A Comparative Study of Robust Regression Designs

by

Mengzhe Wang

Master of Science Project
September 2006

Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, Alberta, Canada

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest sense of gratitude to my supervisor, Dr. Douglas Wiens for his inspirational instruction, tremendous support and invaluable guidance. He is very patient with my report editing and revision.

My sincere thanks are extended to Dr. Edit Gombay, Dr. Eric Woolgar, Dr. A. Adewale, and Dr. Xiaojian Xu for their insightful comments, helpful suggestions, encouragement and understanding during my study.

1 Introduction

Consider an approximate linear regression model with additive, homoscedastic errors:

$$Y(\mathbf{x}) = E[Y|\mathbf{x}] + \varepsilon \tag{1}$$

for a q-dimensional independent variable $\mathbf{x} \in S$ and

$$E[Y|\mathbf{x}] \approx \mathbf{z}^T(\mathbf{x})\boldsymbol{\theta}$$

for a p-dimensional regressor $\mathbf{z}$. We can define

$$\boldsymbol{\theta}_0 = \arg\min \int_{\mathcal{S}} \left\{E[Y|\mathbf{x}] - \mathbf{z}^T(\mathbf{x})\boldsymbol{\theta}\right\}^2 d\mathbf{x} \text{ and } f(\mathbf{x}) = E[Y|\mathbf{x}] - \mathbf{z}^T(\mathbf{x})\boldsymbol{\theta}_0.$$

Because the mean response $E[Y|\mathbf{x}]$ is well approximated by $\mathbf{z}^T(\mathbf{x})\boldsymbol{\theta}$ but may not be exactly linear, the least squares estimator $\boldsymbol{\theta}_0$ is no longer unbiased. Therefore, the mean squared error (MSE) of $\boldsymbol{\theta}_0$ is of interest; it can be given as the sum of squared bias and variance. That is going to be the loss function which will be discussed later. The associated design problem is to choose the observation sites $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ and tell the experimenter what the design support points are and how many subjects should be allocated to each of these points. A design is selected so as to minimize the mean square error averaged over the region $\mathcal{S}$. However, for instance, sometimes the experimenter is fitting a lower degree polynomial over the region in the circumstances where the true response function is a higher degree polynomial. Therefore, we are motivated to explore a "good" design which can detect the departure from fitting model.

Box and Draper (1959) found the dangers of ignoring the bias when fitting an incorrect model. In general, the regression model under the consideration of robustness is assumed to be only approximately known. The true model can be written as $E[Y_i|\mathbf{x}] = \mathbf{z}^T(\mathbf{x}_i)\boldsymbol{\theta}_0 + f(\mathbf{x}_i) + \varepsilon_i$, $i = 1, 2, \cdots, n$ where $f(\mathbf{x})$ is an unknown function from some contamination class $\mathcal{F}$. The designs must be chosen such that the fitted model provides an adequate approximation over a range of possible models.

The model robust design problem has been studied by many authors, whose investigations differ in specification of the contamination class $\mathcal{F}$, the design region, the regressors and the criteria of optimality used. For infinite dimensional $\mathcal{F}$, there are two major types of contamination classes used in the literature. One is the approach for which the contamination class is

$$\mathcal{F}_1 = \{f : |f(\mathbf{x})| \leq \phi(\mathbf{x}), \int_{\mathcal{S}} \mathbf{z}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \mathbf{0}\}.$$

Marcus and Sacks (1978) took $\mathcal{F}_1$ for a specified function ϕ with $\phi(\mathbf{0}) = 0$. Li and Notz (1982), Pesotchinsky (1982), Li (1984) considered their designs using the contamination class $\mathcal{F}_1$ and extended their results to multiple linear regression. However, $\mathcal{F}_1$ is too "thin" because it contains too few functions to be realistic and all those functions are bounded by $\phi(\mathbf{x})$. This class often leads to designs whose mass is concentrated at a small number of points in the design space, hence it has severely limited robustness against realistic departures from the assumed model (Wiens 1992).

The other type of contamination class is

$$\mathcal{F}_2 = \{f : \int_{\mathcal{S}} [f(\mathbf{x})]^2 d\mathbf{x} \leq \eta_s^2, \int_{\mathcal{S}} \mathbf{z}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \mathbf{0}\},$$

where η_s is a constant and assumed "small", so that the parametric term in the regression model is still dominant. This class is used by Huber (1975), and Wiens (1990, 1992, 1996, 1998), and the approach for design construction with this class is called the "Huber approach" in this report. One criticism of this specification is that only designs which are absolutely continuous on $\mathcal{S}$ have a finite maximum loss. Wiens (1992) proved that exact designs in this class have infinite maximum loss. Therefore, robust designs constructed for deviations in $\mathcal{F}_2$ are continuous designs which are approximated by discrete designs in practice. However, in many practical settings, the true model is likely to have some smoothness so that $\mathcal{F}$ considered as in $\mathcal{F}_2$ space is too "broad".

To solve the too "broad" or too "thin" problem, Yue and Hickernell (1999) considered a reproducing kernel Hilbert space (RKHS) as the contamination class $\mathcal{F}$. A reproducing kernel Hilbert space is actually a function space in which

pointwise evaluation is a continuous linear function, i.e., it is the space that can be defined by a reproducing kernel. In Yue and Hickernell (1999), a sharp upper bound for the mean squared error is found in terms of the norm of contamination function which is in RKHS. This upper bound is used to choose a discrete and exact design that is robust against the model bias.

Fang and Wiens (2000) considered the construction of integer-valued designs for approximately linear models. The simulated annealing algorithm is given in this report. Through this annealing approach, Fang and Wiens searched for integer-valued, rather than continuous, designs in a finite design space. One advantage of using simulated annealing for carrying out the numerical minimization problem is to reduce the computing time. Another is that discrete designs are obtained automatically - it is not necessary to do any implementation for practical use after calculating the design. Most importantly however is that the analytic work necessary to obtain the minimizing designs, using other approaches, is feasible only in very simple, well-structured problems - simple linear regression, or multiple regression without interactions, over a spherical design space, for instance. In contrast, annealing can be carried for any response functions, over any design region. It eliminates the restriction imposed by the response function.

Therefore, we have several different designs using different approaches: the approach using $\mathcal{F}_1$, the Huber approach, the RKHS approach and the annealing approach. The goal of this report is to compare some of those designs and discuss with each other. We begin with a summary of RKHS with some fundamental concepts and definitions, followed by a description of design problems for which we seek solutions. In Section 3, we review Yue and Hickernell (1999) and calculate designs for a linear regression model. In Section 4, we give Huber's minimax design for simple linear regression and approximate this continuous design. In Section 5, the annealing approach, presented by Fang and Wiens (2000), is used to get an integer-valued design. And with the comparison between different approaches through variance component and bias component, our conclusion is given in Section 6.

2 Reproducing Kernel Hilbert Spaces

The researchers find that $\mathcal{F}_1$ is too "thin", with the result that the robust designs found generally have only a small number of support points and do not allow exploration of models larger than the fitted one, while $\mathcal{F}_2$ is too broad, resulting in the inclusion of too many functions in the contamination class. So that is the motivation for us to seek a contamination class large enough to be realistic but for minimax design is not necessary to be continuous. In RKHS, the function f can be expressed explicitly and expanded on the basis. The minimax procedure will boil down to dealing with kernels and produce the discrete designs directly. In past work, in order to get a finite maximum loss, the designs have to be continuous. For practical purposes, it is necessary to use discrete designs to approximate continuous ones before implementation. In this case, we are looking for a good contamination class to achieve these two purposes at the same time.

A promising contamination class is provided by the theory of Reproducing Kernel Hilbert Spaces, which is summarized here. A good reference is Wahba (1990). The material in this section is largely taken from Evgeniou, Pontil, and Poggio (2000).

Let $\mathcal{H}$ be a Hilbert space consisting of real-valued functions defined on the design space $\mathcal{S} \subset \mathbb{R}^q$, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. For instance $\mathcal{H} = \mathcal{L}_2(\mathcal{S})$, with

$$\langle f, g \rangle_{\mathcal{H}} = \int_{\mathcal{S}} f(\mathbf{x}) g(\mathbf{x}) d\mathbf{x}.$$

Perhaps more interesting is if $\mathcal{H}$ is a Sobolev space $\mathcal{W}_m(\mathcal{S}) \subset \mathcal{L}_2(\mathcal{S})$, i.e., the space of functions f on $\mathcal{S}$ having partial derivatives up to order m and

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{|\alpha| \leq m} \int_{\mathcal{S}} D^{\alpha} f(\mathbf{x}) D^{\alpha} g(\mathbf{x}) d\mathbf{x},$$

where

$$D^{\alpha} f = \frac{\partial^{|\alpha|}}{\partial x_1 \cdots \partial x_q} f, \ |\alpha| = \alpha_1 + \cdots + \alpha_q.$$

The cases $m = 0, 1, 2$ would presumably be the interesting ones. If $m = 0$, then this is just $\mathcal{L}_2(\mathcal{S})$**.** There are several different RKHS according to the different definitions of kernels. For example, Sobolev-Hilbert space $\mathcal{W}_m$ defined by:

$$\mathcal{W}_m : \mathcal{W}_m[0,1] = \{f : f,\ f^{'}, \cdots, f^{(m-1)} \text{ absolutely continuous, } f^{(m)} \in \mathcal{L}_2[0,1]\}$$

It is known that $\mathcal{W}_m$ is a RKHS. But not all Sobolev space are RKHS. Sobolev space is the general term given for a space obtained by imposing on a function f and its first few derivatives the requirement of a finite L^p norm; it is not necessarily a Hilbert space.

A *reproducing kernel* is a symmetric function $K(\mathbf{x}, \mathbf{y})$ defined on $\mathcal{S} \times \mathcal{S}$, and an associated inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{K}}$, with the property

$$f(\mathbf{x}) = \langle f(\mathbf{y}), K(\mathbf{x}, \mathbf{y}) \rangle_{\mathcal{H}_\mathcal{K}}, \forall f \in \mathcal{H}.$$

These arise as follows. For a set of functions $\left\{\phi_j\right\}_{j=1}^{\infty} \subset \mathcal{H}$, take the Hilbert space $\mathcal{H}_\mathcal{K}$ consisting of all

$$f(\mathbf{x}) = \sum_{j=1}^{\infty} a_j \phi_j(\mathbf{x}), \tag{2}$$

with the inner product

$$\left\langle \sum_{j=1}^{\infty} a_j \phi_j(\mathbf{x}), \sum_{j=1}^{\infty} b_j \phi_j(\mathbf{x}) \right\rangle_{\mathcal{H}_\mathcal{K}} = \sum_{j=1}^{\infty} \frac{a_j b_j}{\lambda_j} \tag{3}$$

for positive, square-summable $\{\lambda_j\}_{j=1}^{\infty}$. Then $K(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\infty} \lambda_j \phi_j(\mathbf{x}) \phi_j(\mathbf{y})$ has the desired property and the set defined by (2) is a Reproducing Kernel Hilbert Space (RKHS) with kernel $K(\mathbf{x}, \mathbf{y})$. The kernel is positive definite. The induced norm in $\mathcal{H}_\mathcal{K}$ is

$$\|f\|^2_{\mathcal{H}_\mathcal{K}} = \langle f, f \rangle_{\mathcal{H}_\mathcal{K}} = \sum_{j=1}^{\infty} \frac{a_j^2}{\lambda_j}. \tag{4}$$

The basis $\left\{\phi_j\right\}_{j=1}^{\infty}$ is a basis for the RKHS (not necessarily orthonormal) and the kernel $K(\mathbf{x}, \mathbf{y})$ is the "correlation" matrix associated with the basis functions. The choice of the$\left\{\phi_j\right\}_{j=1}^{\infty}$ defines a space of functions - the functions spanned by the $\left\{\phi_j\right\}_{j=1}^{\infty}$. The number of basis element of ϕ_j does not need to be infinite. The elements of RKHS are all the functions f that have a finite norm given by equation (4). One could for instance choose $\left\{\phi_j\right\}_{j=1}^{\infty}$ to be an orthonormal basis for the Sobolev-Hilbert space $\mathcal{W}_m(\mathcal{S})$.

Define functions $k_{\mathbf{x}}$ on $\mathcal{S}$ by $k_{\mathbf{x}}(\mathbf{z}) = K(\mathbf{x}, \mathbf{z})$. It follows from (3) that

$$\langle k_{\mathbf{x}}, k_{\mathbf{y}} \rangle_{\mathcal{H}_K} = K(\mathbf{x}, \mathbf{y}).$$

Alternatively, we can start with $K(\mathbf{x}, \mathbf{y})$ such that $\int_{\mathcal{S}\times\mathcal{S}} K^2(\mathbf{x}, \mathbf{y})\, d\mathbf{x}d\mathbf{y} < \infty$. Then there exists an orthonormal sequence $\{\phi_j\}_{j=1}^{\infty} \subset \mathcal{L}_2(\mathcal{S})$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ with

$$K(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\infty} \lambda_j \phi_j(\mathbf{x}) \phi_j(\mathbf{y}), \tag{5}$$

$$\int_{\mathcal{S}} K(\mathbf{x}, \mathbf{y})\, \phi_j(\mathbf{y}) d\mathbf{y} = \lambda_j \phi_j(\mathbf{x}),$$

$$\int_{\mathcal{S}\times\mathcal{S}} K^2(\mathbf{x}, \mathbf{y})\, d\mathbf{x}d\mathbf{y} = \sum_{j=1}^{\infty} \lambda_j^2 < \infty.$$

Now define $f_j = \int_{\mathcal{S}} f(\mathbf{x})\, \phi_j(\mathbf{x}) d\mathbf{x}$ and let $\mathcal{H}_K$ be the set of all functions f with $\sum_{j=1}^{\infty} \frac{f_j^2}{\lambda_j} < \infty$. Then $\mathcal{H}_K$ is a RKHS with kernel $K(\mathbf{x}, \mathbf{y})$ and inner product $\langle f, g \rangle_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \frac{f_j g_j}{\lambda_j}$. Because we take the basis from $\mathcal{H}$, RKHS is a subspace of Hilbert space. Also kernel $k_{\mathbf{x}}(\cdot)$ is belongs to RKHS too if $\mathbf{x}$ is fixed. Wahba (1990) proved this statement and gave the if and only if condition for $f \in H_K$. To see the reproducing property, note that $f(\mathbf{x}) = \sum_{j=1}^{\infty} f_j \phi_j(\mathbf{x})$, so that

$$f(\cdot) = \sum_{j=1}^{\infty} f_j \phi_j(\cdot); \text{ also}$$

$$K(\mathbf{x}, \cdot) = \sum_{j=1}^{\infty} \lambda_j \phi_j(\mathbf{x}) \phi_j(\cdot).$$

Thus

$$\langle f(\mathbf{y}), K(\mathbf{x}, \mathbf{y}) \rangle_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \frac{f_j \cdot \lambda_j \phi_j(\mathbf{x})}{\lambda_j} = \sum_{j=1}^{\infty} f_j \phi_j(\mathbf{x}) = f(\mathbf{x}).$$

One can notice that $k_{\mathbf{x}}(\cdot) \in \mathcal{H}_K$ since $\sum_{j=1}^{\infty} \{\lambda_j \phi_j(\mathbf{x})\}^2 / \phi_j(\cdot) = \sum_{j=1}^{\infty} \lambda_j \phi_j^2(\mathbf{x}) = K(\mathbf{x}, \mathbf{x})\ \forall \mathbf{x} \in \mathcal{S}$. (See Wahba 1990)

3 Summary of Yue & Hickernell (1999)

This section is a summary of the work of Rong-Xian Yue and Fred J. Hickernell (referred to as Y&H). This project considers the design problem for linear regression. The true model is presented in (6) with unknown contamination function

$f(\mathbf{x})$ from some class. Also it is assumed that the contamination function is orthogonal to regressor $\mathbf{z}(\mathbf{x})$ as in (7)

$$Y(\mathbf{x}) = \mathbf{z}^T(\mathbf{x})\boldsymbol{\theta}_0 + f(\mathbf{x}) + \varepsilon, \tag{6}$$

$$\int_{\mathcal{S}} \mathbf{z}(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} = \mathbf{0}. \tag{7}$$

Let the design measure ξ place mass n^{-1} at each of n, not necessarily distinct, points $\mathbf{x}_1, ..., \mathbf{x}_n$. With

$$\mathbf{A} = \int_{\mathcal{S}} \mathbf{z}(\mathbf{x})\mathbf{z}^T(\mathbf{x}) d\mathbf{x},\ \mathbf{B} = \int_{\mathcal{S}} \mathbf{z}(\mathbf{x})\mathbf{z}^T(\mathbf{x}) \xi(d\mathbf{x}),\ \mathbf{b} = \int_{\mathcal{S}} \mathbf{z}(\mathbf{x}) f(\mathbf{x})\, \xi(d\mathbf{x}),$$

we have

$$\begin{aligned}
E\left[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right] &= \mathbf{B}^{-1}\mathbf{b}, \\
\text{COV}\left[\hat{\boldsymbol{\theta}}\right] &= \frac{\sigma_\varepsilon^2}{n}\mathbf{B}^{-1}, \\
\text{MSE}\left[\hat{\boldsymbol{\theta}}\right] &= E\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T\right] \\
&= \frac{\sigma_\varepsilon^2}{n}\mathbf{B}^{-1} + \mathbf{B}^{-1}\mathbf{b}\mathbf{b}^T\mathbf{B}^{-1}, \\
\text{IMSE} &= \int_{\mathcal{S}} E\left[\left\{\hat{Y}(\mathbf{x}) - \mathbf{z}^T(\mathbf{x})\boldsymbol{\theta}_0\right\}^2\right] d\mathbf{x} \\
&= \frac{\sigma_\varepsilon^2}{n} tr\left\{\mathbf{B}^{-1}\mathbf{A}\right\} + \mathbf{b}^T\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{b} \\
&\overset{def}{=} Q_v(\xi) + Q_b(f, \xi).
\end{aligned}$$

Assume now that f varies over a RKHS $\mathcal{H}_\mathcal{K} \subset L_2(\mathcal{S})$, all of whose members satisfy (7), and

$$\|f\|^2_{\mathcal{H}_\mathcal{K}} \leq \eta^2, \tag{8}$$

for a fixed constant η^2. Define a vector of functions on $\mathcal{S}$ by $\mathbf{k}(\mathbf{y}) = (k_{\mathbf{x}_1}(\mathbf{y}), ..., k_{\mathbf{x}_n}(\mathbf{y}))^T$. Set

$$\mathbf{K}_{n\times n} = \left\langle \mathbf{k}, \mathbf{k}^T \right\rangle_{\mathcal{H}_\mathcal{K}},$$

with elements $K_{ij} = \left\langle k_{\mathbf{x}_i}, k_{\mathbf{x}_j} \right\rangle_{\mathcal{H}_\mathcal{K}} = K(\mathbf{x}_i, \mathbf{x}_j)$.

We first maximize $Q_b(f,\xi)$ over $\mathcal{H}_\mathcal{K}$. Write

$$Q_b(f,\xi) = \mathbf{b}^T\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{b} = \mathbf{f}^T\mathbf{C}^T\mathbf{C}\mathbf{f},$$

where

$$\begin{aligned}\mathbf{C}_{p\times n} &= \frac{1}{n}\mathbf{A}^{1/2}\mathbf{B}^{-1}\mathbf{Z}^T,\\ \mathbf{f}^T &= (f(\mathbf{x}_1),...,f(\mathbf{x}_n))\\ &= \left(\langle f, k_{\mathbf{x}_1}\rangle_{\mathcal{H}_K},...,\langle f, k_{\mathbf{x}_n}\rangle_{\mathcal{H}_K}\right)\\ &= \left\langle f, \mathbf{k}^T\right\rangle_{\mathcal{H}_K}.\end{aligned}$$

Then

$$\mathbf{C}\mathbf{f} = \langle f, \mathbf{C}\mathbf{k}\rangle_{\mathcal{H}_K} = \langle f, \boldsymbol{\zeta}\rangle_{\mathcal{H}_K},$$

for $\boldsymbol{\zeta} \overset{def}{=} \mathbf{C}\mathbf{k}$, and we are to maximize

$$Q_b(f,\xi) = \mathbf{f}^T\mathbf{C}^T\mathbf{C}\mathbf{f} = \sum_{j=1}^{p}\left(\left\langle f, \zeta_j\right\rangle_{\mathcal{H}_\mathcal{K}}\right)^2.$$

Decompose f as $f = f_0 + f_1$, where $f_0 = \boldsymbol{\beta}^T\boldsymbol{\zeta} \in span\left\{\zeta_j\right\}_{j=1}^p$ and f_1 is orthogonal to this span. Both f_0 and f_1 should come from the RKHS $\mathcal{H}_0$ and $\mathcal{H}_1$. The two subspaces are spanned by $span\left\{\zeta_j\right\}_{j=1}^p$ and $span\left\{\zeta_j\right\}_{j=p}^\infty$. So $\mathcal{H}_k$=$\mathcal{H}_0 \oplus \mathcal{H}_1$. Then with $\mathbf{e}_j = (0,...,0,\overset{j}{\underset{\downarrow}{1}},0,...0)^T : p\times 1$ and

$$\mathbf{Q}_{p\times p} \overset{def}{=} \mathbf{C}\mathbf{K}\mathbf{C}^T$$

we have

$$\begin{aligned}\left\langle f, \zeta_j\right\rangle_{\mathcal{H}_\mathcal{K}} &= \left\langle f_0, \zeta_j\right\rangle_{\mathcal{H}_\mathcal{K}} = \left\langle \boldsymbol{\beta}^T\boldsymbol{\zeta}, \boldsymbol{\zeta}^T\mathbf{e}_j\right\rangle_{\mathcal{H}_\mathcal{K}} = \left\langle \boldsymbol{\beta}^T\mathbf{C}\mathbf{k}, \mathbf{k}^T\mathbf{C}^T\mathbf{e}_j\right\rangle_{\mathcal{H}_\mathcal{K}}\\ &= \boldsymbol{\beta}^T\mathbf{C}\left\langle \mathbf{k}, \mathbf{k}^T\right\rangle_{\mathcal{H}_\mathcal{K}}\mathbf{C}^T\mathbf{e}_j = \boldsymbol{\beta}^T\mathbf{C}\mathbf{K}\mathbf{C}^T\mathbf{e}_j = \boldsymbol{\beta}^T\mathbf{Q}\mathbf{e}_j,\end{aligned}$$

with

$$Q_b(f,\xi) = \mathbf{f}^T\mathbf{C}^T\mathbf{C}\mathbf{f} = \sum_{j=1}^{p}\left(\boldsymbol{\beta}^T\mathbf{Q}\mathbf{e}_j\right)\left(\boldsymbol{\beta}^T\mathbf{Q}\mathbf{e}_j\right)^T = \boldsymbol{\beta}^T\mathbf{Q}^2\boldsymbol{\beta}. \tag{9}$$

Also,

$$\begin{aligned}\|f\|_{\mathcal{H}_\mathcal{K}}^2 &= \|f_0\|_{\mathcal{H}_\mathcal{K}}^2 + \|f_1\|_{\mathcal{H}_\mathcal{K}}^2 = \left\|\boldsymbol{\beta}^T\mathbf{Ck}\right\|_{\mathcal{H}_\mathcal{K}}^2 + \|f_1\|_{\mathcal{H}_\mathcal{K}}^2 \\ &= \boldsymbol{\beta}^T\mathbf{C}\left\langle\mathbf{k},\mathbf{k}^T\right\rangle_{\mathcal{H}_\mathcal{K}}\mathbf{C}^T\boldsymbol{\beta} + \|f_1\|_{\mathcal{H}_\mathcal{K}}^2 \\ &= \boldsymbol{\beta}^T\mathbf{Q}\boldsymbol{\beta} + \|f_1\|_{\mathcal{H}_\mathcal{K}}^2,\end{aligned}$$

so that

$$\boldsymbol{\beta}^T\mathbf{Q}\boldsymbol{\beta} \leq \|f\|_{\mathcal{H}_\mathcal{K}}^2, \tag{10}$$

with equality iff $f = f_0 = \boldsymbol{\beta}^T\boldsymbol{\zeta}$. The loss function IMSE$= \int_\mathcal{S} E\left[\left\{\hat{Y}(\mathbf{x}) - \mathbf{z}^T(\mathbf{x})\boldsymbol{\theta}_0\right\}^2\right]d\mathbf{x}$ attains its maximum over contamination class

$$\mathcal{F} = \{f : f \in \mathcal{W}_m(\mathcal{S}), ||f||_{\mathcal{H}_\mathcal{K}} \leq \eta_{\mathcal{H}_k}^2, \int_\mathcal{S}\mathbf{z}(\mathbf{x})f(\mathbf{x})d\mathbf{x} = \mathbf{0}\}$$

at $f_{\max} = f_0$ with $||f_{\max}||_{\mathcal{H}_\mathcal{K}}^2 = ||f_0||_{\mathcal{H}_\mathcal{K}}^2 = \eta_{\mathcal{H}_k}^2$.

Y&H proceed as follows. Since

$$\frac{\boldsymbol{\beta}^T\mathbf{Q}^2\boldsymbol{\beta}}{\boldsymbol{\beta}^T\mathbf{Q}\boldsymbol{\beta}} \leq ch_{\max}(\mathbf{Q}), \tag{11}$$

one has

$$Q_b(f,\xi) = \boldsymbol{\beta}^T\mathbf{Q}^2\boldsymbol{\beta} \leq ch_{\max}(\mathbf{Q})\cdot\boldsymbol{\beta}^T\mathbf{Q}\boldsymbol{\beta} \leq ch_{\max}(\mathbf{Q})\cdot\|f\|_{\mathcal{H}_\mathcal{K}}^2.$$

Let $\mathbf{v}_{\max}$ be an eigenvector of $\mathbf{Q}$, with unit Euclidean norm, corresponding to $ch_{\max}(\mathbf{Q})$. Put

$$\boldsymbol{\beta} = c\mathbf{v}_{\max}$$

for an arbitrary non-zero constant c, and put

$$f = \boldsymbol{\beta}^T\boldsymbol{\zeta} = c\mathbf{v}_{\max}^T\mathbf{Ck}.$$

Then equality holds in (11) and so one has a sharp bound: For any $f \in \mathcal{H}_\mathcal{K}$,

$$\begin{aligned}\int_\mathcal{S} E\left[\left\{\hat{Y}(\mathbf{x}) - \mathbf{z}^T(\mathbf{x})\boldsymbol{\theta}_0\right\}^2\right]d\mathbf{x} &= Q_v(\xi) + Q_b(f,\xi) \\ &\leq \frac{1}{n}\sigma_\varepsilon^2 tr\left\{\mathbf{B}^{-1}\mathbf{A}\right\} + ch_{\max}(\mathbf{Q})\cdot\|f\|_{\mathcal{H}_\mathcal{K}}^2 \\ &= \frac{1}{n}[\sigma_\varepsilon^2 tr\left\{\mathbf{B}^{-1}\mathbf{A}\right\} + ch_{\max}\left(\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1}\frac{\mathbf{Z}^T\mathbf{KZ}}{n}\right)\cdot\|f\|_{\mathcal{H}_\mathcal{K}}^2].\end{aligned}$$

The loss function attains its sharp upper bound at $f_{\max} = f_0 = \frac{\eta_{\mathcal{H}_k}}{\sqrt{ch_{\max}(\mathbf{Q})}}\mathbf{v}_{\max}\mathbf{Ck}$, since on one side $||f_{\max}||^2_{\mathcal{H}_\mathcal{K}} = \eta^2_{\mathcal{H}_k}$, and on the other side $||f_{\max}||^2_{\mathcal{H}_\mathcal{K}} = ||c\mathbf{v}^T_{\max}\mathbf{Ck}||^2_{\mathcal{H}_\mathcal{K}} = c^2\mathbf{v}^T_{\max}\mathbf{C} < \mathbf{k}, \mathbf{k} > \mathbf{C}^T\mathbf{v}_{\max} = c^2\mathbf{v}^T_{\max}Q\mathbf{v}_{\max}$, then non-zero constant $c = \frac{\eta_{\mathcal{H}_k}}{\sqrt{ch_{\max}(\mathbf{Q})}}$.

Y&H define

$$
\begin{aligned}
J_v(\xi) &= tr\left\{\mathbf{B}^{-1}\mathbf{A}\right\} = ntr\left\{\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{A}\right\} \\
J_b(\xi) &= ch_{\max}(\mathbf{Q}) = ch_{\max}\left(\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{A}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{KZ}\right) \\
r &= \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + \|f_{\max}\|^2_{\mathcal{H}_\mathcal{K}}} = \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + \eta^2_{\mathcal{H}_k}},
\end{aligned}
$$

so that an (attainable) upper bound is

$$
\int_\mathcal{S} E\left[\left\{\hat{Y}(\mathbf{x}) - \mathbf{z}^T(\mathbf{x})\boldsymbol{\theta}_0\right\}^2\right] d\mathbf{x} \leq \left(\sigma^2_\varepsilon + \eta^2_{\mathcal{H}_k}\right)\left(rJ_v(\xi) + (1-r)J_b(\xi)\right).
$$

The problem now is to choose the design so as to minimize $r \cdot J_v(\xi) + (1-r) \cdot J_b(\xi)$ for fixed $r \in [0,1]$. Coefficient r reflects the relative proportion of the variance to bias. It is the prior belief of the experimenter as to the nature of the true response function.

One observation here is that Y&H are concentrating on the *imse* of $\hat{Y}(\mathbf{x})$ around $\mathbf{z}^T(\mathbf{x})\boldsymbol{\theta}_0$ rather than around $E[Y|\mathbf{x}]$, which is more usual.

According to the definition of Q- optimality, the loss function should be

$$
L_Q(f,\xi) = \int_\mathcal{S} E\left[\left\{\hat{Y}(\mathbf{x}) - E[Y|\mathbf{x}]\right\}^2\right] d\mathbf{x}.
$$

This decomposes as

$$
\begin{aligned}
L_Q(f,\xi) &= Q_v(\xi) + Q_b(f,\xi) + \int_\mathcal{S} f^2(\mathbf{x})d\mathbf{x} \\
&\leq \frac{\sigma^2_\varepsilon}{n} tr\left\{\mathbf{B}^{-1}\mathbf{A}\right\} + ch_{\max}(\mathbf{Q}) \cdot \|f\|^2_{\mathcal{H}_\mathcal{K}} + \int_\mathcal{S} f^2(\mathbf{x})d\mathbf{x} \\
&= \left(\sigma^2_\varepsilon + \eta^2_{\mathcal{H}_k}\right)\left(rJ_v(\xi) + (1-r)J_b(\xi)\right) + \int_\mathcal{S} f^2(\mathbf{x})d\mathbf{x}.
\end{aligned}
$$

To compare this design with others, we will fix the function f such that it is orthogonal to the regressors. For example, in this report, $f(x)$ may be quadratic on $\mathcal{S} = [-\frac{1}{2}, \frac{1}{2}]$:

$$f(x) = \sqrt{\frac{5}{4}}(12x^2 - 1). \tag{12}$$

In this case $\int_{\mathcal{S}} f^2(x)dx = 1$ and we find that $\|f\|^2_{\mathcal{H}_\mathcal{K}} = \int_{\mathcal{S}} \left[f^{'}(x)\right]^2 dx + \left[\int_{\mathcal{S}} f(x)dx\right]^2 = 60$. If $f(x)$ is cubic

$$f(x) = \sqrt{7}(20x^3 - 3x) \tag{13}$$

then $\int_{\mathcal{S}} f^2(x)d\mathbf{v} = 1$ and $\|f\|^2_{\mathcal{H}_\mathcal{K}} = \int_{\mathcal{S}} \left[f^{'}(x)\right]^2 d\mathbf{x} + \int_{\mathcal{S}} \left[f^{''}(x)\right]^2 dx + \left[\int_{\mathcal{S}} f(x)dx\right]^2 = 8568$.

3.1 Construction of $\mathcal{H}_\mathcal{K}$ satisfying (7)

Theorem 2 in Y&H shows a way to construct RKHS which is different from Wahba (1978 Section 3) since Y&H consider the condition of orthogonality when they construct RKHS.

As it is known, choosing a RKHS is equivalent to choosing a basis $\{\phi_j\}_{j=1}^{\infty}$ and λ_j. It is also equivalent to choosing a reproducing kernel K. Now we want to build up a space which is a RKHS but also satisfies $\int_{\mathcal{S}} \mathbf{z}(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} = \mathbf{0}$. The basic idea is to build a RKHS $\mathcal{F}$ with kernel K_0. The RKHS $\mathcal{F}$ can be decomposed into two parts: $\mathcal{H}_\mathcal{K}$ which satisfies the orthogonal condition as above, and others are the span$\{z_1, \cdots, z_p\}$, i.e.,

$$\mathcal{F} = \{z_1, \cdots, z_p\} \oplus \mathcal{H}_\mathcal{K}, \qquad \int_{\mathcal{S}} \mathbf{z}(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} = \mathbf{0}, \quad f(\mathbf{x}) \in \mathcal{H}_\mathcal{K}$$

where $z_1, \cdots, z_p$ is a set of linearly independent functions from $\mathcal{F}$. The objective is to find the kernel K for the RKHS $\mathcal{H}_\mathcal{K}$.

One of the useful properties of reproducing kernels is that from them one can obtain the representer of any bounded linear functional. Define

$$T_j(f) = \int_{\mathcal{S}} z_j(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}, \quad f \in \mathcal{F}, \quad j = 1, 2, \cdots, p. \tag{14}$$

Let η_j be the representer for T_j on $\mathcal{F}$, that is

$$< \eta_j, f >= T_j(f), \quad f \in \mathcal{F}, \quad j = 1, 2, \cdots, p \tag{15}$$

For fixed $z_j \in \mathcal{F}$, We have

$$\eta_j =< \eta_j, K_0(\cdot, \mathbf{x}) >= T_j(K_0(\cdot, \mathbf{x})) = \int_{\mathcal{S}} z_j(\mathbf{w}) K_0(\mathbf{w}, \mathbf{x}) d\mathbf{w}$$

Further from (15) and (14), we have

$$< \eta_i, \eta_j >= T_i(\eta_j) = \int_{\mathcal{S}} z_i(\mathbf{x})\eta_j(\mathbf{x})d\mathbf{x} = \int_{\mathcal{S}\times\mathcal{S}} z_i(\mathbf{w})z_j(\mathbf{w})K_0(\mathbf{w},\mathbf{x})d\mathbf{x}d\mathbf{w},$$

which forms a matrix Ψ with the (i,j)th entry $< \eta_i, \eta_j >$. Let $\boldsymbol{\eta}$ be a vector of p functions $\int_{\mathcal{S}} z_j(\mathbf{w})K_0(\mathbf{w},\mathbf{x})d\mathbf{w}$, $j = 1, 2, \cdots p$. Y&H showed that kernel K for the RKHS $\mathcal{H}_{\mathcal{K}}$ has the form

$$K(\mathbf{x},\mathbf{w}) = K_0(\mathbf{x},\mathbf{w}) - \boldsymbol{\eta}^T(\mathbf{x})\Psi^{-1}\boldsymbol{\eta}(\mathbf{w}) \tag{16}$$

It is easy to show that $\eta_1, \cdots \eta_p$ are linearly independent because $\mathbf{A} = \int_{\mathcal{S}} \mathbf{z}(\mathbf{x})\mathbf{z}^T(\mathbf{x})d\mathbf{x}$ is nonsingular. Hence Ψ is nonsingular.

The function K given in (16) satisfies an orthogonality condition. For any fixed $\mathbf{w} \in \mathcal{S}$, we have for each j

$$\begin{aligned}\int_{\mathcal{S}} z_j(\mathbf{x})K_0(\mathbf{x},\mathbf{w})d\mathbf{x} &= < \eta_j, K(\cdot,\mathbf{w}) > \\ &=< \eta_j, K_0(\mathbf{x},\mathbf{w}) > - < \eta_j, \boldsymbol{\eta}^T(\mathbf{x}) > \Psi^{-1}\boldsymbol{\eta}(\mathbf{w}) \\ &= \eta_j(\mathbf{w}) - \mathbf{e}_j^T\Psi\Psi^{-1}\boldsymbol{\eta}(\mathbf{w}) = 0\end{aligned}$$

It follows that $K(\cdot,\mathbf{w}) \in \mathcal{H}_{\mathcal{K}}$, for any $f \in \mathcal{H}_{\mathcal{K}} \subset \mathcal{F}$,

$$< f, K(\cdot,\mathbf{w}) >=< f, K_0(\mathbf{x},\mathbf{w}) > - < f, \boldsymbol{\eta}^T(\mathbf{x}) > \Psi^{-1}\boldsymbol{\eta}(\mathbf{w}) = f(\mathbf{w})$$

since $< f, \eta_j >= \int_{\mathcal{S}} z_i(\mathbf{x})f(\mathbf{x})d\mathbf{x} = 0$, for $j = 1, 2, \cdot\cdot, p$. Therefore, K is a reproducing kernel for $\mathcal{H}_{\mathcal{K}}$.

3.2 Examples

Rong-Xian Yue and Fred J. Hickernell gave several examples in their paper. Here we summarize and discuss Example 2 in Section 3 of Y&H with $s = 1$. This is a linear regression model with a one-dimensional design space. For $\mathbf{x} = x \in [0,1]$ we set $\Phi_1(x) = 2\sqrt{3}B_1(x)$, where $s = 1$, and where B_1 is the first Bernoulli polynomial $B_1(x) = x - 0.5$. The regressor here is $z(x) = (1, \Phi_1(x))^T$. From (16), the reproducing kernel for $\mathcal{H}_{\mathcal{K}}$ is

$$K(x,w) = [1 + B_1(x)B_1(w) + \frac{1}{2}B_2(\{x-w\})] - 1 - \frac{5}{6}B_{1,3}(x)B_{1,3}(w),$$

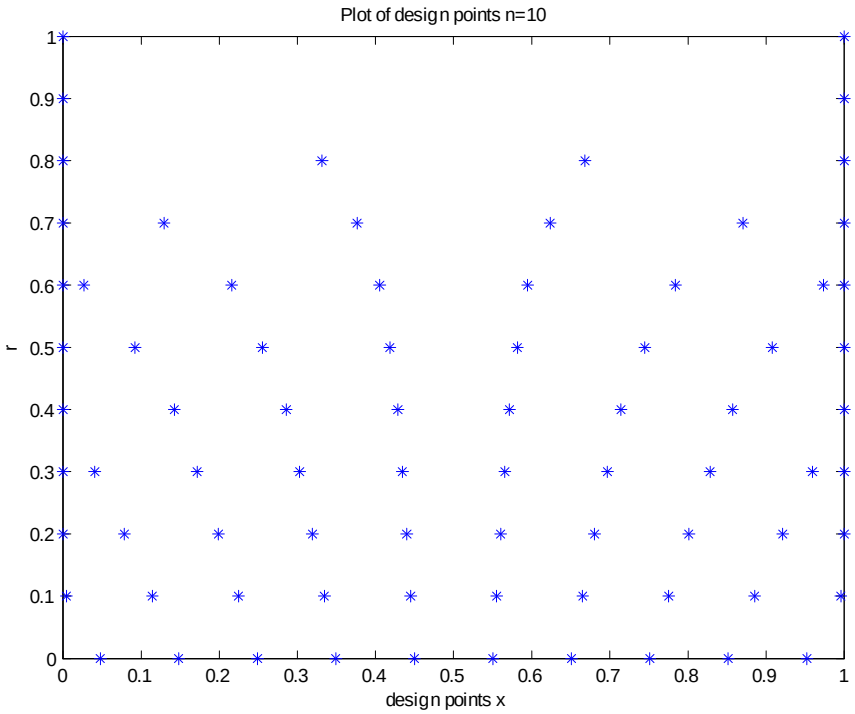

Figure 1: RKHS design for linear regression with $n = 10$.

where $B_{1,3}(\cdot) = B_1(\cdot) - 2B_3(\cdot)$ is given by

$$B_{1,3}(\cdot) = -2x^3 + 3x^2 - 0.5,$$

since

$$\begin{aligned} B_2(x) &= x^2 - x + \frac{1}{6}, \\ B_3(x) &= x^3 - \frac{3}{2}x^2 + \frac{1}{2}x. \end{aligned}$$

Here B_l is the l^{th} Bernoulli polynomial, and $\{\cdot\}$ is the fractional part of a real number.

The function $f_{\max}$ is the function in RKHS in Y&H Example 2 that maximizes the loss function $\int_{\mathcal{S}} E\left[\left\{\hat{Y}(x) - \mathbf{z}^T(x)\boldsymbol{\theta}_0\right\}^2\right]$ over $\mathcal{H}_{\mathcal{K}}$ subject to $||f||_{\mathcal{H}_{\mathcal{K}}} = 1$. This function is determined by the reproducing kernel K from

$$f_{\max} = [K(w,w)]^{1/2} K(x,w).$$

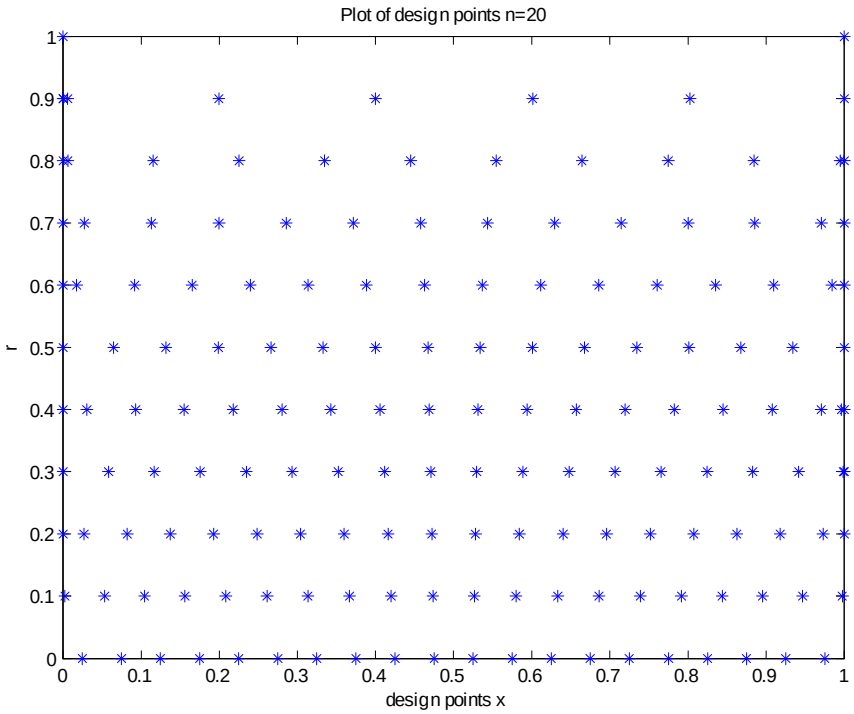

Figure 2: RKHS design for linear regression with $n = 20$.

We numerically calculate the optimal design that minimizes $rJ_v(\xi) + (1-r)J_b(\xi)$ over contamination class in design space $[0,1]$, for a given $r \in [0,1]$. After obtaining the design points, we use a linear transformation to translate the design to $[-0.5, 0.5]$. For a fixed particular contamination function (12) or (13), we calculated the variance component and bias component with model prior value $r \in \{0, 0.1, 0.3, 0.5, 0.8\}$. Here the variance and bias are defined as:

$$variance = tr\left\{\mathbf{B}^{-1}\mathbf{A}\right\},\ bias = \mathbf{b}^T\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{b}, \tag{17}$$

Table 1 Bias and variance values for RKHS design at (12) and (13)

contamination function as (12)				contamination function as (13)			
	r	variance	bias		r	variance	bias
$n = 10$	0	1.9999	3.8713e-09	$n = 10$	0	1.9999	0.0009
	.1	1.8323	0.0507		.1	1.8323	0.1803
	.3	1.6808	0.2748		.3	1.6808	0.5033
	.5	1.5683	0.7212		.5	1.5683	0.9347
	.8	1.4052	2.6939		.8	1.4052	2.0569
	.9	1.3333	5.0000		.9	1.3333	2.3333
	1	1.3333	5.0000		1	1.3333	2.3333
$n = 20$	0	2.0014	2.3535e-06	$n = 20$	0	2.0014	0.0001
	.1	1.9022	0.0147		.1	1.9022	0.0473
	.3	1.7644	0.1187		.3	1.7644	0.3013
	.5	1.6617	0.3267		.5	1.6617	0.6184
	.8	1.4768	1.5052		.8	1.4768	1.4615
	.9	1.3972	2.8787		.9	1.3972	1.9576
	1	1.3333	5.0000		1	1.3333	2.3333
$n = 40$	0	1.9996	2.2291e-07	$n = 40$	0	1.9996	0.0001
	.1	1.8838	0.0216		.1	1.8838	0.0401
	.3	1.7654	0.1174		.3	1.7654	0.2956
	.5	1.6561	0.3434		.5	1.6561	0.6027
	.8	1.4691	1.6011		.8	1.4691	1.4949
	.9	1.3990	2.8350		.9	1.3990	1.9738
	1	1.3333	5.0000		1	1.3333	2.3333

where

$$\begin{aligned}
\mathbf{A} &= \int_{-1/2}^{1/2} \mathbf{z}(\mathbf{x})\, \mathbf{z}^T(\mathbf{x})\, d\mathbf{x} = \begin{pmatrix} 1 & 0 \\ 0 & 1/12 \end{pmatrix}, \\
\mathbf{B} &= \int_{-1/2}^{1/2} \mathbf{z}(\mathbf{x})\, \mathbf{z}^T(\mathbf{x})\, \xi(d\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}(\mathbf{x}_i)\, \mathbf{z}^T(\mathbf{x}_i), \\
\mathbf{b} &= \int_{-1/2}^{1/2} \mathbf{z}(\mathbf{x})\, f(\mathbf{x})\, \xi(d\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}(\mathbf{x}_i)\, f(\mathbf{x}_i).
\end{aligned}$$

In Table 1, we calculate the variance and bias for particular contamination functions. The variance component values of RKHS designs are less than those for Huber's design. The bias component values of RKHS design are similar to those for Huber's design.

In Figures 1 and 2, RKHS designs are presented for $n = 10$ and $n = 20$,

respectively, for value of q ranging from 1 (all-variance design) to 0 (all-bias design). The design starts with uniform or close uniform design when one does not know the model prior and obtain the classical design with all mass at the two endpoints, when one knows the exact model. The RKHS design is a symmetric design.

4 Huber's minimax design for simple linear regression

Huber (1975) obtained the minimax design for fitting a linear regression model with misspecification of the form as (6). Upon minimizing the maximum, over $f \in \mathcal{F}_2$, value of $IMSE = L_Q(f, \xi)$, Huber gave robust optimal designs with $\mathbf{z}(x) = (1, x)^T$ and $\mathcal{S} = [-\frac{1}{2}, \frac{1}{2}]$. With

$$\begin{aligned} L_Q(f,\xi) &= \int_{\mathcal{S}} E\left[\left\{\hat{Y}(\mathbf{x}) - E[Y|\mathbf{x}]\right\}^2\right] d\mathbf{x} \qquad (18)\\ &= \sigma_\varepsilon^2 tr\left\{(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{A}\right\} + \mathbf{b}^T\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{b} + \int_{\mathcal{S}} f^2(\mathbf{x})d\mathbf{x}\\ &= (v\eta_s^2+1)\left(\frac{v\eta_s^2}{v\eta_s^2+1} tr\left\{\mathbf{B}^{-1}\mathbf{A}\right\} + \frac{1}{v\eta_s^2+1}\mathbf{b}^T\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{b}\right) + \int_{\mathcal{S}} f^2(\mathbf{x})d\mathbf{x}\\ &= const \cdot (q \cdot variance + (1-q)\cdot bias) + \eta_s^2 \end{aligned}$$

where $q = \frac{v\eta_s^2}{v\eta_s^2+1}$, $const = v\eta_s^2 + 1$, $\int_{\mathcal{S}} f^2(\mathbf{x})d\mathbf{x} = \eta_s^2 \le 1$, $v = \frac{\sigma_\varepsilon^2}{n\eta_s^2}$, the minimax design measure ξ_0 has a density function of the form $m_0 = (ax^2 + b)^+$. We denote by $f_{\max}$ the least favourable function.

The following description of the design is from Huber (1981)**.** The dependence of $(f_{\max}, \xi_0)$ on η_s can be described in parametric form with everything depending on the parameter γ. If $\gamma \in \left[\frac{1}{12}, \frac{3}{20}\right]$, the design measure is

$$\xi_0(x) = \int_{-\frac{1}{2}}^{x} m_0(t)dt = x + 0.5 + \frac{5}{4}(12\gamma - 1)(4x^3 - x) \qquad (19)$$

where the density function is

$$m_0(x) = 1 + \frac{5}{4}(12\gamma - 1)(12x^2 - 1) \qquad (20)$$

Table 2 Variance and bias values for Huber's design at (12) and (13)

contamination function as (12)				contamination function as (13)			
	q	variance	bias		q	variance	bias
$n = 10$	0	1.8182	0.0617	$n = 10$	0	1.8182	0.2223
	.1	1.7955	0.0826		.1	1.7955	0.2292
	.3	1.7481	0.1418		.3	1.7481	0.2477
	.5	1.6933	0.2447		.5	1.6933	0.2744
	.8	1.5659	0.7358		.8	1.5659	0.3412
	.9	1.4872	1.3850		.9	1.4872	0.4226
	1	1.3333	5.0000		1	1.3333	2.3333
$n = 20$	0	1.9048	0.0139	$n = 20$	0	1.9048	0.0542
	.1	1.8743	0.0258		.1	1.8743	0.0656
	.3	1.8125	0.0665		.3	1.8125	0.0927
	.5	1.7436	0.1486		.5	1.7436	0.1293
	.8	1.5931	0.5881		.8	1.5931	0.2319
	.9	1.5055	1.1965		.9	1.5055	0.3310
	1	1.3333	5.0000		1	1.3333	2.3333
$n = 40$	0	1.9512	0.0033	$n = 40$	0	1.9512	0.0133
	.1	1.9165	0.0104		.1	1.9165	0.0215
	.3	1.8467	0.0410		.3	1.8467	0.0439
	.5	1.7700	0.1115		.5	1.7700	0.0778
	.8	1.6063	0.5269		.8	1.6063	0.1828
	.9	1.5132	1.1245		.9	1.5132	0.2874
	1	1.3333	5.0000		1	1.3333	2.3333

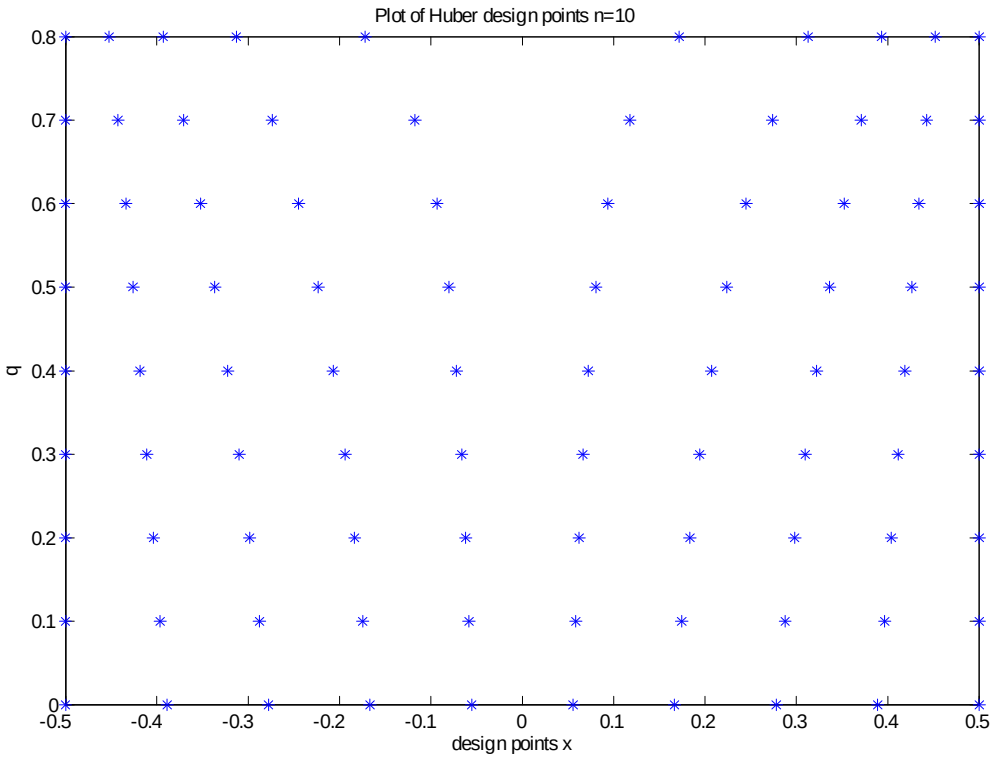

Figure 3: Huber's design for linear regression with $n = 10$.

and the least favourable function $f_{\max}$ is

$$f_{\max}(x) \propto 12x^2 - 1. \tag{21}$$

The parameters v and γ are related by

$$v = 360\gamma^2(12\gamma - 1).$$

If $\gamma \in \left[\frac{3}{20}, \frac{1}{4}\right]$, Huber changes the parameter to $c \in [0, 1)$, with no interpretation

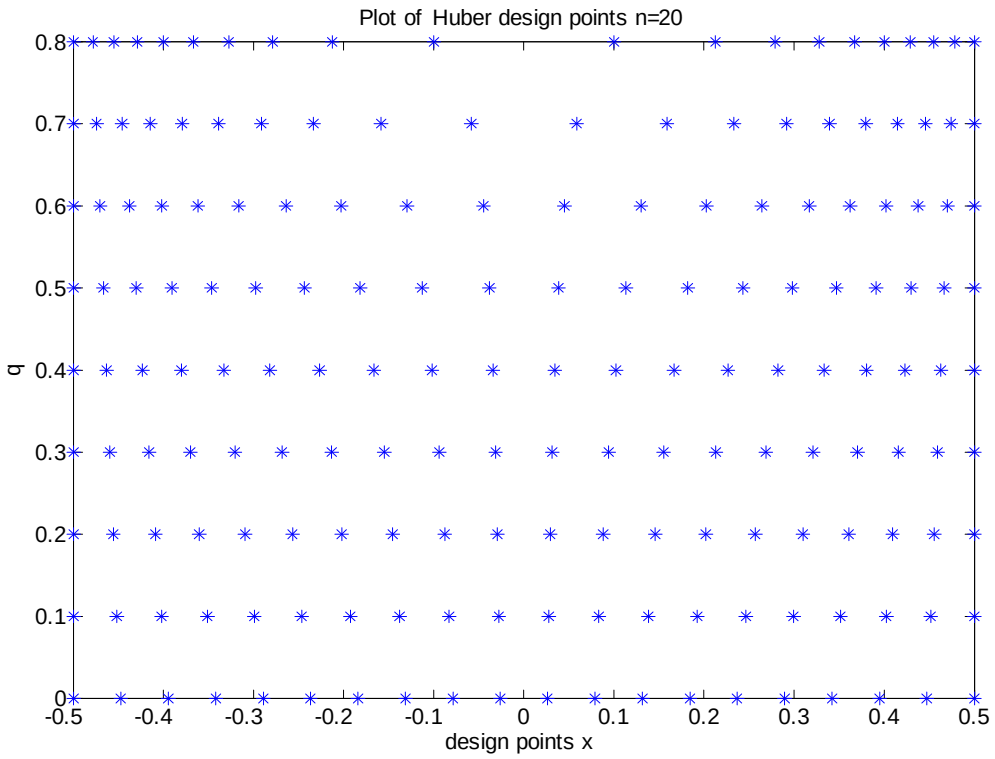

Figure 4: Huber's design for linear regression with $n = 20$.

of c. Then

$$
\begin{aligned}
a &= \frac{12}{(1+2c)(1-c)^2} \\
m_0(x) &= \frac{a}{4}(4x^2 - c^2)^+ = \begin{cases} \frac{a}{4}(4x^2 - c^2) & \text{if } |x| > \frac{c}{2} \\ 0 & \text{otherwise} \end{cases} \\
\gamma &= \frac{3 + 6c + 4c^2 + 2c^3}{20(1+2c)} \\
f_{\max}(x) &= [m_0(x) - 1]\,\varepsilon \\
\varepsilon^2 &= \frac{5(1-c)^3(1+2c)^5}{4(1+3c+6c^2+5c^3)} \\
v &= \frac{18(3+6c+4c^2+2c^3)^2}{25(1-c)^2(1+2c)^3}
\end{aligned}
$$

and

$$\xi_0(x) = \int_{-\frac{1}{2}}^{x} m_0(t)dt = \begin{cases} \frac{a}{24}(1 - 6c^2x - 3c^2 + 8x^3) & \text{if } x \in \left[\frac{-1}{2}, \frac{c}{2}\right] \\ \frac{1}{2} & \text{if } x \in \left[\frac{-c}{2}, \frac{c}{2}\right] \\ 1 - \frac{a}{24}(-8x^3 + 6c^2x + 1 - 3c^2) & \text{if } x \in \left[\frac{c}{2}, \frac{1}{2}\right] \end{cases} \quad (22)$$

This measure will minimize the loss function (18). But it cannot be implemented. We will choose quantiles $x_i = \xi_0^{-1}(\frac{i-1}{n-1})$, then the empirical distribution of the design tends to the true optimal design. We solve the equation

$$\xi_0(x) = \frac{i-1}{n-1} \quad i = 1, 2, 3..., n \quad (23)$$

where $\xi_0(x)$ is defined in (19) and (22). This approximates the optimal design $\xi_0(x)$ by a symmetric and discrete measure. Solving equation (23) for $n = \{10, 20, 40\}$ by MATLAB command "solve", we can get the design points. Using those design points, we calculate the variance component and the bias component at particular contamination functions (12) and (13). Table 2 shows the numerical results. The variance and bias are as defined at (17).

In Figures 3 and 4, Huber's designs are presented for $n = 10$ and 20, respectively, for values of q ranging from 0 to 0.1. When model prior q is equal to 1, Huber's design becomes the classical, variance minimizing design. At each of the point $\pm\frac{1}{2}$, the pointmass is $\frac{1}{2}$. In this case, the value of bias is 5.

5 Integer-Valued, Minimax Robust Designs of Fang/Wiens (2000)

Fang and Wiens (2000) found exact designs which are robust against departures from the assumed linear response function. They use an easily implemented simulated annealing algorithm that yields near-optimal solutions with no restriction on the fitted model or on the structure of the design space. Huber's approach has generally resulted in "designs" that are arbitrary and possibly continuous probability functions $\xi_0(x)$ on the design space. The number of observations allocated to a particular design point x_i is $n\xi_0(x)$, which need not be an integer. In this case, the annealing algorithm gives exact integer-valued designs so that $n\xi_0(x)$ is an integer. In this integer-valued design, we suppose to have a finite

set $\{x_i\}_{i=1}^N$ of possible design points from which the experimenter is interested in choosing n, not necessarily distinct, points at which to observe the response. The experimenter makes $n_i \geq 0$ observations at x_i such that $\sum_{i=1}^N n_i = n$. The design problem is to choose $n_1, ..., n_N$ in an optimal manner. Equivalently, the objective is to find an optimal probability distribution $\{p_i\}_{i=1}^N$, with $p_i = n_i/n$, on the design space $\{x_i\}_{i=1}^N$. The resulting design is said to be integer valued.

5.1 Simulated annealing algorithm for minimax design

We consider models with regressors $(1, x)^T$ where $x \in [-1, 1]$. Given the desired number of observations n to be taken and the number of points in the design space N, we seek design to minimize the average mean squared error (AMSE) loss function. Our algorithm accommodates all (n, N) combinations. We take φ to be the set $\{x_i = -1/2 + 1 \times (i-1)/(N-1)\}_{i=1}^N$ of equally spaced points in $\left[-\frac{1}{2}, \frac{1}{2}\right]$.

Simulated annealing is used to search for optimal designs. The simulated annealing algorithm seeks to assign integers $n_i \geq 0$ to each of the design points x_i in such a way that the maximum, over contaminants f, of the average mean squared error (AMSE) is a minimum. The AMSE I is defined in Fang and Wiens (2000); for our purposes it is given by

$$\begin{aligned} I &= \frac{1}{N}\sum_{i=1}^{N} E\left[\left(\hat{Y}(\mathbf{x}_i) - E\left[Y(\mathbf{x}_i)\right]\right)^2\right] \\ &= \frac{\sigma_\varepsilon^2}{n} tr\left\{\mathbf{B}^{-1}\mathbf{A}_N\right\} + \mathbf{b}^T\mathbf{B}^{-1}\mathbf{A}_N\mathbf{B}^{-1}\mathbf{b} + \frac{1}{N}\sum_{i=1}^{N} f^2\left(\mathbf{x}_i\right), \end{aligned}$$

with

$$\mathbf{A}_N = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}\left(\mathbf{x}_i\right)\mathbf{z}^T\left(\mathbf{x}_i\right).$$

Simulated annealing is a direct search optimization algorithm which has been quite successful at finding the global extreme of a function, possibly nonsmooth, that has many local extrema. In an unpublished University of Alberta Ph.D. thesis, Adewale (2006) describes the algorithm as a biased random walk consisting of three steps:

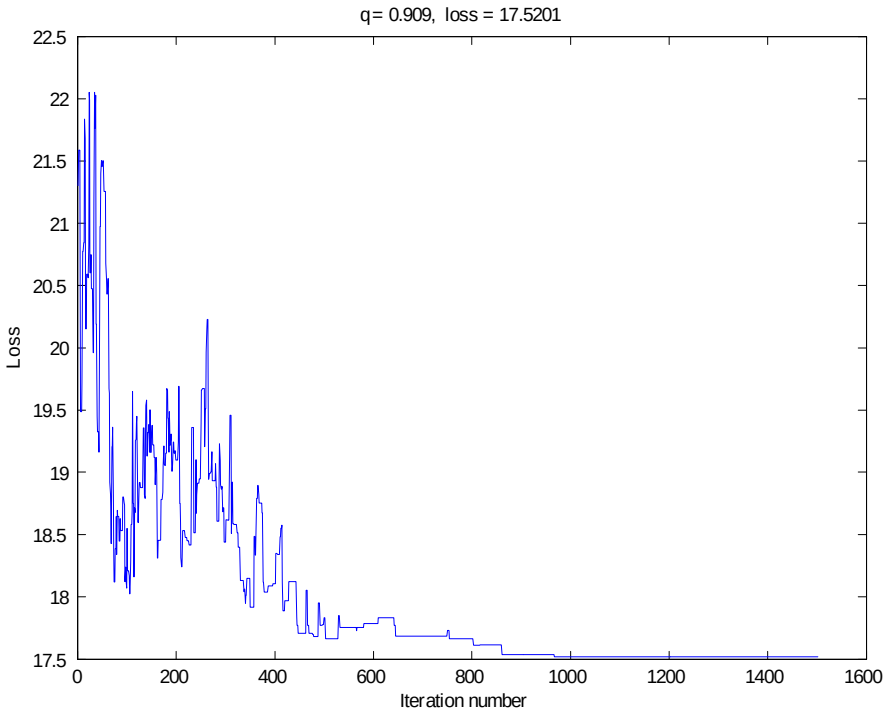

Figure 5: Simulated annealing trajectory for simple linear regression with $N = 40, n = 20$ and $q = 10/11$.

1. A description of the initial state of the process; that is, of the starting vector of allocation $\mathbf{n}_0 = (n_1, ..., n_N)$. We can calculate the corresponding initial value of the loss function $L_Q(\mathbf{n}_0)$ based on the initial state.

2. The second step is the random choice of the next state of the process from the optimization space. The subsequent states are generated.

3. A criterion according to which these subsequent states are accepted or rejected.

Fang and Wiens (2000) assumed that one of (n, N) is a multiple of the other and then chose the initial design to be as uniform as possible. If $n > N$, then initial state is the uniform design, with $n_i = n/N$ for $i = 1, ..., N$. If $n \leq N$, then this vector of frequencies assigned to $x_1, ...x_N$ starts with the vector formed by repeating the vector $(1, 0, ...0)$ (with $N/n - 1$ 0's) $[n/2]$ times. This is followed by the same vector with the order of its elements reversed. If N is odd, then also

a vector $(0, ...0, 1, 0, ...0)$ of length N/n is inserted in the middle. Thus in either case the initial design is symmetric and at least close to uniform.

To generate a new design, we perturb the current state as following. Define vector $\mathbf{v}$ to be the $N \times 1$ current allocation vector. For symmetric designs redefine $\mathbf{v}$ to be the $[N/2] \times 1$ vector consisting of the initial segment $(n_1, ..., n_{[N/2]})$ of the current allocation vector. Let $J_+ = \{i | v_i > 0\}, J_+ = \{i | v_i > 0\}$, $J_0 = \{i | v_i = 0\}$ with cardinalities $j_+ \geq 1$ and j_0. If $j_+ \geq 2$, generate a Bernoulli random variable

$$B = \begin{cases} 1 & \text{with probability } j_0/(j_0 + j_+) \\ 0 & \text{with probability } \ j_+/(j_0 + j_+) \end{cases}$$

choose two indices (t_1, t_2) from J_+, at random without replacement, choose an index t_0 from J_0, at random and modify the selected components of $\mathbf{v}$ as follows:

$$v_{t_0} = v_{t_0} + B; \ v_{t_1} = v_{t1} - 1, \ v_{t_2} = v_{t_2} + 1 - B$$

If $j_+ = 1$, choose t_0 from J_0 at random, let t_1 be the index in the singleton set J_+, and then replace v by

$$v_{t_0} = v_{t_0} + 1, \ v_{t_1} = v_{t1} - 1.$$

This completes the perturbation scheme for general designs. For symmetric designs, we complete the scheme as follows. If N is even, let $\mathbf{n} = (n_1, ..., n_N) = (v_1, ..., v_{N/2}, v_{N/2}, ...v_1)$. If N is odd, then generate a uniform random variable u. If $u < 1/N$, with probability $\frac{1}{2}$ increase $n_{[N/2]+1}$ by 2 then randomly and symmetrically reduce the remaining n_i by 2; with probability $\frac{1}{2}$ reduce $n_{[N/2]+1}$ by 2 then randomly and symmetrically increase the remaining n_i. This step is omitted if $n_{[N/2]+1} < 2$. We then construct $\mathbf{n}$ as described above, with the inclusion of the new frequency $n_{[N/2]+1}$.

The value of loss function $L_Q = L_Q(\widetilde{\mathbf{n}})$ is evaluated at the new state $\widetilde{\mathbf{n}}$ and the state is accepted with probability π, defined as

$$\pi = \begin{cases} 1 & \text{if } \triangle L_Q \leq 0 \\ \exp(- \triangle L_Q/T) & \text{if } \triangle L_Q > 0 \end{cases}$$

where $\triangle L_Q = L_Q(\widetilde{\mathbf{n}}) - L_Q(\mathbf{n})$. Thus, a favourable state ($\triangle L_Q \leq 0$) is accepted with certainty and an unfavourable state is accepted according to a separate

Bernoulli experiment with success probability $\exp(-\triangle L_Q/T)$. We choose T such that initially the inequality $0.5 < \exp(-\triangle L_Q/T) < 0.9$ is satisfied. As long as $\exp(-\triangle L_Q/T) > 0$ an unfavourable state could be accepted, thus providing the possibility of the path leading out of local minima. To ensure that the process settles at a global minimum we progressively decrease T. Fang and Wiens (2000) decrease T by a factor of 0.9 after each 100 iterations.

In Figure 5 we present the example of designs obtained using $v = \frac{q}{1-q} = 10$, $f = \sqrt{\frac{5}{4}}(12x^2 - 1), \eta_s = 1$. The plot shows that loss function value will converge to its minimum value after long enough iteration.

After a large number of iterations (3000 was used), it is expected that the algorithm converges to a design with minimum loss (loss = 17.5201). In the following examples, we will get the designs for particular contamination functions.

5.2 Example

For the purposes of comparing the Fang and Wiens design with RKHS designs and Huber's design, we consider approximate simple linear regression with contamination functions as (12) and (13). Table 3 gives the values of bias and variance components. We continue to define variance and bias as at (17), but now

$$\begin{aligned} \mathbf{B} &= \int \mathbf{z}(\mathbf{x})\,\mathbf{z}^T(\mathbf{x})\,\xi(d\mathbf{x}) = \frac{1}{n}\sum_i n_i \mathbf{z}(\mathbf{x}_i)\,\mathbf{z}^T(\mathbf{x}_i), \\ \mathbf{b} &= \int \mathbf{z}(\mathbf{x})\,f(\mathbf{x})\,\xi(d\mathbf{x}) = \frac{1}{n}\sum_i n_i \mathbf{z}(\mathbf{x}_i)\,f(\mathbf{x}_i), \end{aligned}$$

where the summations are over the *distinct* design points $\{\mathbf{x}_i\}$ and n_i is the number of observations made at x_i.

From Table 3, we find that the variance values of the integer-valued designs of Fang/Wiens (2000) for both quadratic and cubic contamination functions are slightly less those for Huber's design. For the quadratic contamination function (12), the bias for Huber's design is smaller than bias component values of Fang/Wiens' integer-valued designs. But for the cubic contamination function (13), the bias component values of Fang/Wiens integer-valued designs are smaller than those for Huber's design for $q = \{0.5, 0.8\}$. Note that we are comparing these

Table 3 Variance and Bias values for integer-valued design of Fang/Wiens(2000)

contamination function as (12)				contamination function as (13)			
	q	variance	bias		q	variance	bias
$n = 10$	0	1.7948	0.0833	$n = 10$	0	1.7948	0.0547
	.1	1.5165	1.0958		.1	1.5165	0.2524
	.3	1.5165	1.0958		.3	1.5165	0.2524
	.5	1.5165	1.0958		.5	1.5165	0.2524
	.8	1.5165	1.0958		.8	1.5165	0.2524
	.9	1.4282	2.2284		.9	1.4282	0.8031
	1	1.3333	5.0000		1	1.3333	2.3333
$n = 20$	0	1.7454	0.1459	$n = 20$	0	1.7454	0.2623
	.1	1.5434	0.8825		.1	1.5434	0.1393
	.3	1.5434	0.8825		.3	1.5434	0.1393
	.5	1.5434	0.8825		.5	1.5434	0.1393
	.8	1.5434	0.8825		.8	1.5434	0.1393
	.9	1.4870	1.3867		.9	1.4870	0.3694
	1	1.3333	5.0000		1	1.3333	2.3333
$n = 40$	0	1.7885	0.0899	$n = 40$	0	1.7885	0.0863
	.1	1.5585	0.7817		.1	1.5585	0.0956
	.3	1.5573	0.7889		.3	1.5573	0.0964
	.5	1.5573	0.7889		.5	1.5573	0.0956
	.8	1.5573	0.7889		.8	1.5573	0.0964
	.9	1.4995	1.2552		.9	1.4995	0.2893
	1	1.3333	5.0000		1	1.3333	2.3333

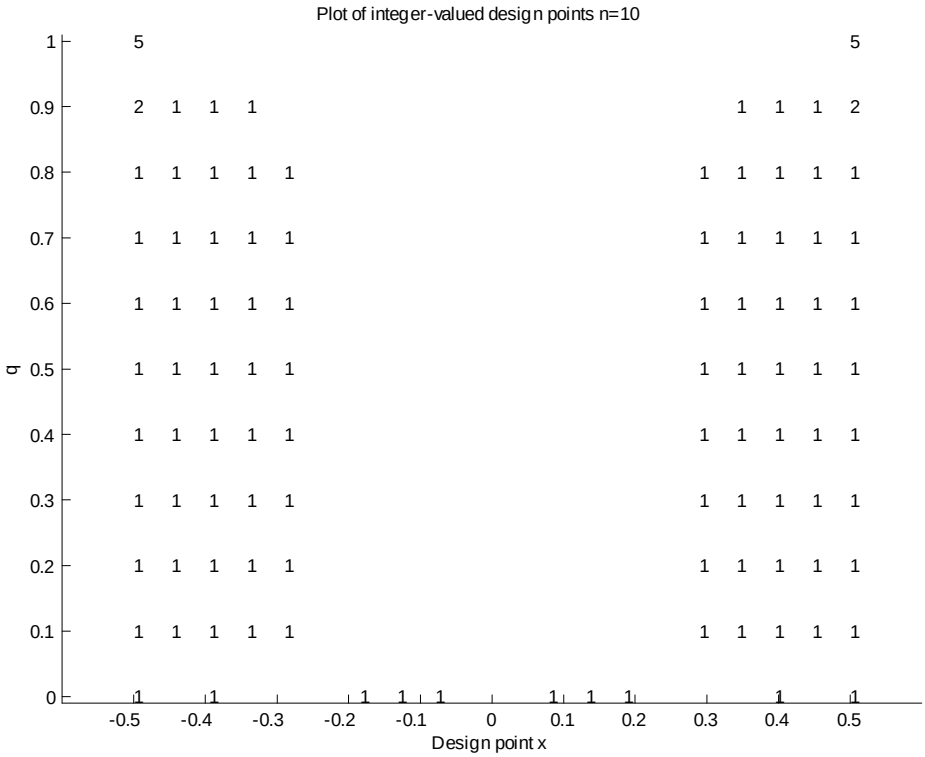

Figure 6: Integer-valued designs for linear regression with $N = 20$ and $n = 10$.

two designs from two design spaces. The integer-valued design is constructed on a finite design space. Huber's design is constructed in a continuous design space. At (13), the bias of annealing approach is smaller than that of RKHS design for $q = \{0.3, .0.5, 0.8, 0.9, 1\}$.

In Figure 6 and 7, we present designs for $n = 10$ and $n = 20$, respectively, for values of q ranging from 0 (all-bias design) to 1 (all-variance design) with contamination function (12). Instead of printing the symbol "*", we print the number of observations to be made at each distinct design points.

We find that the annealing algorithm will converge with a large number of iteration (3000 times). The resulting designs range from an almost uniform design ($q = 0$) to the classical design ($q = 1$). The classical design is the design that takes the assumed model to be exact and all pointmass are concentrate at the two endpoints of the design space. Also we find the designs are same when q changes from 0.1 to 0.8.

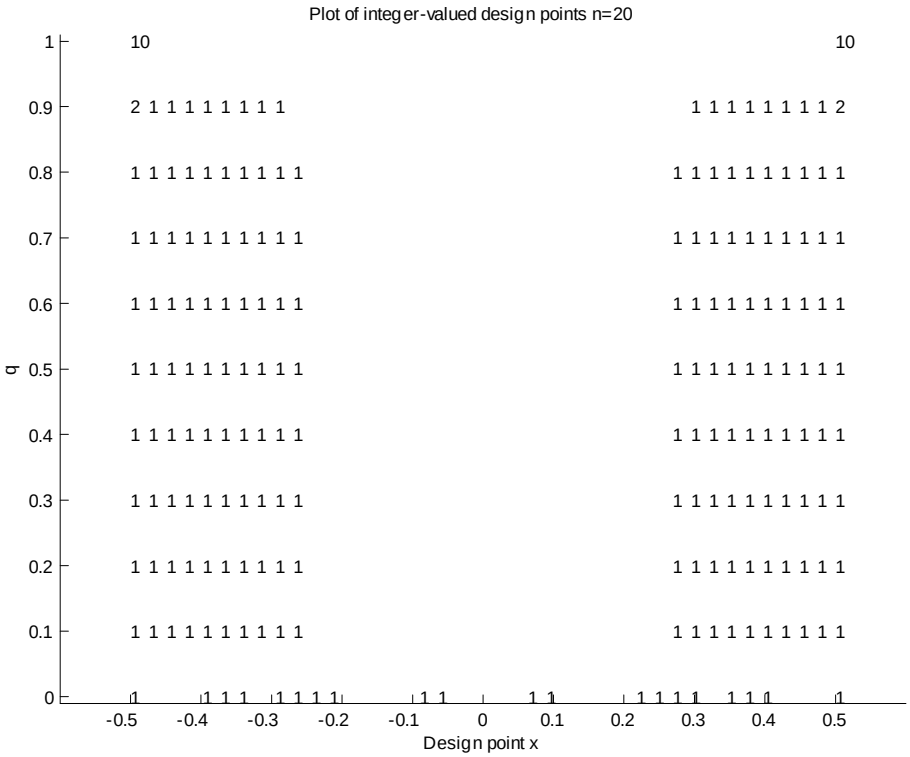

Figure 7: Integer-valued designs for linear regression with $N = 40$ and $n = 20$.

6 Conclusion

In this project we have implemented the RKHS, Huber, and Fang & Wiens design for straight line regression and compared the designs by variance and bias component values. In comparing these designs, we find RKHS design and Huber's design produce similar variance and bias component values, integer-valued design of Fang &Wiens gives larger bias. Fang and Wiens use finite design space but RKHS and Huber's design use continuous design space. Theoretically, Huber's design is a continuos design which can not be used in real life directly. In this project, practically we approximate it using a discrete design measure. RKHS design and integer-valued design are exact designs which are implementable. Huber's approach is feasible to only simple problems. But annealing approach and RKHS approach can be carried out for more complicated response function such as high degree polynomial regression. When model prior q increases from 0 to

1, all of the three methods produce classical design. That is, the design puts pointmass $\frac{1}{2}$ at each of the end points of design space. Symmetry is a common property for all of three designs.

From the computational point of view, Huber's design is simple to calculate. The annealing algorithm is less complicated compared to the RKHS method. It will take around 161 seconds in network server to calculate integer-valued design using annealing algorithm for 40 observations in 80 design point space. RKHS will take much longer time to calculate up to 40 observations (about 1917 seconds) because using optimization technique to solve nonlinear objective function is time-consuming. We use a restriction of symmetry; this reduces the computing time by about half. Also this searching solution procedure will heavily depend on initial points and is restricted by the number of design points.

References

Adewale A.J. (2006), Ph.D. thesis, "Robust Integer-Valued Designs for Generalized Linear Models", University of Alberta Department of Mathematical and Statistical Sciences.

Adewale A.J., Wiens D.P. (2006), "New Criteria for Robust Integer-Valued Designs in Linear Models," *Computational Statistics and Data Analysis,* in press.

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, 622-654.

Evgeniou, T., Pontil, M., and Poggio,T. (2000), "Regularization Networks and Support Vector Machines," *Advances in Computational Mathematics* 13, 1-50.

Fang, Z., and Wiens, D.P. (1999), "Robust Extrapolation Designs and Weights for Biased Regression Models With Heteroscedastic Errors," *The Canadian Journal of Statistics*, 27, 751-770.

Fang, Z., and Wiens, D.P. (2000), "Integer-valued, Minimax Robust Designs for Estimation and Extrapolation in Heteroscedastic, Approximately Linear Models," *Journal of the American Statistical Association*, 95, 807-818.

Fang, Z., and Wiens, D.P. (2003), "Robust Regression Designs for Approximate Polynomial Models," *Journal of Statistical Planning and Inference*, 117, 305 - 321.

Heo, G., Schmuland, B., and Wiens, D.P. (2001), "Restricted Minimax Robust Designs for Misspecified Regression Models," *The Canadian Journal of Statistics*, 29, 117-128.

Huber, P.J. (1975), "Robustness and Designs," in: *A Survey of Statistical Design and Linear Models*, ed. J.N. Srivastava, Amsterdam: North Holland, pp. 287-303.

Li, K.C., and Notz, W. (1982), "Robust Designs for Nearly Linear Regression," *Journal of Statistical Planning and Inference*, 6, 135-151.

Li, K.C. (1984), "Robust Regression Designs When the Design Space Consists of Finitely Many Points," *The Annals of Statistics*, 12, 269-282.

Marcus, M.B., and Sacks, J. (1976), "Robust Designs for Regression Problems," in: *Statistical Theory and Related Topics II*, ed. S.S. Gupta and D.S. Moore, New York: Academic Press, pp. 245-268.

Notz, W. (1989), "Optimal Designs for Regression Models With Possible Bias," *Journal of Statistical Planning and Inference*, 22, 43-54.

Oyet, A.J., and Wiens, D.P. (1997), "Robust Designs for Wavelet Approximations of Regression Models," *Journal of Nonparametric Statistics,* 12, 837-859.

Pesotchinsky, L. (1982), "Optimal Robust Designs: Linear Regression in R^k," *The Annals of Statistics*, 10, 511-525.

Schumaker, L.L. (1981), *Spline Functions: Basic Theory*, Wiley, New York.

Sinha, S. and Wiens, D.P. (2002), "Robust Sequential Designs for Nonlinear Regression," *The Canadian Journal of Statistics*, 30, 601-618.

Wahba, G. (1990), *Spline Models for Observational Data*, Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia.

Wahba, G. (1978), "Improper priors, spline smoothing and the problem of guarding against model errors in regression," *J. Roy. Stat. Soc., Ser. B.*, 40, 3, 364-372.

Wiens, D.P. (1991), "Designs for approximately linear regression: two optimality properties of uniform designs," *Statistics and Probability Letters*; 12, 217-221.

Wiens, D.P. (2000), "Bias Constrained Minimax Robust Designs for Misspecified Regression Models," *Selected Proceedings of the Third St. Petersburg Workshop on Simulation*; 117-133, Birkhauser, Boston.

Wiens, D.P. (2005), "Robustness in Spatial Studies II: Minimax Design," *Environmetrics*, 16, 205-217.

Wiens, D.P. (2005), "Robust Allocation Schemes for Clinical Trials With Prognostic Factors," *Journal of Statistical Planning and Inference*, 127, 323-340.

Wiens, D.P., and Zhou, J. (1997), "Robust Designs Based on the Infinitesimal Approach," *Journal of the American Statistical Association*, 92, 1503-1511.

Yue, R.-X., and Hickernell, F.J. (1999), "Robust Designs for Fitting Linear Models with Misspecification," *Statistica Sinica*, 9, 1053-1069.