Statistics and Applications
Volume 11, Nos. 1&2, 2013 (New Series), pp. 33-54

Optimal design for prediction using local linear regression and the D_{SI}-criterion

Verity A. Fisher, David C. Woods and Susan M. Lewis
University of Southampton, Southampton, UK

Abstract When it is anticipated that data to be collected from an experiment cannot be adequately described by a low-order polynomial, alternative modelling and new design methods are required. Local linear regression, where the response is approximated locally by a series of weighted linear regressions, is an effective nonparametric smoothing method that makes few assumptions about the functional form of the response. We present new methods for the optimal design of experiments for local linear regression, including a new criterion, called D_{SI}-optimality, to find designs that enable precise prediction across a continuous interval. Designs are found numerically for weights defined through the Gaussian and uniform kernels. Theoretical results are presented for the uniform kernel and the special case of prediction at a single point. The sensitivity of the designs to the choice of bandwidth in the local linear regression is studied, and it is found that designs for the Gaussian kernel with large bandwidth have a small number of distinct design points. The methodology is motivated by, and demonstrated on, an experiment from Tribology.

Key words: D_S-optimality, linear smoothing, nonparametric regression, weighted least squares.

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. For such data, the more flexible approach of nonparametric regression is preferred where fewer assumptions are required on the functional form of the response.

Previous research (Müller, 1992, 1996; Fedorov, Montepiedra and Nachtsheim, 1999) has found designs for local linear regression tailored to prediction at a finite set of points within the design region. The research in this paper gives, for the first time, methods for the efficient design of experiments for prediction of the whole curve across a continuous interval using local linear regression. We introduce a new design selection criterion, D_{SI}-optimality,

and present properties of the resulting designs, together with some examples. The work is motivated by, and demonstrated on, an experiment from Tribology.

1.1 Linear smoothing

Assume the response, y, depends on a single explanatory variable, x, through

$$y_j = g(x_j) + \varepsilon_j \,, \tag{1}$$

where $g(\cdot)$ is an unknown function, x_j and y_j are the jth design point and associated response, respectively, and ε_j is an error random variable with constant variance σ^2 ($j = 1, \ldots, n$). Further, we assume ε_i and ε_j are independent.

The function $g(x)$ is estimated by a linear smoother $\hat{g}(x)$, defined through the following weighted linear combination of the observations y_j,

$$\hat{g}(x) = \sum_{j=1}^{n} S_j(x) y_j \,. \tag{2}$$

Here, $S_j(x)$ is a smoothing weight that determines the influence of y_j on a prediction at x (see, for example, Ramsay and Silverman, 2005, ch. 4). The simplest example of a linear smoother is simple linear regression. Further, more complex, examples can be found in Buja, Hastie and Tibshirani (1989), Wand and Jones (1995, ch. 5) and Simonoff (1996, ch. 5).

1.2 Tribology application

To motivate the methodology, we consider an experiment from the field of Tribology, which is the study of interacting surfaces in relative motion. The experiment was a pilot study to assess how different controllable factors affect the wear of a pin and disc assembly when the surface of the disc is lubricated by an oil.

Data from one run of the process was obtained by measuring the total wear of the disc and pin at a given number of time points after the disc has started spinning. The data from different runs of the process may differ due to random variation and the various different settings chosen for the controllable factors which included disc material, pin material and the addition of various contaminants to the lubricant. Data from two runs of the process are shown in Figure 1. Notice that wear can actually decrease at some time points due to a build up of contamination in the groove, or "wear scar", worn into the disc by the pin.

Fig 1: Data from run 1 (left) and run 2 (right) of the wear experiment with examples of a locally linear smooth fit (red).

In this paper, we find optimal designs for predicting curves such as those given in Figure 1 using only a small number of design points; that is, we choose the 'best' subset $\{x_1, \ldots, x_n\} \subset \mathbb{R}$ of points at which to observe the process. In Section 2 we describe prediction via the local linear estimator. Section 3 introduces optimal design for this prediction method and, in particular, D_{SI}-optimal designs for predicting the curve across an interval. Results are presented in Section 4 for the special case of prediction at a single point, as well as prediction across an interval. In Section 5, D_{SI}-optimal designs are found for a simulated wear experiment, motivated by the Tribology application. We make our concluding remarks in Section 6.

2 Local linear regression

We employ local regression methods (Pelto, Elkins and Boyd, 1968 and Cleveland, 1979) to estimate $g(\cdot)$. Local fitting weights the observations to ensure that points closer (or more local) to x have larger influence on $\hat{g}(x)$. For prediction at a point x^*, local weighted regression fits a pth degree polynomial using weighted least squares.

Suppose that the $(p+1)$th derivative of $g(x)$ exists in a small neighbourhood about the point

x^*. Then, from a Taylor series expansion of $g(x)$ about x^*,

$$\begin{aligned} g(x) &\approx g(x^*) + g^{(1)}(x^*)(x - x^*) + \frac{g^{(2)}(x^*)}{2}(x - x^*)^2 + ... + \frac{g^{(p)}(x^*)}{p!}(x - x^*)^p \\ &= \beta_0(x^*) + \beta_1(x^*)(x - x^*) + ... + \beta_p(x^*)(x - x^*)^p \,, \end{aligned} \tag{3}$$

where $g^{(p)}(x)$ denotes the pth derivative. We define $u = x - x^*$ to obtain $g(x) = \beta_0(x^*) + \beta_1(x^*)u+, \ldots, +\beta_p(x^*)u^p$. On setting $x = x^*$, it becomes clear that prediction of $g(x^*)$ is equivalent to estimation of $\beta_0(x^*)$.

Weights for each observation, y_j, are introduced via a symmetric kernel function $K(u; h)$ that satisfies $\int K(u; h)\,\mathrm{d}u = 1$ and depends on a pre-specified constant h, known as the bandwidth. This constant controls the degree of the locality in the prediction $\hat{g}(x^*)$. Estimators $\hat{\boldsymbol{\beta}}(x^*) = [\hat{\beta}_0(x^*), ..., \hat{\beta}_p(x^*)]^T$ are then found to minimise

$$\sum_{j=1}^{n} \frac{1}{h} \left[y_j - \beta_0(x^*) - \beta_1(x^*)u_j - ... - \beta_p(x^*)u_j^p \right]^2 K(u_j; h) \,, \tag{4}$$

where $u_j = x_j - x^*$ for $j = 1, \ldots, n$. When $p = 1$, we can write the solution to (4) as

$$\hat{\boldsymbol{\beta}}(x^*) = (X^T W X)^{-1} X^T W \boldsymbol{Y} \,, \tag{5}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, $W = \mathrm{diag}\{K(u_j; h)/h\}$ and $X = [\mathbf{1}, \boldsymbol{u}]$, with $\mathbf{1}^T = (1, \ldots, 1)$ and $\boldsymbol{u}^T = (u_1, \ldots, u_n)$ (Wand and Jones, 1995, p.114). The prediction is $\hat{g}(x) = \hat{\beta}_0(x^*) + \hat{\beta}_1(x^*)u$ and hence the prediction at x^*, when $u = 0$, is given by $\hat{g}(x^*) = \hat{\beta}_0(x^*)$, where

$$\hat{\beta}_0(x^*) = \frac{1}{nh} \frac{\sum_{j=1}^{n} \left\{ \hat{s}_2(x^*; h) - \hat{s}_1(x^*; h)u_j \right\} K(u_j; h) y_j}{\hat{s}_2(x^*; h)\hat{s}_0(x^*; h) - \{\hat{s}_1(x^*; h)\}^2} \,, \tag{6}$$

and $\hat{s}_r(x^*; h) = \sum_{k=1}^{n} u_k^r K(u_k; h)/(nh)$, for $r = 0, 1, 2$. From (2), the smoothing weights at x^* are given by

$$S_j(x^*) = \frac{1}{nh} \frac{\sum_{j=1}^{n} \left\{ \hat{s}_2(x^*; h) - \hat{s}_1(x^*; h)u_j \right\} K(u_j; h)}{\hat{s}_2(x^*; h)\hat{s}_0(x^*; h) - \hat{s}_1(x^*; h)^2} \,.$$

Note that when $K(u_j; h) = h$ for all u_j, then $\hat{\beta}_0(x^*)$ reduces to the ordinary least squares estimator of the intercept in a simple linear regression. A further special case, the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964), is obtained when $p = 0$. Fan (1992) established the advantages of the local linear smoother ($p = 1$) over the Nadaraya-Watson estimator in terms of expected squared error.

In this paper, we find designs for the uniform and Gaussian kernel functions:

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

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

We choose these kernels to demonstrate the differences in optimal designs for truncated and non-truncated kernel functions. Clearly, both kernel functions are non-increasing functions of $|u| = |x - x^*|$. For the uniform kernel, observations y with corresponding $|x - x^*| > h$ will not influence prediction at x. The Gaussian kernel function is monotonically decreasing with u^2 but monotonically increasing with h. These two kernels present some different issues for design; for example, with a truncated kernel such as the uniform, at least two design points are required to lie within distance h of x^* to enable the local linear smoother to be estimated. For both kernels, lower prediction variance is obtained when more design points lie within h of the point x.

3 Experimental design for local prediction

In this section, we define a design selection criterion, the D_{SI}-criterion, for prediction of $g(x^*)$ for $x^* \in I = [a, b] \subset \mathbb{R}$ using local linear smoothing. We find designs $\xi_n = \{x_1, ..., x_n\}$, with $x_1 \leq \ldots \leq x_n$.

Similar problems were considered by Müller (1992), Müller (1996) and Fedorov et al. (1999) for local linear regression. Müller (1992) and Müller (1996) found designs that minimised a weighted sum of variances of the estimator $\hat{\beta}_0(x_i)$, for prediction at points x_i^*, $i = 1, ..., q$:

$$\Psi_M(\xi) = \text{trace}\left\{\sum_{i=1}^{q} A_i (X_i^T W_i X_i)^{-1}\right\}\,, \tag{7}$$

where $A_i = a_i A$, a_i is a given scalar and $A = \boldsymbol{e}_1 \boldsymbol{e}_1^T$ with $\boldsymbol{e}_1$ the first column of a $p \times p$ identity matrix. The $p \times p$ matrix, $X_i^T W_i X_i$, is the information matrix for the linear model

$$\boldsymbol{Y} = X_i \boldsymbol{\beta}(x_i^*) + \boldsymbol{\eta}\,, \tag{8}$$

where $X_i = [\mathbf{1}_n, \boldsymbol{u}_i, \ldots, \boldsymbol{u}_i^p]$, $\boldsymbol{u}_i^T = (u_{i1}, \ldots, u_{in})$ and $u_{ij} = x_j - x_i^*$. The n-vector $\boldsymbol{\eta} \sim N(\mathbf{0}, W^{-1})$, with W the diagonal matrix of kernel weights. That is, (8) is the linear model that would lead to estimators (5) of $\boldsymbol{\beta}(x_i)$. Müller (1992) found designs by maximising (7) for examples when prediction was required at a set of $q = 9$ equally spaced points in $[-1, 1]$ and when x could only take values from this same set.

Fedorov et al. (1999) also found designs for local linear smoothing but adopted a mean squared error criterion where the true model was assumed to be a local linear regression of higher order. We follow Müller (1992) and make the assumption that sufficient prior information is available to choose an appropriate value of h to render the bias negligible.

Designs for alternative local smoothing methods have been considered by Cheng, Hall and Titterington (1998) and Biedermann and Dette (2001), among other authors.

3.1 D_{SI}-optimal designs for prediction

Our design selection criterion is motivated by model (8). Under this model, and using notation that suppresses the dependence of X and W on x^*, the variance-covariance matrix of $\hat{\boldsymbol{\beta}}(x^*)$ is given from equation (5) by

$$\begin{aligned} \text{Var}\left(\hat{\boldsymbol{\beta}}(x^*)\right) &= \sigma^2 (X^T W X)^{-1} X^T W W^{-1} W X (X^T W X)^{-1} \\ &= \sigma^2 (X^T W X)^{-1}, \end{aligned} \tag{9}$$

with corresponding information matrix $M(\xi_n) = X^T W X$.

For the prediction of $g(x^*)$, interest is only in $\beta_0(x^*)$. Hence, we apply D_S-optimality (see, for example, Atkinson et al., 2007), which selects designs that minimise the variance of a subset of model parameter estimators whilst regarding the remaining parameters as nuisance parameters. The information matrix can be partitioned as

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

In general, the D_S-optimality criterion seeks a design that minimises the determinant of $M_{11}^{-1}(\xi_n)$ or, equivalently, maximises

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

For local linear regression, and estimation of $\beta_0(x^*)$,

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

and

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

Hence, a D_S-optimal design for prediction at a point x^* for local linear regression maximises the objective function

$$\Psi_S(\xi_n) = \frac{\sum_{j=1}^{n} K\left(u_j;\, h\right) \sum_{j=1}^{n} u_j^2 K\left(u_j;\, h\right) - [\sum_{j=1}^{n} u_j K\left(u_j;\, h\right)]^2}{h \sum_{j=1}^{n} u_j^2 K\left(u_j;\, h\right)}. \tag{10}$$

This objective function is equivalent to (7) with $q = 1$.

To select designs for predicting the whole curve over an interval, we construct a compound criterion (Atkinson et al., 2007, ch.21) by integrating (10) across $I = [a, b]$.

Definition 1: A D_{SI}-optimal design ξ_n^* for prediction across the interval $[a, b]$ for local linear regression maximises

$$\begin{aligned}\Psi_{SI}(\xi_n) &= \frac{1}{h}\int_a^b \log\left(\sum_{j=1}^{n} K\left(x_j - x^*;\, h\right) - \frac{\left[\sum_{j=1}^{n}\left(\frac{x_j - x^*}{h}\right) K\left(x_j - x^*;\, h\right)\right]^2}{\sum_{j=1}^{n}\left(\frac{x_j - x^*}{h}\right)^2 K\left(x_j - x^*;\, h\right)}\right) \mathrm{d}x^* \\ &= \int_a^b \log\left[L(x^*)\right] \mathrm{d}x^*, \end{aligned} \tag{11}$$

where

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

We integrate the logarithm of (10) as an alternative to integrating the D_S-efficiency across $[a, b]$. Here taking logs compensates for any differences in scale in (10) for different values of x^* (see also Woods et al., 2006).

Note that we do not restrict the design region to be the interval $[a, b]$ (cf Box and Draper, 1959 and the concept of the operability region). However, most D_{SI}-optimal designs have all points within, or close to, the interval $[a, b]$.

Evaluation of (11) is analytically intractable for both the uniform and Gaussian kernels and hence we find designs by applying a Legendre-Gauss quadrature approximation (see Golub and Welsch, 1969). Thus, we maximise a weighted sum of the logarithm of objective function (10) at p_a abscissa values,

$$\Psi_{SI}(\xi_n) \approx \sum_{i=1}^{p_a} \kappa_i \log\left[L(x_i^*)\right], \tag{12}$$

where x_i^* are chosen as solutions to the Legendre polynomials and κ_i are Legendre-Gauss weights. By comparing values of (12) calculated using different values of p_a, we concluded that $p_a = 25$ was a sufficient in most cases (including for large n); a value of $p_a = 500$ was used for a few, more difficult, integrals.

4 Theoretical and numerical results

In this section, we apply the D_{SI}-criterion to find designs for different values of run sizes n. First, we consider the special case of D_S-optimality for predicting at a single point x^*, where some theoretical results on optimal designs can be derived.

4.1 Designs for prediction at a single point

We now derive sufficient conditions for a design to be D_S-optimal, i.e. maximising (10), for prediction at a single point using local linear regression. We then demonstrate the results for the uniform kernel. Our discussion of the Gaussian kernel is delayed to Section 4.2 where we apply D_{SI}-optimality. We first prove a simple but useful result.

Lemma 1. *An upper bound, U, for objective function* (10) *is given by*

$$U = \frac{n}{h} K(0;\, h) \geq \max_{\xi_n \in \mathcal{D}_n} \{\Psi_S(\xi_n)\} \ .$$

where $\mathcal{D}_n$ is the set of all possible designs of run size n.

Proof. By definition, $K(0;\, h)$ is the maximum value of K. Hence, $K\left(u_j;\, h\right) = \left[K(0;\, h) - f(u_j)\right]$ where $f(x) \geq 0$ for all x and $f(0) = 0$. Hence (10) can be expressed as

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

As the kernel is a non-negative function,

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

and hence

$$\Psi_S(\xi) = \frac{1}{h}\left(\sum_{j=1}^{n}\left[K(0;\,h) - f(u_j)\right] - \frac{[\sum_{j=1}^{n} u_j K\left(u_j;\,h\right)]^2}{\sum_{j=1}^{n} u_j^2 K\left(u_j;\,h\right)}\right) \tag{13}$$

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

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

$$\leq \frac{n}{h}K(0;\,h)\,. \tag{14}$$

□

Lemma 1 leads directly to a sufficient condition for a design to be D_S-optimal.

Theorem 1. *A sufficient condition for design ξ_n^* to be D_S-optimal for prediction at a single point using local linear regression is that $\Psi_S(\xi_n^*) = nK(0;\,h)/h$.*

Proof. Proof follows directly from Lemma 1. □

4.1.1 Optimal designs using the uniform kernel

We now find D_S-optimal designs for prediction at a single point under the uniform kernel. Using this kernel, objective function (10) simplifies to:

$$\Psi_S(\xi_n) = \frac{1}{2h}\left[\sum_{j=1}^{n} 1_A(u_j) - \frac{[\sum_{j=1}^{n} u_j 1_A(u_j)]^2}{\sum_{j=1}^{n} u_j^2 1_A(u_j)}\right], \tag{15}$$

where

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

and

$$A = \left\{u_j; |u_j| \leq h\right\}\,. \tag{16}$$

It is straightforward to establish the form of D_S-optimal designs that maximise (15) as a corollary to Theorem 1.

Corollary 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) $|x_j - x^*| \leq h$ *for* $j = 1, \ldots, n$

(ii) $\bar{x} = \sum_{j=1}^{n} x_j/n = x^*$

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

has $\Psi_S(\xi_n) = \frac{n}{2h}$ *and is* D_S*-optimal.*

Proof. Conditions (i)-(iii) imply that (a) $1_A(u_j) = 1$ for all $j = 1, \ldots, n$, (b) $\sum_{j=1}^{n}(x_j - x^*) = 0$, and (c) $\sum_{j=1}^{n}(x_j - x^*)^2 > 0$. Hence, it follows from the form of (15) that

$$\Psi_S(\xi_n^*) = \frac{nK(0;\, h)}{h} = \frac{n}{2h}\,. \tag{17}$$

The D_S-optimality of the design follows directly from Theorem 1. □

The result from Corollary 1 was verified numerically by finding a variety of D_S-optimal designs via minimisation of $-\Psi_S(\xi_n)$ using the `fminsearch` and `ksrlin` routines in MATLAB (2010) and the Nelder-Mead algorithm (Nelder and Mead, 1965). Every design found satisfied the conditions in Corollary 1, with $\Psi_S(\xi_n) = n/2h$. Fedorov et al. (1999) found designs under a mean squared error criterion for the uniform kernel on $[-1, 1]$ with $h = 2$. For this choice of kernel and bandwidth, every design point in $[-1, 1]$ has equal weight for prediction at any point in $x^* \in [-1, 1]$. Their design had equally replicated support points $\{-1, 1\}$ and, from Corollary 1, this design is also D_S-optimal.

4.2 Designs for prediction across an interval

We apply the D_{SI}-optimality criterion using local linear regression. For the numerical results, we take the interval $[-1, 1]$, that is $a = -1$ and $b = 1$. Designs for predicting at a finite number of given points can be found in Fisher (2012).

4.2.1 D_{SI}-optimal designs for the uniform kernel

Table 1 gives D_{SI}-optimal and efficient designs found by numerical search for the uniform kernel and a variety of values of n and h. The choices of n and h were restricted to ensure sufficient points in each sub-interval $[x - h, x + h]$ (with $x - h \geq -1$ and $x + h \leq 1$). For

Table 1: D_{SI}-optimal designs for prediction over $[-1, 1]$ using a uniform kernel for n design points and bandwidth h.

n								
				$h = 0.2$				
12	±0.08	±0.27	±0.45	±0.62	±0.80	±0.98		
15	±1.00	±0.83	±0.68	±0.54	±0.38	±0.24	±0.11	0.00
				$h = 0.5$				
5	±1.00	±0.50	0.00					
6	±0.99	±0.55	±0.20					
7	±1.03	±0.61	±0.30	0.00				
8	±1.04	±0.71	±0.39	±0.16				
12	±1.12	±0.86	±0.65	±0.45	±0.27	±0.06		
15	±1.12	±0.95	±0.78	±0.54	±0.40	±0.30	±0.17	0.00
				$h = 0.75$				
4	±0.94	±0.25						
5	±1.09	±0.53	0.00					
6	±1.10	±0.63	±0.13					
7	±1.10	±0.82	±0.22	0.00				
8	±1.17	±0.83	±0.46	±0.08				
12	±1.18	±1.04	±0.80	±0.55	±0.16	±0.05		
15	±1.20	±1.09	±0.90	±0.70	±0.52	±0.18	±0.10	0.00
				$h = 1$				
3	±1.00	0.00						
4	±1.12	±0.32						
5	±1.16	±0.52	0.00					
6	±1.19	±0.65	±0.30					
7	±1.30	±0.85	±0.36	0.00				
8	±1.30	±0.90	±0.47	±0.24				
12	±1.34	±1.10	±0.82	±0.52	±0.37	±0.21		
15	±1.38	±1.17	±0.97	±0.72	±0.51	±0.38	±0.25	0.00

example, when $h = 0.2$ at least 11 points are required to predict over an interval of length two. In general, the minimum number of design points required is given by

$$n_{min} = \frac{b-a}{h} + 1\,, \tag{18}$$

The designs found are all symmetric about 0, resulting from the symmetry of the kernel function and the prediction interval. The points span an interval just wider than the prediction interval $[-1, 1]$ and, although well spread across the interval, they are not equally spaced. Generally, the range of the design points increases with n; the pattern with changing h is less clear. For smaller n and larger h, the numerical optimisation was more straightforward and faster.

4.3 D_{SI}-optimal designs using the Gaussian kernel

As in the previous section, Table 2 presents D_{SI}-optimal designs for the Gaussian kernel found via numerical search. Once again all the designs are symmetric about 0, due to the symmetry of both the kernel function and prediction interval. However, for larger values of h, the designs now have substantial numbers of repeated points. Larger values of h represent less local behaviour, and hence fewer support points are required. Large values of h also result in design points closer to the extremes of $[-1, 1]$; for larger h, these points will still have substantial weight for predictions near the centre of the interval. However, unlike designs for the uniform kernel, no points from these designs lie outside the prediction interval.

To assess the robustness of designs to the choice of kernel function, we calculate the efficiency under the Gaussian kernel of a design found assuming the uniform kernel would be used for the modelling. The efficiency is given by

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

where $\Psi_G(\xi^u)$ and $\Psi_G(\xi^G)$ are the respective values of objective function (12) calculated using the Gaussian kernel for two designs: (a) ξ^u, the D_{SI}-optimal design under the uniform kernel; and (b) ξ^G, the D_{SI}-optimal design using the Gaussian kernel. Two examples are considered: (i) $h = 0.5$ and $n = 5$ and (ii) $h = 0.5$ and $n = 15$. For both cases, design ξ^u can be found in Table 1 and ξ^G in Table 2. For case (i), the two designs have very similar design points and the uniform kernel design has very high efficiency of 0.998. For case (ii), the uniform design has a substantially greater number of distinct design points than the design found using the Gaussian kernel but still has high efficiency of 0.932.

In general, designs found for the Gaussian kernel may not perform well under the uniform kernel, especially for large n, as they have fewer than n_{min} distinct design points. Hence, designs for the uniform kernel are more robust to the choice of kernel function.

5 Application to the tribology experiment

We now demonstrate the D_{SI}-criterion using simulated data sets motivated by the Tribology experiment and assess the performance of the resulting designs. Recall that the aim of the experiment was to predict the wear curve over an interval. For each of the two runs of the process (Figure 1), the original data sets had 1900 data points, to which we fit local linear regression models. The bandwidths, $h = 0.2$ and $h = 0.1$ for runs 1 and 2, respectively, were chosen "by eye" to achieve adequate descriptions of the mean response.

D_{SI}-optimal designs for these two bandwidths are given in Table 3 for the Gaussian kernel and a variety of values of n. For design selection, we scaled the prediction interval to $[-1, 1]$. Notice once again the symmetry of the designs. For these relatively complex curves and smaller bandwidths, the optimal designs have many distinct points; the designs for $h = 0.1$ have more distinct points than the designs for $h = 0.2$.

Table 2: D_{SI}-optimal designs for prediction over $[-1, 1]$ using a Gaussian kernel for n design points and bandwidth h. When design points are repeated, the number of repetitions is given 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				±0.90(7)	0.00	

Table 3: Further D_{SI}-optimal designs for predicting over $[-1, 1]$ using a Gaussian kernel for n design points and bandwidth h. When design points are repeated, the number of repetitions is given in parentheses.

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	±0.58	±0.48	±0.41	±0.32
	±0.24	±0.16	±0.08	0.00				
30	±0.98(2)	±0.81(2)	±0.71	±0.60(2)	±0.50	±0.42	±0.39	±0.26(2)
	±0.16	±0.09	±0.05					
				$h = 0.2$				
15	±0.95(2)	±0.61(2)	±0.43	±0.20(2)	0.00			
20	±0.94(3)	±0.57(3)	±0.38	±0.16				
25	±0.96(3)	±0.79	±0.54(4)	±0.21(3)	0.16			

To allow assessment of the performance of these designs relative to using all the points in the original data set, we simulate wear observations from independent normal distributions with mean given by the local linear regressions (Figure 1) and variance 2.25×10^{-8}, commensurate with background variability exhibited by the process. We then compared the local linear regression models obtained from the D_{SI}-optimal designs and the whole data sets. Figure 2 shows the smoothed fits for these runs, using $h = 0.2$ and $h = 0.1$, respectively, for both the whole data set and the data from the corresponding D_{SI}-optimal designs with $n = 25$ (run 1) and $n = 30$ (run 2).

For run 1, the prediction from the D_{SI}-optimal design with $n = 25$ points slightly over-predicts for $x \in [600, 1200]$. The choice of $h = 0.2$ for this run perhaps over-smoothes the data in this sub-interval. However, a smaller choice of h leads to under-smoothing for larger values of x. For run 2, the D_{SI}-optimal design with $h = 0.1$ and $n = 30$ provides a reasonably accurate fit.

To assess quantitatively the performance of the D_{SI}-optimal designs, we calculated "moving window" mean squared errors (MSEs):

$$\delta_i = \sum_{k=i-100}^{k=i+100} \left[\hat{g}(x_k) - y_k\right]^2 , \quad i = 101, \ldots, 1800 . \tag{19}$$

For each run, (19) was calculated for the whole data set, labelled $\delta_i(w)$, and the D_{SI}-optimal

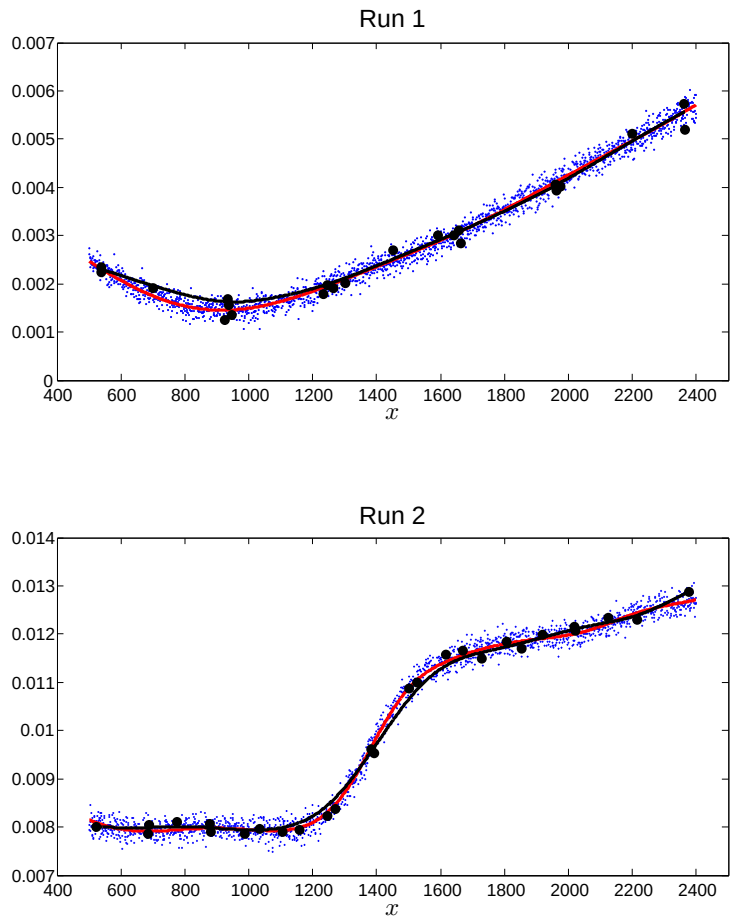

Fig 2: Simulated data (small blue dots) and D_{SI}-optimal designs (large black dots) for $n = 25$ (run 1) and $n = 30$ (run 2) with smooth fitted curves using the whole data set (red) and the data from the design points (black).

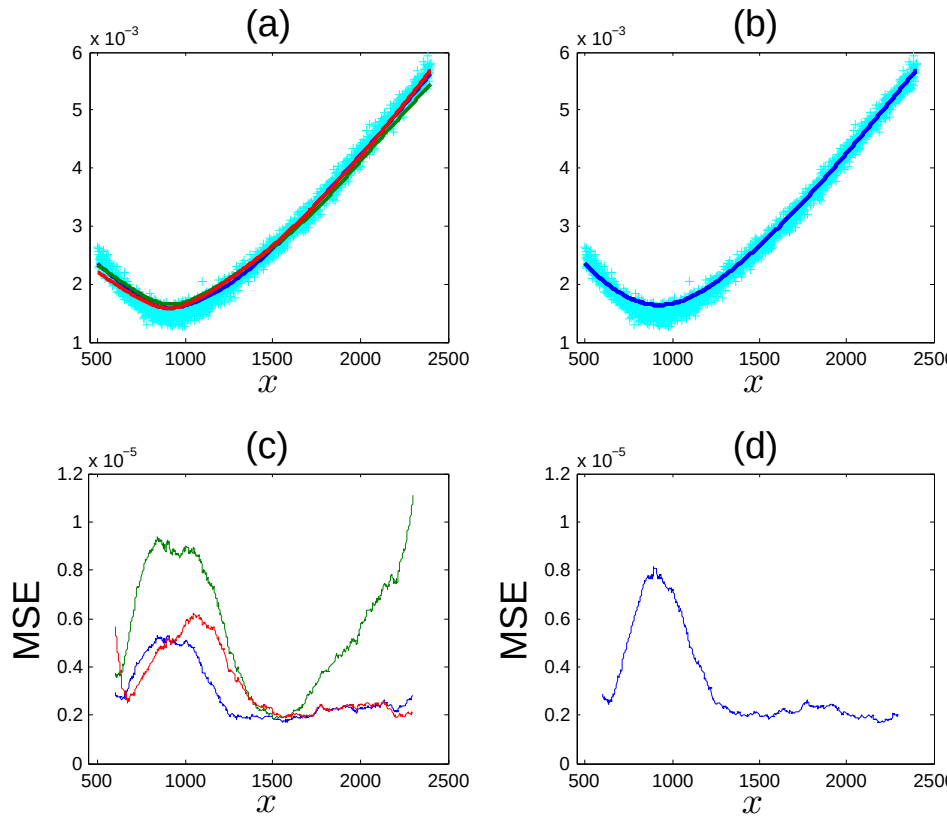

Fig 3: Run 1: Smooth fits from local linear regressions and MSE plots (a) $\hat{g}(x)$ using data corresponding to D_{SI}-optimal designs with 15 (blue), 20 (green) and 25 (red) design points, (b) $\hat{g}(x)$ for the whole data set, (c) MSE for $\hat{g}(x)$ for 15, 20 and 25 design points and (d) MSE for $\hat{g}(x)$ for the whole data set.

design, labelled $\delta(\xi)$. For an overall comparison, we used the average standardised difference

$$\bar{\Delta} = \frac{1}{1700} \sum_{i=101}^{1800} \Delta_i \,,$$

where

$$\Delta_i = \frac{\delta_i(\xi) - \delta_i(w)}{\delta_i(\xi)} \,. \tag{20}$$

In all these comparisons, it should be recognised that a D_{SI}-optimal design was found to minimise the prediction variance and not the mean squared error.

For run 1, Figure 3 shows the fitted curves from D_{SI}-optimal designs with $n = 15, 20, 25$ points and $h = 0.2$, the fitted curves for the whole data set, and plots of δ_i for the optimal

designs and the whole data set. Figure 4 gives equivalent plots for run 2 with D_{SI}-optimal designs for $n = 15, 20, 25, 30$ points and $h = 0.1$.

For both runs, there is little qualitative difference between the regression curves for the optimal designs with different numbers of runs (Figures 3(a) and 4(a)). The moving window MSEs, δ_i, for each design (Figures 3(c) and 4(c)) are two-or three-orders of magnitude smaller than the mean predictions, reflecting the strong signal to noise ratio. The MSE for the whole data set for each run (Figures 3(d) and 4(d)) is of similar magnitude to that obtained from the optimal designs. The common peaks in the plots for δ_i, in the sub-interval $[600, 1200]$ for run 1 and $[1000, 1500]$ for run 2, are indicative of areas where prediction is harder, for example where the curve has its steepest slope for run 2.

The average standardised difference $\bar{\Delta}$ is smallest for run 1, $\bar{\Delta} = 0.189$, when $n = 15$, and smallest for run 2, $\bar{\Delta} = 0.166$, when $n = 25$. While it would be expected that larger designs would produce smaller values of $\bar{\Delta}$, it should be noted that these results are for a single simulated set of data. A larger simulation study would likely produce more intuitive results.

5.1 Robustness of prediction to bandwidth selection

In general, there is likely to be some uncertainty in the correct choice of bandwidth h when designing an experiment. Even when studying run 1 and run 2, with data available, this choice was not completely clear. To assess the robustness of the quality of the model fit from D_{SI}-optimal designs to the choice of h, we use run 2 and compare predictions from using the whole data set and $h = 0.1$ (considered the best "by eye" choice of bandwidth) to the predictions obtained from optimal designs found for $h = 0.2$ and $h = 0.3$.

Figures 5 and 6 show the predicted curves and moving window MSE (δ_i) for $h = 0.2$ and $h = 0.3$ respectively. Although overall the prediction error is still small, it is clear from Figures 5(a) and 6(a) that for both $h = 0.2$ and $h = 0.3$, the local linear regression overestimates the response on the left of the inflection point, and underestimates the response on the right of the inflection point; that is, the prediction is over-smoothed. This is also clear from the plots of δ_i (Figures 5(c) and 6(c)), where the peak values of δ_i from the misspecified optimal designs are around 10 times larger than from the whole data set. Compare this to Figures 3 and 4, where peak values of δ_i for the optimal designs are less than twice that from the whole data set.

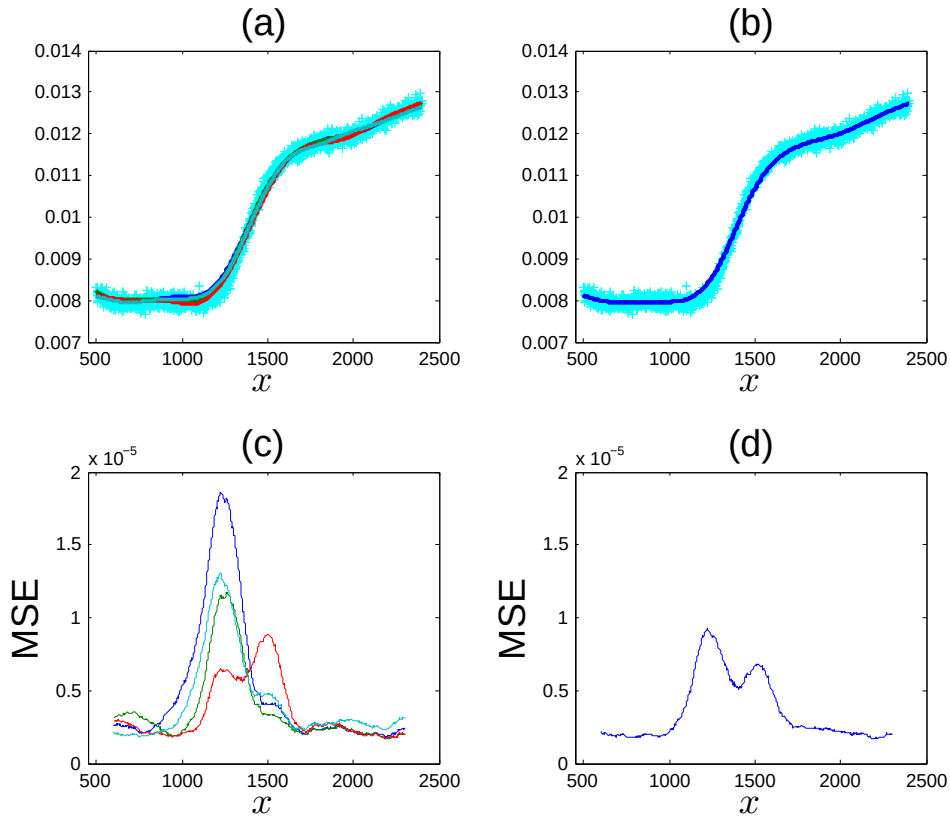

Fig 4: Run 2: Smooth fits from local linear regressions and MSE plots (a) $\hat{g}(x)$ using data corresponding to D_{SI}-optimal designs with 15 (blue), 20 (green), 25 (red) and 30 (light blue) design points, (b) $\hat{g}(x)$ for the whole data set, (c) MSE for $\hat{g}(x)$ for 15, 20, 25 and 30 design points, and (d) MSE for $\hat{g}(x)$ for the whole data set.

6 Concluding Remarks

We have presented a new optimality criterion, the D_{SI}-criterion, for selecting designs to enable accurate prediction across a predetermined interval using local linear regression. The new designs have been compared and assessed. The criterion makes use of a weighted least squares approximation to the local linear regression. For the uniform kernel and model (1), with identically distributed errors, the approximation to the prediction variance made via model (8) differs from the true variance only by a multiplicative constant. For the Gaussian kernel, the approximation also makes an adjustment to the bandwidth.

A clear direction for future work, motivated by the Tribology application, is to find designs that assume a varying bandwidth across the prediction interval. Such models and

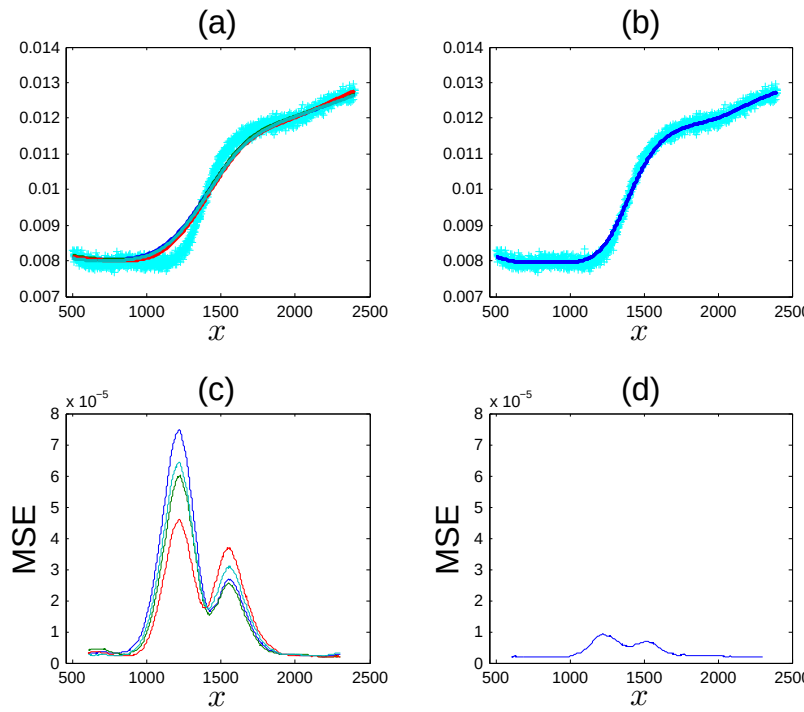

Fig 5: Run 2: Smooth fits from local linear regressions and MSE plots (a) $\hat{g}(x)$ using data corresponding to D_{SI}-optimal designs with 15 (blue), 20 (green), 25 (red) and 30 (light blue) points for $h = 0.2$, (b) $\hat{g}(x)$ for the whole data set with true bandwidth, $h = 0.1$, (c) MSE for $\hat{g}(x)$ for 15, 20, 25 and 30 design points, and (d) MSE for $\hat{g}(x)$ for the whole data set.

designs would be better able to approximate responses that display marked differences in smoothness.

Acknowledgments

This work was supported by the UK Engineering and Physical Sciences Research Council through a PhD studentship (Fisher) and a Fellowship (Woods; EP/J018317/1). The authors thank Professor Robert Wood, Dr Ling Wang and Dr Terry Harvey (National Centre for

Advanced Tribology, University of Southampton) and Dr Ramkumar Penchaliah (Indian Institute of Technology Madras) for providing the data for the example, and for related discussions. The authors also acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

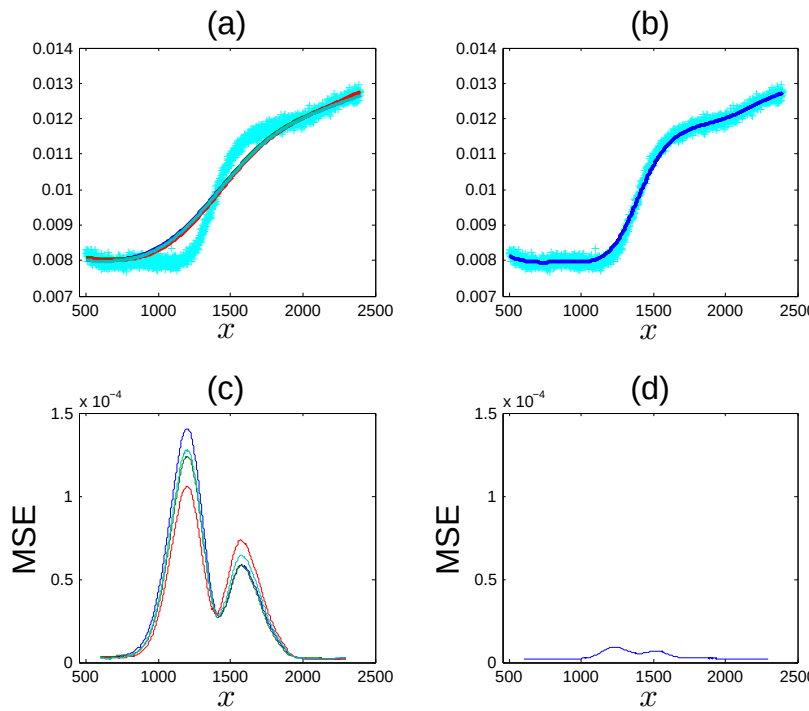

Fig 6: Run 2: Smooth fits from local linear regressions and MSE plots (a) $\hat{g}(x)$ using data corresponding to D_{S1}-optimal designs with 15 (blue), 20 (green), 25 (red) and 30 (light blue) design points for $h = 0.3$, (b) $\hat{g}(x)$ for the whole data set with true bandwidth, $h = 0.1$, (c) MSE for $\hat{g}(x)$ for 15, 20, 25 and 30 design points, and (d) MSE for $\hat{g}(x)$ for the whole data set.

References

Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007) *Optimum Experimental Designs, with SAS*. Oxford: Oxford University Press, 2nd edn.

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.

Box, G. E. P. and Draper, N. R. (1959) A basis for the selection of a response surface design. *Journal of the American Statistical Association*, **54**, pp. 622–654.

Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear smoothers and additive-models. *Annals of Statistics*, **17**, 453–510.

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.

Fan, J. Q. (1992) Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.

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.

Fisher, V. A. (2012) *Optimal and efficient experimental design for nonparametric regression with application to functional data*. Ph.D. thesis, University of Southampton, UK.

Golub, G. H. and Welsch, J. H. (1969) Calculation of Gauss quadrature rules. *Mathematics of Computation*, **23**, 221–230.

MATLAB (2010) *Version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts.

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.

Pelto, C. R., Elkins, T. A. and Boyd, H. A. (1968) Automatic contouring of irregularly spaced data. *Geophyisics*, **33**, 424–430.

Ramsay, J. O. and Silverman, B. (2005) *Functional Data Analysis*. New York: Springer, 2nd edn.

Simonoff, J. S. (1996) *Smoothing Methods in Statistics*. New York: Springer-Verlag.

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.

Woods, D. C., Lewis, S. M., Eccleston, J. A. and Russell, K. G. (2006) Designs for generalized linear models with several variables and model uncertainty. *Technometrics*, **48**, 284–292.

Verity A. Fisher, David C. Woods, Susan M. Lewis
Southampton Statistical Sciences Research Institute
University of Southampton
Southampton, SO17 1BJ, UK
Email: {V.Fisher,D.Woods,S.M.Lewis}@southampton.ac.uk