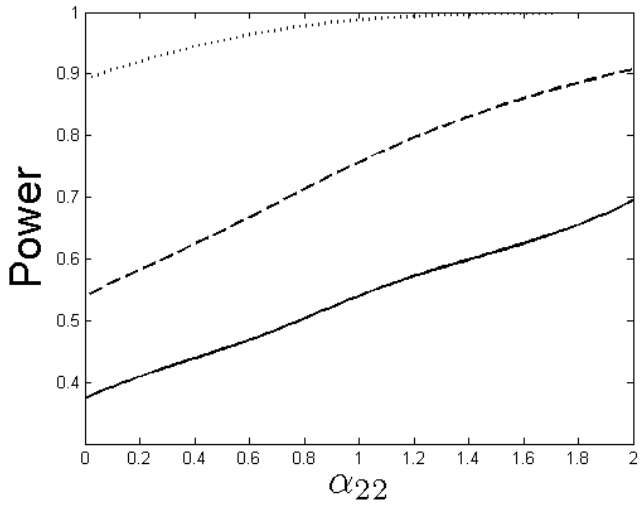

Figure 5.9: Example 1: Power against α_{22} from 1000 simulations for the D-optimal design, ξ^a, for $n = 12$ (–), $n = 24$ (- -) and $n = 72$ ($\cdots$).

$$\xi^b = \left\{ \begin{array}{cccc} -1 & -\frac{1}{3} & \frac{1}{3} & 1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{array} \right\},$$

(c) the discrete uniform 6-point design on [-1,1] with equal weights

$$\xi^c = \left\{ \begin{array}{cccccc} -1 & -0.6 & -0.2 & 0.2 & 0.6 & 1 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{array} \right\}.$$

We fixed the parameters $\alpha_{20} = \alpha_{21} = 1$, $\sigma_a^2 = 0.1$, $\sigma_b^2 = 2$ and $\rho = 0.75$ and, as before, investigated how the power changed for $n = 12, 24, 72$ runs as α_{22} varied. The results are shown in Figures 5.9, 5.10 and 5.11 for designs ξ^a, ξ^b and ξ^c respectively. The corresponding plot for the T-optimal design is the central plot of Figure 5.7. As expected, the power of the three alternative designs is generally lower than that of the T-optimal design, for each choice of number of runs. However, the power when the D-optimal design ξ^a, is used is very similar to that for the T-optimal design for all n investigated. The power was however, lower for $n = 12$. These results are explained by the fact that the D-optimal and T-optimal designs only differ in their weights.

Overall, the equally-spaced four-point design, ξ^b, and the equally spaced six-point design, ξ^c, have considerably lower power than the T-optimal design with maximum shortfalls of 21% and 37% for $n = 12$, 21% and 36% for $n = 24$, and 12% and 25% for $n = 72$ respectively. For $n = 72$, the power is reasonably high for both designs and increases almost to 1 for the highest values of α_{22}. Both designs performed similarly for each value

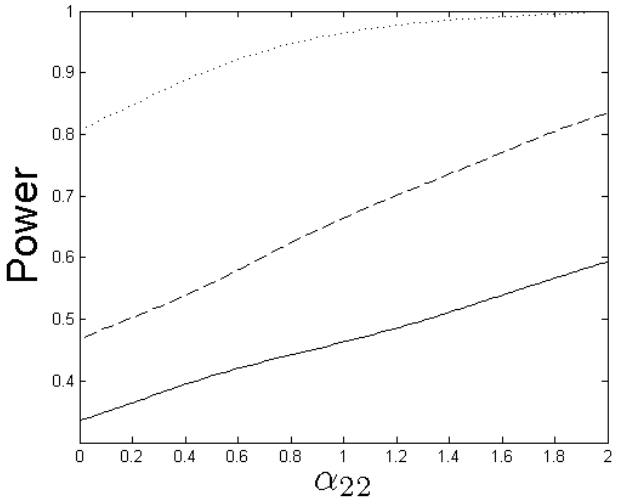

Figure 5.10: Example 1: Power against α_{22} from 1000 simulations for design ξ^b, for $n = 12$ (–), $n = 24$ (- -) and $n = 72$ ($\cdots$).

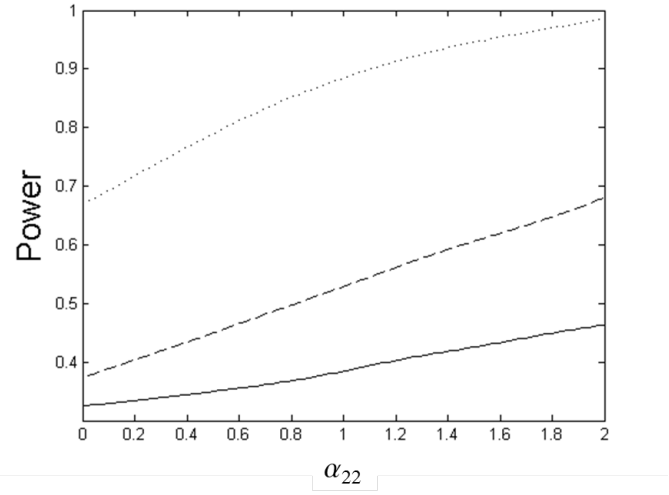

Figure 5.11: Example 1: Power against α_{22} from 1000 simulations for design ξ^c, $n = 12$ (–), $n = 24$ (- -) and $n = 72$ ($\cdots$).

.

of n investigated.

Study 4: To investigate an alternative error structure

We performed a limited investigation into the impact on power of changing the variance covariance matrix and the selection of m_2 points used to calculate the residual sum of squares (5.55). The m_2 points here are chosen to be distinct from the m points where observations are made which were used previously in (5.55). Instead of using the autoregressive correlation structure (5.53), as in the previous studies, we generated the data using a covariance function, which is often used in practice.

$$r(t,s) = Cov(\eta(t), \eta(s)) = \exp\left[-|(2(t-s))^2|\right] \tag{5.57}$$

for $t, s \in [-1, 1]$.

We wish to calculate the residual sum of squares using m_2 points where the response has not been observed. To obtain predictions for the m_2 points, we used cubic spline interpolation. The predictions may then be used to estimate the covariance function $r(s,t)$, where the jkth element of the corresponding variance-covariance matrix is given by

$$\hat{r}(t_j, t_k) = \frac{1}{n-p}\left(\boldsymbol{Y}_{t_j} - F(F^TF)^{-1}F^T\boldsymbol{Y}_{t_j}\right)^T\left(\boldsymbol{Y}_{t_k} - F(F^TF)^{-1}F^T\boldsymbol{Y}_{t_k}\right),$$

with $\boldsymbol{Y}_{t_j}$ as the t_jth column of the matrix Y, generated from (5.52). The matrix of errors R_2, used to generate Y, satisfies $\text{vec}(R_2) \sim N(0, I \otimes \Sigma + J \otimes I\sigma_b^2)$ where Σ has entries defined by (5.57) and J is an $n \times n$ matrix of ones.

We no longer use quadrature to estimate the integrals in the residual sum of squares formula for the assumed and ANOVA models given in (5.54) and (5.56). Therefore, we no longer need weights and hence the values z_j in (5.55) are set to one.

The vectors of m_2 values at which we predict the response clearly affect the estimated covariance structure. We considered three different vectors and investigated how they affected the power. We set $m_2 = 2m$ and used

(a) Equally spaced points on the interval $[-1, 1]$

(b) 39 points equally spaced points on $[-1, 0]$ and one point at -1

(c) 39 points equally spaced on $[-1, 0.5]$ and one point at -1.

The parameter values of $\alpha_{20} = \alpha_{21} = \alpha_{22} = 2.5$ $\sigma_a^2 = 0.1$, $\sigma_b^2 = 2$ and $\rho = 0.75$ were fixed in the simulation which followed the algorithm described earlier with $n = 32$.

The results showed that the power was largest, 0.7060, when the $m_2 = 40$ prediction points were equally spaced over $[-1, 1]$. The power obtained using the prediction points in (b) and (c) were lower at 0.3510, and 0.5040 respectively. We also calculated the average of the 1000 values of the degrees of freedom adjustment factors , $\hat{\lambda}$, for (a)-(c), which were very similar for each case: (a) 1.49, (b) 1.46, (c) 1.49. We might expect the power to increase with the value of $\hat{\lambda}$ because larger $\hat{\lambda}$ leads to larger degrees of freedom and a smaller critical value for the functional F-test. This may be difficult to establish from this simulation since we calculate an average over the 1000 simulations.

Changing the method of estimating the covariance structure allowed us to gain some information on the influence of the location of the prediction points on the power of the test to discriminate between the linear and quadratic functional models. The study which varied the prediction points was very limited and could be further explored in future work.

5.8.2 Example 2

In this example, we investigate the power of the test to reject a two-factor model with interaction (model 1), when the data comes from a two-factor model with an interaction and both quadratic terms.

This example considers two models where model 1 and model 2 differ in two terms. We assess how the T-optimal designs change according to the parameter values in model 2 and also the degree of correlation, ρ. We then use simulation to investigate the effect of the number of runs on the power of the test. We also investigated some alternative designs and calculated their power in order to compare their performance to that of the functional T-optimal design.

Specifically we test the hypothesis

$$
\begin{aligned}
H_0 : y(t) &= y_1(t) = \beta_0(t) + \beta_1(t)x_1 + \beta_2(t)x_2 + \beta_{12}x_1^T x_2 + \epsilon(t) \\
H_1 : y(t) &= y_2(t) = \theta_0(t) + \theta_1(t)x_1 + \theta_2(t)x_2 + \theta_{12}x_1 x_2 \\
&+ \theta_{11}(t)x_1^2 + \theta_{22}(t)x_2^2 + \eta(t).
\end{aligned}
$$

We assume a linear model for each run i

$$
\text{model 1:} \quad y_{1i}(t) = \beta_0(t) + \beta_1(t)x_{1i} + \beta_2(t)x_{2i} + \beta_{12}x_{1i}x_{2i} + \epsilon_i(t), \tag{5.58}
$$

but simulate a response from the model

$$
\begin{aligned}
\text{model 2:} \quad y_{2i}(t) &= \theta_0(t) + \theta_1(t)x_{1i} + \theta_2(t)x_{2i} + \theta_{12}x_{1i}x_{2i} \\
&+ \theta_{11}(t)x_{1i}^2 + \theta_{22}(t)x_{2i}^2 + \eta_i(t).
\end{aligned} \tag{5.59}
$$

We assume an AR1 auto-regressive covariance function where $\epsilon_i(t), \eta_i(t) \sim N(0, \sigma^2)$ and $Cov(\epsilon_i(t_j), \epsilon_i(t_k)) = Cov(\eta_i(t_j), \eta_i(t_k)) = \rho^{|j-k|}$ for t_j and t_k on some real interval. In this study, we have $t_j, t_k \in [0, 1]$ to ensure easier assessment of the effect of the linear terms in functions $\theta_{11}(t)$ and $\theta_{22}(t)$ defined below.

In a similar way to Example 1, the parameter functions $\theta_0(t), \theta_1(t), \theta_2(t), \theta_{12}(t), \theta_{11}(t), \theta_{22}(t)$ are defined in terms of quadratic functions

$$
\begin{aligned}
\theta_0(t) &= \alpha_{00} + \alpha_{01}t + \alpha_{02}t^2 \\
\theta_1(t) &= \alpha_{10} + \alpha_{11}t + \alpha_{12}t^2 \\
\theta_2(t) &= \alpha_{20} + \alpha_{21}t + \alpha_{22}t^2 \\
\theta_{12}(t) &= \alpha_{30} + \alpha_{31}t + \alpha_{32}t^2 \\
\theta_{11}(t) &= \alpha_{40} + \alpha_{41}t + \alpha_{42}t^2 \\
\theta_{22}(t) &= \alpha_{50} + \alpha_{51}t + \alpha_{52}t^2.
\end{aligned}
$$

The parameter values $\boldsymbol{\alpha_4} = (\alpha_{40}, \alpha_{41}, \alpha_{42})^T$ and $\boldsymbol{\alpha_5} = (\alpha_{50}, \alpha_{51}, \alpha_{52})^T$ determine, through

θ_{11} and θ_{22}, the difference between model 1 and model 2. Corollary 5.1 shows that the functional T-optimal design depends only on the subset of terms appearing in model 2, but not model 1. Hence $\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \boldsymbol{\alpha_3}$ have no influence on the choice of design. We therefore fix them to be $(1, 1, 1)$.

The response is observed at m points per run and therefore the model can be written as

$$Y_2 = GT + R_2, \tag{5.60}$$

where Y_2 is an $n \times m$ response matrix and T is calculated from

$$T = \begin{pmatrix} \alpha_{00} & \alpha_{01} & \alpha_{02} \\ \alpha_{10} & \alpha_{11} & \alpha_{12} \\ \alpha_{20} & \alpha_{21} & \alpha_{22} \\ \alpha_{30} & \alpha_{31} & \alpha_{32} \\ \alpha_{40} & \alpha_{41} & \alpha_{42} \\ \alpha_{50} & \alpha_{51} & \alpha_{52} \end{pmatrix} \begin{pmatrix} 1 & \dots & 1 \\ t_1 & \dots & t_m \\ t_1^2 & \dots & t_m^2 \end{pmatrix},$$

and $\text{vec}(R_2) \sim N(0, I \otimes \Sigma + J \otimes I\sigma_b^2)$ where J is an $n \times n$ matrix of ones. The variance-covariance matrix Σ has an autoregressive (AR1) autocorrelation structure defined in (5.53).

5.8.2.1 Optimal design

We found functional T-optimal designs as described in Section 5.7.2.2 using the Nelder Mead simplex algorithm. We tried a variety of different $\boldsymbol{\alpha_4}$ and $\boldsymbol{\alpha_5}$ values to see whether the parameter choices affected the design. To investigate the effect of ρ on the T-optimal designs, we conducted a small study in which $\boldsymbol{\alpha_4}$ and $\boldsymbol{\alpha_5}$ were fixed and found designs numerically for a variety of ρ values. The results (not shown) indicated that the value of ρ had little effect on the choice of optimal design, providing some evidence that the T-optimal design is robust to the degree of correlation.

Case 1

We fixed either $\boldsymbol{\alpha_4}$ or $\boldsymbol{\alpha_5}$ and varied the elements in the other parameter vector. In the case where all the parameters in $\boldsymbol{\alpha_4}$ and $\boldsymbol{\alpha_5}$ were positive, the T-optimal approximate

design was found to be

$$\xi = \left\{ \begin{array}{ccccc} (-1,-1) & (1,-1) & (0,0) & (-1,1) & (1,1) \\ 0.1250 & 0.1250 & 0.5 & 0.1250 & 0.1250 \end{array} \right\}. \tag{5.61}$$

We found that the functional T-optimal design and objective function values (5.43) for the parameter vectors $-\boldsymbol{\alpha_4}$ and $-\boldsymbol{\alpha_5}$ were the same as those for the parameter vectors $\boldsymbol{\alpha_4}$ and $\boldsymbol{\alpha_5}$, for any choice of entries in $\boldsymbol{\alpha_4}$ and $\boldsymbol{\alpha_5}$. For example, when $\boldsymbol{\alpha_4} = (6,5,4)^T$ and $\boldsymbol{\alpha_5} = (5,5,3)^T$ the value of the objective function was 5.8993. The same value was obtained when $\boldsymbol{\alpha_4} = (-6,-5,-4)^T$ and $\boldsymbol{\alpha_5} = (-5,-5,-3)^T$.

Case 2

We fixed either $\boldsymbol{\alpha_4}$ or $\boldsymbol{\alpha_5}$ to be positive and set the other parameter vector to be negative. In this case we found the T-optimal design to have the form

$$\xi = \left\{ \begin{array}{ccccc} (0,-1) & (-1,0) & (1,0) & (0,1) & (\tilde{x_1},\tilde{x_2}) \\ 0.25\gamma & 0.25\gamma & 0.25\gamma & 0.25\gamma & (1-\gamma) \end{array} \right\}, \tag{5.62}$$

where $(\tilde{x_1},\tilde{x_2})$ has $(\tilde{x_1},\tilde{x_2}) \in [-1,1]$ with $\gamma \approx 0.9999$ so that weight $(1-\gamma)$ is very small. This design would be very poor in practice as most realised exact designs would only have four support points unless n is very large, and four support points is not enough to test whether model 1 is the 'true' model.

We found that the functional T-optimal design and objective function values (5.43) for the parameter vectors $-\boldsymbol{\alpha_4}$ and $\boldsymbol{\alpha_5}$ were the same as those for the parameter vectors $-\boldsymbol{\alpha_4}$ and $\boldsymbol{\alpha_5}$, for any choice of entries in $\boldsymbol{\alpha_4}$ and $\boldsymbol{\alpha_5}$. For example, when $\boldsymbol{\alpha_4} = (6,5,4)^T$ and $\boldsymbol{\alpha_5} = (-5,-5,-3)^T$, and $\boldsymbol{\alpha_4} = (-6,-5,-4)^T$ and $\boldsymbol{\alpha_5} = (5,5,3)^T$ the optimal design was (5.62) and the value of the objective function was 5.8993.

Case 3

We set either $\boldsymbol{\alpha_4} = \mathbf{0}$ or $\boldsymbol{\alpha_5} = \mathbf{0}$, resulting in factor 1 or 2, respectively, being deleted from model 2. In both cases, the optimal design has 15 support points. If $\boldsymbol{\alpha_4} = \mathbf{0}$ then the design has weight 0.25 at design points (-1,$\star$), weight 0.5 at design points (0,$\star$) and weight 0.25 at design points (1,$\star$), where $\star$ is a defining any level of factor 2. Note that the projection of this design onto the first factor gives the T-optimal design (5.45).

The power for each example was calculated by following the algorithm described earlier

except for steps 1, 2 and 5 which are:

1. Generate responses from model 2 (5.60)

2. Fit the first order model (5.58) and the ANOVA model

5. Compare the test statistic to the reference distribution, an F distribution with $\hat{\lambda}(q-p)$ and $\hat{\lambda}(n-q)$ degrees of freedom where $p=4$ and $q=6$ for the optimal design and $\hat{\lambda}$ is the degrees-of-freedom-adjustment-factor in (5.17)

We investigated 8 different values of n and 6 combinations of parameter vectors $\boldsymbol{\alpha_4}$ and $\boldsymbol{\alpha_5}$, where all parameters were chosen to be positive. Hence the T-optimal design was given by (5.61). The other parameters were kept fixed with $m=20$, $\sigma_a^2=0.1$ and $\sigma_b^2=2$ and $\rho=0.75$. We calculated the power for three designs:

(a) the T-optimal design (5.61)

(b) a D-optimal design for model 2

$$\xi^b = \left\{ \begin{array}{ccccccccc} (-1,-1) & (0,-1) & (1,-1) & (-1,0) & (0,0) & (1,0) & (-1,1) & (0,1) & (1,1) \\ 0.146 & 0.080 & 0.146 & 0.080 & 0.096 & 0.080 & 0.146 & 0.080 & 0.146 \end{array} \right\}.$$

(c) the 9-point design:

$$\xi^c = \left\{ \begin{array}{ccccccccc} (-1,-1) & (0,-1) & (1,-1) & (-1,0) & (0,0) & (1,0) & (-1,1) & (0,1) & (1,1) \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \end{array} \right\}.$$

5.8.2.2 Results

Figure 5.12 shows that in all cases the functional T-optimal design had higher power to reject model 1. We also see that examples with larger values of parameters, $\boldsymbol{\alpha_4}$ and $\boldsymbol{\alpha_5}$, had larger power, e.g. in the bottom row of Figure 5.12, and the power generally increased with n. The trend in Example 2 is the same as that seen in Example 1, Study 1. Larger

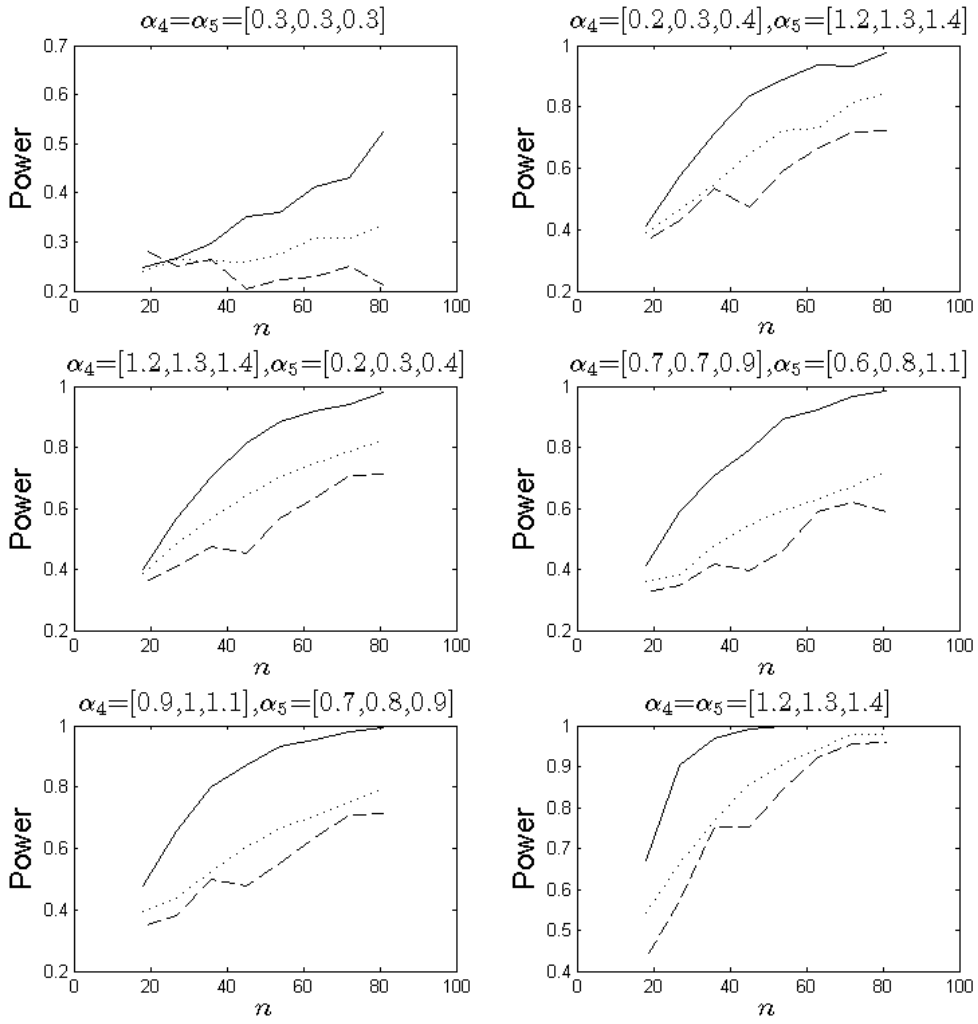

Figure 5.12: Example 2: Power against the number of runs, n from 1000 simulations for three designs: T-optimal design in (5.61) (–); D-optimal design for model 2 (- -); 9-point equally weights design $(\cdots)$ and six choices of $\boldsymbol{\alpha_4}$ and $\boldsymbol{\alpha_5}$.

parameters in $\tilde{\theta}$ indicated a larger difference in the two models being compared. In all cases the D-optimal design for model 2 out performed the equally spaced nine point design. It is important to note that the power for the D-optimal design is variable in all plots. The weights do not give consistent proportions of design points once we have rounded to get an integer number of runs. Therefore, the exact designs, used to calculate the power, change quite substantially for these values of n. However, we did investigate larger n for this design and we found that as expected the power did increase slowly with n.

5.9 Concluding remarks

In this chapter we have developed novel results on optimal designs for discriminating between two functional linear models when, for each response variable, observations made on the same run may be correlated. Specifically we have:

- proposed a T-optimality criterion for discriminating between two such models
- proved that, if two linear models differ by only one term, then the same design is T-optimal for discriminating between pairs of models that are both univariate or both multivariate or both functional
- established analytically a T-optimal design for discriminating between a first and second order functional linear model with a single explanatory variable.

We have assessed the power of the test for discriminating between the above first and second order functional linear models via simulation studies. We have also found numerically designs to discriminate between two functional linear models that differ by more than one term, and compared their performances. Simulation studies showed that the power resulting from the use of a T-optimal design was greater than that from competing designs, including a D-optimal design.

We also carried out some investigations on how the power varied according to the choice of prediction points. However, this work was limited and there is scope to extend these ideas to further investigate the problem.

Chapter 6

Conclusions and future work

6.1 Conclusions

In this thesis we have investigated two aspects of experimental design for functional data. First, the selection of points at which to take observations in order to reconstruct the functional response from a single run of the experiment using nonparametric techniques. Secondly, the choice of points that enable effective discrimination between two functional linear models.

In Chapters 3 and 4, we found optimal designs to 'best' predict a smooth function g using criteria based on prediction variance. We considered two different methods of nonparametric prediction using the local linear and Gasser and Müller estimators. Chapter 5 developed theory and methods for finding designs to enable discrimination between two functional linear models by testing the fit of one model given data generated from the alternative model.

6.1.1 Optimal design for nonparametric prediction of a curve

The aim of the work in both Chapters 3 and 4 was to find ordered design points to enable nonparametric prediction. In both chapters we investigated a variety of methods and different optimality criteria.

Chapter 3 found optimal designs which minimised a compound D_s-optimality criterion for prediction across a specified interval. Initially, we found designs for prediction at a finite number of points which were then generalised to optimal designs for prediction

across an interval. Application of designs found using the Gaussian kernel were then demonstrated using data from the tribology experiment. A study, drawing on the tribology data, indicated that these designs performed similarly to equally spaced designs in enabling a model which was a good fit to the data to be found, where mean squared error was used to measure goodness of fit. In the tribology application the choice of bandwidth was difficult without prior knowledge of the curve. The use of different bandwidths on different intervals may have achieved a better fit for predicting different sections of the response.

The work in Chapter 4 found designs by trading-off integrated prediction variance against the complexity of the fitted model as quantified by the trace of the smoothing matrix. We minimised an objective function that was a weighted sum of these two components. This enabled designs to be tailored to different complexities of models to be fitted in the data analysis. We conducted a robustness study to investigate the effect of misspecifying the kernel.

6.1.2 T-optimal designs for functional linear models

In Chapter 5 we obtained the first results on T-optimal designs for functional linear models. We showed that the choice of an optimal design for discriminating between two nested functional linear models, which differ by only one term, is independent of the parameter values in each model and the correlation structure of each of the functional responses, see Proposition 5.3 of Section 5.7.2.3. Where the models differ by more than one term, the design depends on the parameters for the additional terms in the larger model (Corollary 5.1).

The T-optimal designs were then used in simulation studies to calculate the power of the test for assessing the fit of model 1 given data obtained from model 2 for two specific examples, where model 1 is nested in model 2. We found that tests had larger power for larger numbers of runs and smaller between run errors led to larger power. Another intuitive result was that the power increased when the parameters for the additional terms in model 2 were larger. The correlation structure and the location of the prediction points were briefly investigated in the power studies; there is scope to investigate further these influences on the choice of optimal design.

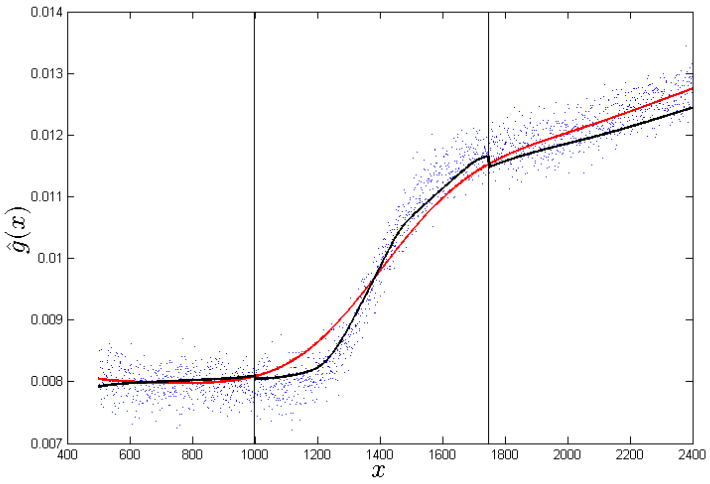

Figure 6.1: Smooth fits using a design for run 19 (a) data from the optimal design with varying h, $h = 0.2$ on $[501, 1000]$ and $[1751, 2400]$ and $h = 0.1$ on $[1001, 1750]$ (black) (b) whole dataset with $h = 0.1$ (red)

6.2 Future Work

6.2.1 Optimal design for local linear regression

6.2.1.1 Varying the bandwidth in local linear regression

In Section 3.5.3 we introduced the idea of allowing for a variable bandwidth in design selection. This would be appropriate when it is anticipated that data to be collected will have features such as a turning point or a point of inflection. Allowing the bandwidth h to have different values for different ranges of x allows information on complexity to be introduced into model (1.1) in a similar way to constraining the complexity of the smooth fit, as seen in Section 4.3. We require a small value of h to predict complex features in our data. For example, the prediction of the data in run 19 of the tribology experiment (Figure 6.1) would benefit from a smaller bandwidth on the interval $[1001, 1750]$. On the other hand, a larger bandwidth is required for the remaining parts of the interval so the data are not oversmoothed.

We explore these issues for simulated data from run 19, with $\epsilon_j \sim N(0, 0.00015^2)$ and ϵ_j, ϵ_k independent for $j, k = 1, ..., n$, using an optimal design found by Criterion 3.3 and the search method in Chapter 3, for the following varying bandwidth: $h = 0.1$ for $1001 \leq x \leq 1750$ and $h = 0.2$ elsewhere. Figure 6.1 shows the resulting fit from the data generated

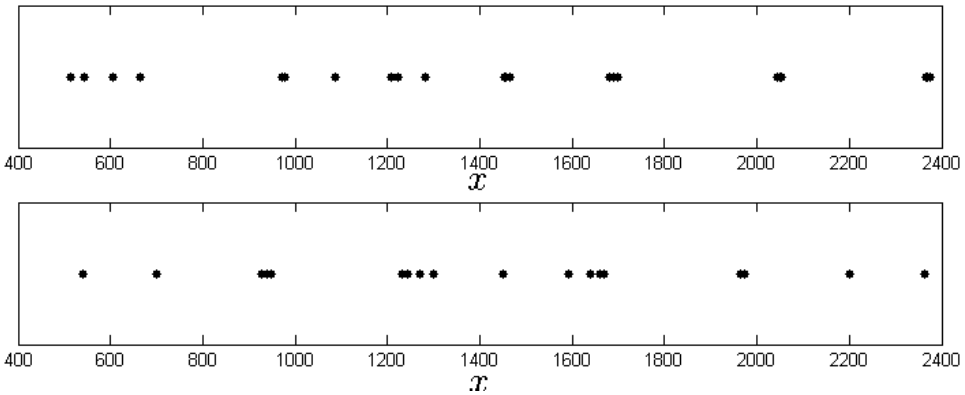

Figure 6.2: Designs for constant bandwidth with $h = 0.2$ (bottom) and varying bandwidth: $h = 0.1$ on $[1001, 1750]$ and $h = 0.2$ otherwise (top)

via the optimal design found for varying bandwidth

$$[513, 544, 605, 665, 973, 973, 979, 1086, 1208, 1223, 1283, 1452, 1457,$$

$$1464, 1680, 1689, 1690, 1697, 2045, 2050, 2051, 2052, 2367, 2369, 2375].$$

The constant bandwidth design with $h = 0.2$ has points

$$[539, 539, 539, 700, 928, 928, 938, 947, 1232, 1242, 1270, 1299, 1451,$$

$$1593, 1640, 1659, 1669, 1963, 1963, 1963, 1973, 2201, 2362, 2362, 2362].$$

The difference in distributions of the points in each design are shown in Figure 6.2. The design obtained from prediction using a variable bandwidth had two more points in the interval $[1001, 1750]$. There are also more points closer to the centre of the interval for the varying bandwidth, as expected, due to the more complex fit enforced by $h = 0.1$.

Figure 6.1 also shows the smooth fit obtained from the whole dataset and bandwidth $h = 0.2$. The fit using data from the design is better for prediction over $[1001, 1750]$ and does not oversmooth the remaining data. However, there is a discontinuity where the bandwidth changes at $x = 1001$ and $x = 1751$.

An interesting problem for the future is the need to find an appropriate method of avoiding this sudden change in bandwidth and then to develop efficient designs for this type of data analysis. One possible method is block-wise least squares parabolic fitting, introduced by Härdle and Marron (1995) which sets a bandwidth for each block and then smooths the bandwidth over the blocks using local linear regression techniques.

6.2.1.2 Correlated errors

In this thesis, we have found optimal designs to enable the 'best' prediction using the local linear estimator with the assumption (Section 2.1) that the error variables are independent. When observations have a natural ordering, e.g. over time, this assumption may not hold (Simonoff, 1996). It is important that if data are likely to be correlated, then this feature is incorporated into the model, as it affects the choice of bandwidth (Opsomer, Wang and Yang, 2001).

In general, if we make no assumptions about either the form of the mean function g or the correlation structure then it is impossible to estimate either function separately (Opsomer et al., 2001). Therefore, to find optimal designs for unknown g we would have to make an assumption about the correlation structure. A simple starting point would be to assume the errors follow an AR(1) process. Designs could also be found for other correlation structures and the robustness of designs to different types of correlation could be investigated.

6.2.2 Designs to minimise the integrated variance subject to a constraint

In Chapter 4, we developed a new criterion, which was applied to designs under the Gasser and Müller estimator. The criterion minimised a weighted sum of the integrated variance and the inverse trace of the smoothing matrix. A sensible extension would be to find designs using this criterion for the local linear estimator. We would then be able to compare these designs with those found for the Gasser and Müller estimator. In addition, this comparison would give some indication of how D_s-optimal designs from Chapter 3 differ from designs obtained via the new constrained criterion in Chapter 4.

Another possible avenue of future work would be to extend methods in Chapter 3 and 4 to find designs for prediction for more than one variable. In particular, it would be interesting to find designs using Criterion 4.4 for a multi-variable Gaussian process model (Rasmussen and Williams, 2006).

6.2.3 Further work on T-optimality for functional linear models

It would be useful to gain a more general understanding of T-optimal designs for discriminating between two functional linear models through expanding the range of models for

which designs are found in Chapter 5. In particular, more complex models and how the choice of covariance structure influences the design when the models differ by more than one term could both be considered.

In Chapter 5 we briefly discussed the effect of the choice of the set of points where predictions are made for each run on the power of the test for rejecting a model given data from a competing model. We considered three simple sets of points. This research could be taken further by incorporating the work in Chapters 3 and 4 to create a two stage design problem. Initially, we find for each run an optimal design for predicting the functional response. Then these designs can be used as the prediction points in the T-optimality power studies.

We could also construct an example using a factorial experiment in order to assess how the optimal design and form of $\hat{g}$ change with different treatments. We may wish to investigate how the treatment affects the choice of h used in finding an optimal set of points at which to observe the functional response.

Bibliography

Arfken, G. B., Weber, H. J. and Harris, F. (2012) *Mathematical Methods for Physicists*. Waltham, MA: Academic Press, 7th edn.

Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007) *Optimum Experimental Designs, with SAS*. Oxford: Oxford University Press, 2nd edn.

Atkinson, A. C. and Fedorov, V. V. (1975) The design of experiments for discriminating between two rival models. *Biometrika*, **62**, 57–70.

Biedermann, S. and Dette, H. (2001) Minimax optimal designs for nonparametric regression - a further optimality property of the uniform distribution. In *6th International Workshop on Model-Oriented Design and Analysis* (eds. A. C. Atkinson, P. Hackl and W. G. Müller), 13–20.

Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear smoothers and additive-models. *Annals of Statistics*, **17**, 453–510.

Cardot, H., Ferraty, F. and Sarda, P. (1999) Functional linear model. *Statistics &; Probability Letters*, **45**, 11 – 22.

Cardot, H., Goia, A. and Sarda, P. (2004) Testing for no effect in functional linear regression models, some computational approaches. *Communications in Statistics- Simulation and Computation*, **33**, 179–199.

Chatfield, C. (2004) *The Analysis of Time Series: An Introduction*. London: Chapman and Hall, 6th edn.

Cheng, M. Y., Hall, P. and Titterington, D. M. (1998) Optimal design for curve estimation by local linear smoothing. *Bernoulli*, **4**, 3–14.

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Clyde, M. and Chaloner, K. (1996) The equivalence of constrained and weighted designs in multiple objective design problems. *Journal of the American Statistical Association*, **91**, 1236–1244.

Cuevas, A., Febrero, M. and Fraiman, R. (2004) An ANOVA test for functional data. *Computational Statistics & Data Analysis*, **47**, 111 – 122.

Diggle, P. J., Heagerty, P. J., Liang, K. and Zeger, S. L. (2002) *Analysis of Longitudinal Data*. Oxford: Oxford University Press, 2nd edn.

Draper, N. and Smith, H. (1998) *Applied Regression Analysis*. New York: Wiley, 3rd edn.

Eubank, R. L. (1999) *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker, 2nd edn.

Fan, J. and Gijbels, I. (1995) Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society Series B*, **57**, 371–394.

Fan, J. and Lin, S.-K. (1998) Test of significance when data are curves. *Journal of the American Statistical Association*, **93**, 1007–1021.

Fan, J. Q. (1992) Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.

Faraway, J. J. (1997) Regression analysis for a functional response. *Technometrics*, **39**, 254–261.

Fedorov, V. V. and Hackl, P. (1997) *Model-Oriented Design of Experiments*. New York: Springer.

Fedorov, V. V., Montepiedra, G. and Nachtsheim, C. J. (1999) Design of experiments for locally weighted regression. *Journal of Statistical Planning and Inference*, **81**, 363–382.

Gasser, T. and Müller, H.-G. (1979) Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation* (eds. T. Gasser and M. Rosenblatt), 23–68. Heidelberg: Springer.

— (1984) Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, **11**, 171–185.

Golub, G. H. and Welsch, J. H. (1969) Calculation of Gauss quadrature rules. *Mathematics of Computation*, **23**, 221–230 s1–s10.

Härdle, W. and Marron, J. (1995) Fast and simple scatterplot smoothing. *Computational Statistics & Data Analysis*, **20**, 1–17.

Johnson, R. A. and Wichern, D. W. (1998) *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice-Hall, 4th edn.

Lagarias, J. C., Reeds, J. A., Wright, M. H. and Wright, P. E. (1998) Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, **9**, 112–147.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*. London: Chapman and Hall.

McCulloch, C., Searle, S. and Neuhaus, J. (2008) *Generalized, Linear, and Mixed Models*. Hoboken, NJ: John Wiley and Sons, 2nd edn.

McLain, D. H. (1971) Drawing contours from arbitrary data points. *The Computer Journal*, **17**, 318–324.

Müller, W. G. (1992) Optimal design for moving local regressions, unpublished technical report. URL `http://epub.wu.ac.at/932/ (accessed 09/12/12)`.

— (1996) Optimal design for local fitting. *Journal of Statistical Planning and Inference*, **55**, 389–397.

Nadaraya, E. A. (1964) On estimating regression. *Theory of Probability and its Applications*, **10**, 186–190.

Nelder, J. and Mead, R. (1965) A simplex method for function minimization. *The Computer Journal*, **7**, 308–313.

Opsomer, J., Wang, Y. and Yang, Y. (2001) Nonparametric regression with correlated errors. *Statistical Science*, **16**, 134–153.

Pelto, C. R., Elkins, T. A. and Boyd, H. A. (1968) Automatic contouring of irregularly spaced data. *Geophyisics*, **33**, 424–430.

Priestly, M. B. and Chao, M. T. (1972) Nonparametric function fitting. *Journal of the Royal Statistical Society, Series B*, **34**, 384–392.

Ramsay, J., Hooker, G. and Graves, S. (2009) *Functional Data Analysis with R and MATLAB*. New York: Springer.

Ramsay, J. O. and Dalzell, C. J. (1991) Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society, Series B*, **53**, 539–572.

Ramsay, J. O. and Silverman, B. (2005) *Functional Data Analysis*. New York: Springer, 2nd edn.

Rasmussen, C. E. and Williams, C. (2006) *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Shen, Q. and Faraway, J. (2004) An F test for linear models with functional responses. *Statistica Sinica*, **14**, 1239–1257.

Shen, Q. and Xu, H. (2006) Diagnostics for linear models with functional responses. *Technometrics*, **49**, 26–33.

Simonoff, J. S. (1996) *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Uciński, D. and Bogacka, B. (2005) T-optimum designs for discrimination between two multiresponse dynamic models. *Journal of the Royal Statistical Society: Series B*, **67**, 3–18.

Wand, M. P. and Jones, M. (1995) *Kernel Smoothing*. London: Chapman and Hall.

Watson, G. S. (1964) Smooth regression analysis. *Sankhya A*, **26**, 101–116.

West, M., Harrison, P. J. and Migon, H. S. (1985) Dynamic generalized linear models and bayesian forecasting (with discussion). *Journal of the American Statistical Association*, **80**, 73–97.