The Canadian Journal of Statistics
Vol. 40, No. 2, 2012, Pages 362–377
La revue canadienne de statistique

Sequential design for nonparametric inference

Zhibiao ZHAO[1]* and Weixin YAO[2]

[1]*Department of Statistics, Penn State University, University Park, PA, USA*
[2]*Department of Statistics, Kansas State University, Manhattan, KS, USA*

Key words and phrases: Conditional distribution; conditional heteroscedasticity; nonparametric regression; optimal design density; quantile regression; sequential design.

MSC 2010: Primary 62L05; secondary 62G05.

Abstract: The performance of nonparametric function estimates often depends on the choice of design points. Based on the mean integrated squared error criterion, we propose a sequential design procedure that updates the model knowledge and optimal design density sequentially. The methodology is developed under a general framework covering a wide range of nonparametric inference problems, such as conditional mean and variance functions, the conditional distribution function, the conditional quantile function in quantile regression, functional coefficients in varying coefficient models and semiparametric inferences. Based on our empirical studies, nonparametric inference based on the proposed sequential design is more efficient than the uniform design and its performance is close to the true but unknown optimal design. *The Canadian Journal of Statistics* 40: 362–377; 2012 © 2012 Statistical Society of Canada

Résumé: La performance de l'estimation non paramétrique d'une fonction dépend souvent du choix des points dans le plan d'expérience. En nous basant sur le critère de l'erreur quadratique moyenne intégrée, nous proposons un plan d'expérience séquentiel qui met à jour la connaissance sur le modèle et la densité du devis optimal de façon séquentielle. Cette méthodologie est développée dans un cadre général regroupant une grande variété de problèmes d'inférence non paramétrique tels que : les fonctions moyennes et de variance conditionnelles, la fonction de répartition conditionnelle, la fonction quantile conditionnelle en régression quantile, les coefficients fonctionnels en modèles de coefficients variables et l'inférence semi-paramétrique. Basé sur nos études empiriques, l'inférence non paramétrique, basée sur le plan d'expérience séquentiel proposé, est plus efficace que celle faite en utilisant un devis uniforme et sa performance est proche de celle du plan d'expérience optimal inconnu. *La revue canadienne de statistique* 40: 362–377; 2012 © 2012 Société statistique du Canada

1. INTRODUCTION

In many applications, researchers want to draw inferences about an unknown model M based on available data. For a given input signal X_i to model M, we observe the output Y_i. Based on the inputs and outputs (X_i, Y_i), $i = 1, \ldots, n$, we use statistical methods to draw inferences about M. This procedure can be best described using the diagram:

$$\text{Input } X_i \rightarrow \text{ Model M} \rightarrow \text{Output } Y_i \rightarrow \text{Draw inference about M.} \tag{1}$$

In the above diagram, we have control of the inputs $\{X_i\}$, also known as the design points, and the statistical methods used to draw inferences.

To draw inferences about M, two popular classes of methods are the parametric and nonparametric approaches. In the literature on nonlinear experimental design, a great deal of research has

* *Author to whom correspondence may be addressed.*
E-mail: zuz13@stat.psu.edu

been done on sequential design for parametric inference. It is assumed that the model dynamics has a parametric form with unknown finite dimensional parameters, and the goal is to construct optimal design procedures based on certain criteria, for instance the Fisher information matrix. For a partial list, see Kiefer (1959), Ford & Silvey (1980), Abdelbasit & Plackett (1983), Wu (1985), Ford, Titterington, & Wu (1985), Chaudhuri & Mykland (1993), Dette, Melas, & Pepelyshev (2004), and the review paper by Ford, Titterington, & Kitsos (1989). Here we consider the nonparametric approach, which, by imposing no specific parametric model structure, allows model flexibility while reducing the risk of model mis-specification in parametric approaches; see the monograph by Li & Racine (2007) for nonparametric inferences based on uncontrolled experiments.

To understand the effect of design points on nonparametric inference, consider a simple case in which we want to estimate a piecewise constant function based on noisy observations. If we have no prior knowledge of how the underlying model responds to the design point or input, then one reasonable approach is to assign design points uniformly over the design interval. However, the uniform design strategy may not be the best choice when some model knowledge is available. For example, we may assign fewer points to regions where the model dynamics has a low noise level and more points to regions with large variations. As shown in Section 2.2, in some cases there could be a substantial loss of efficiency when one blindly uses the uniform design. Similarly, in Equation (1), if we have some knowledge about model M, then certain desirable statistical properties may be achieved by choosing proper design points $\{X_i\}$. In practice, however, we often have no prior model knowledge, and any knowledge has to be learned from the inputs and outputs. Therefore, it is important to study design strategy that is adaptive to unknown model dynamics.

Despite the vast literature on parametric experimental design, there are few references in the literature on experimental design in nonparametric inference. For nonparametric regression models, Müller (1984) was among the first to study optimal design for the derivatives of the mean regression function; however, his method was not adaptive and we must assume a priori the known conditional variance function of the errors. For nonparametric regression models with homoscedastic errors, Cheng, Hall, & Titterington (1998) considered sequential design for a local linear estimate of the mean function. More recently, Efromovich (2008) studied optimal design for nonparametric regression models with conditional heteroscedasticity using the Fourier series approach. For other contributions see Faraway (1990), Müller (1996), Park & Faraway (1998) and Biedermann & Dette (2001).

Our main purpose is to study adaptive optimal design strategies for a wide range of nonparametric inference problems. While most existing works deal with nonparametric mean regression function estimation, we proceed under a unified framework that covers several popular nonparametric inference problems, including the nonparametric mean regression function, the conditional variance function, the conditional distribution function, the conditional quantile function in quantile regression, functional coefficients in varying coefficient models and semiparametric inferences, among others. Under a general setting, we propose a sequential design procedure that updates model knowledge and design-point assignment sequentially. To implement the procedure, we obtain the optimal design density as an explicit function of a model dynamics related quantity. The proposed algorithm then works such that at each step the design points are drawn from the sequentially estimated optimal design density. As demonstrated through simulation studies, the proposed sequential design performs much better than the uniform design and is comparable to the optimal design.

The rest of this article is organized as follows. In Section 2 we study the efficiency loss of uniform design and present sequential design under a general framework. Section 3 concerns applications in several nonparametric inference problems. Simulation studies are carried out in Section 4 to illustrate the empirical performance of the proposed method.

2. METHODOLOGY

2.1. Optimal Design

We consider nonparametric inference under a general framework. For model M in Equation (1), denote by $m(x)$ a generic nonparametric function of interest; see Section 3 for examples. The kernel regression based nonparametric estimate often involves a kernel function $K(\cdot)$ and a bandwidth $b_n > 0$. Let $\hat{m}(x)$ be a nonparametric estimate of $m(x)$ using bandwidth b_n based on independent and identically distributed samples (X_i, Y_i), $1 \le i \le n$, from the population (X, Y). Popular choices of nonparametric estimation methods include the Nadaraya-Watson kernel smoother, local linear method, quantile regression and local M-estimation among others. For simplicity we assume that the design interval is $X \in [0, 1]$.

Assumption 1. Let $\hat{m}(x)$ be a nonparametric estimate of $m(x)$ using bandwidth b_n such that $b_n + (nb_n)^{-1} \to 0$. Denote by $f(x)$ the density function of X. Assume that there exist some functions $\rho(\cdot) \ge 0$ and $W(\cdot) \ge 0$ such that the mean squared error, abbreviated hereafter as MSE, of $\hat{m}(x)$ has the asymptotic expansion

$$\mathrm{MSE}\{\hat{m}(x)\} = E\big\{[\hat{m}(x) - m(x)]^2\big\} = b_n^4 \rho(x) + \frac{W(x)}{nb_n f(x)} + o\big\{b_n^4 + (nb_n)^{-1}\big\}, \tag{2}$$

uniformly over $x \in [0, 1]$. Further assume that $\rho(x)$ and $W(x)$ depend only on the model dynamics and do not depend on the design density $f(x)$.

Equation (2) is the well-known bias and variance decomposition of the mean squared error of the nonparametric estimates, with $b_n^2\sqrt{\rho(x)}$ and $W(x)/\{nb_n f(x)\}$ being the bias and variance, respectively, and $o\{b_n^4 + (nb_n)^{-1}\}$ the negligible error term. In Section 3 we show that Assumption 1 holds for many local linear nonparametric function estimates. For many nonparametric estimates, Equation (2) holds on a compact set $[\epsilon, 1 - \epsilon]$ for any $\epsilon > 0$ and may not hold at the boundaries. For simplicity we do not distinguish this.

Note that $\mathrm{MSE}\{\hat{m}(x)\}$ measures the performance of $\hat{m}(\cdot)$ at x. To evaluate the overall performance over the interval $[0, 1]$, consider the mean integrated squared error

$$\mathrm{MISE}\{\hat{m}\} = \int_0^1 \mathrm{MSE}\{\hat{m}(x)\}\mathrm{d}x = \overline{\mathrm{MISE}}\{\hat{m}|b_n, f\} + o\big\{b_n^4 + (nb_n)^{-1}\big\}, \tag{3}$$

where the leading term is

$$\overline{\mathrm{MISE}}\{\hat{m}|b_n, f\} = b_n^4 \int_0^1 \rho(x)\mathrm{d}x + \frac{1}{nb_n} \int_0^1 \frac{W(x)}{f(x)}\mathrm{d}x,$$

which depends on both the bandwidth b_n and the design density f. In an uncontrolled experiment, we choose the optimal bandwidth by minimizing $\overline{\mathrm{MISE}}\{\hat{m}|b_n, f\}$. Let

$$b_n^* = \operatorname*{argmin}_{b_n} \overline{\mathrm{MISE}}\{\hat{m}|b_n, f\} \quad \text{and} \quad \mathrm{MISE}^*\{\hat{m}|f\} = \overline{\mathrm{MISE}}\{\hat{m}|b_n^*, f\}. \tag{4}$$

After inserting the optimal bandwidth b_n^*, the goal of the controlled experiment is to find the optimal design density minimizing $\mathrm{MISE}^*\{\hat{m}|f\}$.

Theorem 1. *The optimal design density f^* minimizing* $\mathrm{MISE}^*\{\hat{m}|f\}$ *in Equation (4) is*

$$f^*(x) = \operatorname*{argmin}_f \mathrm{MISE}^*\{\hat{m}|f\} = \frac{\sqrt{W(x)}}{\int_0^1 \sqrt{W(x)}\mathrm{d}x}, \quad x \in [0,1]. \tag{5}$$

By Theorem 1, the optimal design density $f^*(x)$ is proportional to $W^{1/2}(x)$ or, equivalently, the asymptotic standard deviation of $\hat{m}(x)$, which agrees with the intuition that more design points are needed for areas where the model dynamics has a higher noise level. In particular, if $W(x)$ is a constant function, then $f^*(x)$ is the uniform density. In the special case of the mean regression function in conditional heteroscedastic models, Efromovich (2008) obtained the same optimal design density using the Fourier series approach. Here our framework is more general.

2.2. Relative Efficiency Loss of Uniform Design

By the definition of f^* in Equation (5), f^* is the most efficient design based on the MISE criterion. For any sub-optimal design f, there is a loss of efficiency.

Definition 1. *Let f be any design density. We define its relative efficiency loss, compared to the optimal design density f^* in Equation (5), as*

$$\mathrm{REL}(f) = \left[1 - \frac{\mathrm{MISE}^*\{\hat{m}|f^*\}}{\mathrm{MISE}^*\{\hat{m}|f\}}\right] \times 100\%.$$

Clearly, $0 \le \mathrm{REL}(f) \le 1$. If $\mathrm{REL}(f) \approx 0$, f is close to the optimal design. If $\mathrm{REL}(f) \approx 1$, there is almost a complete loss of efficiency. Throughout the rest of this article, denote by $f_U(x) = \mathbf{1}_{x\in[0,1]}$ the uniform density on $[0, 1]$, where and hereafter $\mathbf{1}$ is the indicator function. In many applications, we have no prior model knowledge and the uniform design is a reasonable choice. Proposition 1 below studies the relative efficiency loss of the uniform design in comparison with the optimal design.

Proposition 1. *For the uniform design f_U, its relative efficiency loss is*

$$\mathrm{REL}(f_U) = 1 - \left[\frac{\left\{\int_0^1 \sqrt{W(x)}\mathrm{d}x\right\}^2}{\int_0^1 W(x)\mathrm{d}x}\right]^{4/5}. \tag{6}$$

In the examples below we calculate the relative efficiency loss for some choices of $W(\cdot)$. We denote by $c > 0$ a generic constant that may vary from place to place.

Example 1. In Equation (5), let $W(x) = cx^r$, $x \in [0, 1]$, for $r \ge 0$. Then $f^*(x) = (r/2 + 1)x^{r/2}$ and $\mathrm{REL}(f_U) = 1 - \{4(r + 1)/(r + 2)^2\}^{4/5}$. If $r \to \infty$, $\mathrm{REL}(f_U) \to 1$ and there is almost 100% loss of efficiency. For $r = 1, 2, \ldots, 6$, $\mathrm{REL}(f_U) \approx 9\%, 20\%, 30\%, 38\%, 44\%, 48\%$.

Example 2. In Equation (5), let $W(x) = c\cos^2(k\pi x)$, $x \in [0, 1]$, for $k \in \mathbb{Z}$. Then $f^*(x) = 0.5\pi|\cos(k\pi x)|$ and $\mathrm{REL}(f_U) = 1 - (8/\pi^2)^{4/5} \approx 15\%$ for all $k \ne 0$.

Example 3. In Equation (5), let $W(x) = c\rho^x$, $x \in [0, 1]$, for $\rho > 0$. Then

$$f^*(x) = \frac{\rho^{x/2}\log(\rho)}{2(\sqrt{\rho} - 1)} \quad \text{and} \quad \mathrm{REL}(f_U) = 1 - \left\{\frac{4(\sqrt{\rho} - 1)}{(\sqrt{\rho} + 1)\log(\rho)}\right\}^{4/5}.$$

Clearly, $\mathrm{REL}(f_U) \to 1$ as $\rho \to 0$ or $\rho \to \infty$. If $\rho = 0.001$, $\mathrm{REL}(f_U) \approx 39\%$.

Example 4. In Equation (5), let $W(x) = \mathbf{1}_{x<0.5} + c\mathbf{1}_{x\geq 0.5}$, $x \in [0, 1]$. Then the noise level is different over the two intervals [0, 0.5) and [0.5, 1]. We have

$$f^*(x) = \frac{2}{1+c}\{\mathbf{1}_{x<0.5} + c\mathbf{1}_{x\geq 0.5}\} \quad \text{and} \quad \mathrm{REL}(f_U) = 1 - \left\{\frac{1+c+2\sqrt{c}}{2(1+c)}\right\}^{4/5}.$$

Then $\mathrm{REL}(f_U) \to 1 - 0.5^{4/5} \approx 43\%$ as $c \to 0$ or ∞.

As demonstrated by the above examples, under certain conditions, the uniform design may suffer from a substantial loss of efficiency. In Section 2.3 we propose a batch-sequential design procedure and prove that its efficiency loss goes to zero.

2.3. Batch-Sequential Design

If $W(x)$ is known, we can draw design points $X_1, \ldots, X_n$ from the optimal design density $f^*(x)$ in Equation (5). In practice, $W(x)$ is often unknown and Equation (5) is not directly applicable. To overcome this issue, we propose a batch-sequential design procedure.

First, we introduce the basic idea. If we have a consistent estimate $\hat{W}(x)$ of $W(x)$, then we can use the plug-in estimator by replacing $W(x)$ in Equation (5) with $\hat{W}(x)$. In the initial step, we have no knowledge about $W(x)$ and use the uniform design density $f_U(x)$. Based on the initial observations, we can obtain the estimate $\hat{W}(x)$ and the plug-in design density estimator $\hat{f}^*(x)$. In the second step, we draw design points from $\hat{f}^*(x)$ and update estimates $\hat{W}(x)$ and $\hat{f}^*(x)$, taking into account the new observations. The above procedure is repeated until the desired sample size n is achieved. We summarize this using the diagram

$$\text{Input } X_i \to \text{Model M} \to \text{Output } Y_i \to \text{Draw inference about M} \to \text{Input } X_i \to \cdots$$

To implement the above procedure, let k_n be a positive integer representing the batch size. For notational simplicity, we assume $\ell_n = n/k_n$ is an integer. Define the batches $\mathcal{I}_j = \{(j-1)k_n + 1, \ldots, jk_n\}$, $j = 1, \ldots, \ell_n$. We propose the following algorithm:

(P1) Let the initial design density $\hat{f}^*(x) = f_U(x)$ be the uniform density on [0, 1].
(P2) For each step $j = 1, \ldots, \ell_n$, repeat the following procedure:
  (S1) draw k_n random design points X_i, $i \in \mathcal{I}_j$, from the estimated optimal design density $\hat{f}^*(x)$ and record the corresponding observed outputs Y_i, $i \in \mathcal{I}_j$.
  (S2) based on (X_i, Y_i), $i \in \mathcal{I}_1 \cup \cdots \cup \mathcal{I}_j$, obtain updated estimates $\hat{m}(x)$ and $\hat{W}(x)$.
  (S3) update the design density

$$\hat{f}^*(x) = \frac{\hat{W}^{1/2}(x)}{\int_0^1 \hat{W}^{1/2}(x)\mathrm{d}x}. \tag{7}$$

  (S4) update j to $j+1$ and go to step S1.
(P3) After completing all iterations, we record the final estimates $\hat{m}(x)$, $\hat{W}(x)$, $\hat{f}^*(x)$.

We make some comments about the proposed algorithm.

First, for parametric models, batch-sequential designs have been studied in, for instance, Draper & Hunter (1967), Hohmann & Jung (1975), Ford & Silvey (1980), Abdelbasit & Plackett (1983), and references in Ford, Titterington, & Kitsos (1989). Our sequential design can be viewed as a nonparametric version of the parametric sequential design. Cheng, Hall, & Titterington (1998) and Efromovich (2008) studied sequential design for nonparametric models under different contexts.

Second, in step P2, if $\hat{W}(x)$ is a consistent estimate of $W(x)$ at each step j, then $\hat{f}^*(x)$ will be a consistent estimate of the optimal design density $f^*(x)$. Therefore, the design points are drawn from the asymptotically optimal design density at each step. Methods for constructing consistent estimates $\hat{W}(x)$ are discussed in Section 3.

Third, the proposed sequential design has appealing features of both the uniform design and the optimal design. When $W(\cdot) \approx 0$ at some region $\mathcal{A}$, it is difficult to draw design points near $\mathcal{A}$ from the optimal design density. In contrast, for the sequential design, the first batch of uniform design points copes well with this issue. On the other hand, after the first batch, the sequential design becomes asymptotically the optimal design and can adapt to unknown model dynamics. Thus, the sequential design has the appealing sampling property of the uniform design and the adaptiveness of the unknown optimal design.

When $k_n \to \infty$, excluding the first batch, all subsequent design points are drawn from the asymptotically consistent optimal design density. More formally, for each $j = 1, \ldots, \ell_n$, denote by $(X_{j,s}, Y_{j,s})$, $s = 1, \ldots, k_n$, the design points and outputs in batch j, then $X_{j,1}, \ldots, X_{j,\ell_n}$ are random samples from a common density, denoted by f_j. In general, it is difficult to study the asymptotic properties of the sequential design procedure due to the dependence in the inputs, and here we shall consider a slightly simplified problem. Notice that, conditioning on the design points, the outputs are conditionally independent. Intuitively, at the end of the sequential design procedure, all pooled inputs $X_{1,1}, \ldots, X_{1,k_n}, \ldots, X_{\ell_n,1}, \ldots, X_{\ell_n,k_n}$ can be viewed as samples from the mixture density $\bar{f} = (f_1 + \cdots + f_{\ell_n})/\ell_n$ with $f_1 = f_U$. In Theorems 2–3 below, we assume that $\inf_{x\in[0,1]} W(x) > 0$ and $\sup_{x\in[0,1]} W(x) < \infty$.

Theorem 2. *Suppose that there exists some $\theta \in (0, 1)$ such that $f_j = f^* + O\{(jk_n)^{-\theta}\}$ uniformly. Assume that n IID design points are drawn from the mixture density $\bar{f} = (f_U + f_2 + \cdots + f_{\ell_n})/\ell_n$. As $k_n \to \infty$ and $\ell_n \to \infty$, $\mathrm{REL}(\bar{f}) = O(1/\ell_n + n^{-\theta})$.*

For all the nonparametric estimates in Section 3 below, the optimal rate corresponds to $\theta = 2/5$. Thus, when $\ell_n \asymp n^{2/5}$, $\mathrm{REL}(\bar{f}) = O(n^{-2/5})$, which can not be further improved using larger ℓ_n. Our empirical studies show that $\ell_n = 3, 4$ works well for many applications with sample sizes $n \le 1{,}200$. A similar phenomenon has also been observed in the batch-sequential design for parametric models (Ford, Titterington, & Kitsos, 1989); also see the discussion in Section 4.5.

To understand this phenomenon, Theorem 3 studies the ideal case where $f_j = f^*$, $j \ge 2$, are the exact optimal design density so that $\bar{f} = f_U/\ell_n + (1 - 1/\ell_n)f^* \equiv \tilde{f}$.

Theorem 3. *Assume that n IID design points are drawn from the mixture density $\tilde{f} = \lambda_n f_U + (1 - \lambda_n) f^*$, where $\lambda_n = 1/\ell_n$. As $\lambda_n \to 0$,*

$$\lim_{\lambda_n \to 0} \frac{\mathrm{REL}(\tilde{f})}{\lambda_n^2} = \frac{4}{5}\left[\int_0^1 \sqrt{W(x)}\mathrm{d}x \int_0^1 \frac{1}{\sqrt{W(x)}}\mathrm{d}x - 1\right]. \tag{8}$$

By Theorem 3, $\mathrm{REL}(\tilde{f})$ converges to zero at rate $O(\ell_n^2)$. Consider $W(x) = 0.01 + x^r$, $x \in [0, 1]$. In Table 1 below we tabulate $\mathrm{REL}(f_U)$ using Equation (6) and $\mathrm{REL}(\tilde{f})$ using Equation (8) with $\ell_n = 4$. We see that $\ell_n = 4$ provides reasonably good performance and a full sequential procedure (Efromovich, 2008) with a higher cost may be unnecessary.

TABLE 1: Relative efficiency loss in percentages for f_U and $\tilde{f}$ with $\ell_n = 4$.

r	1	2	3	4	5	6	7	8	9	10
REL($\tilde{f}$)	1.1	2.8	3.6	4.0	4.0	3.9	3.8	3.6	3.5	3.3
REL(f_U)	8.4	18.0	24.9	29.8	33.3	35.8	37.7	39.2	40.3	41.2

3. APPLICATIONS IN NONPARAMETRIC INFERENCE

In this section denote by $K(\cdot)$ a symmetric kernel function with bounded support. For a bandwidth b_n, write $K_{b_n}(u) = K(u/b_n)$. For ease of presentation, we introduce the notation

$$C_K = \int u^2 K(u)\mathrm{d}u, \qquad D_K = \int K^2(u)\mathrm{d}u.$$

Throughout the rest of this section we write $K_b(u) = K(u/b)$.

3.1. Inference for Mean Regression Function

For a given input $X = x$, we are interested in the mean regression function $\mu(x) = \mathbb{E}(Y|X = x)$. Given samples (X_i, Y_i) from (X, Y), where X has density function $f(x)$, consider the local linear estimate $\hat{\mu}(x)$ of $\mu(x)$:

$$\left[\hat{\mu}(x), \hat{\beta}(x)\right] = \underset{(\alpha,\beta)}{\operatorname{argmin}} \sum_{i=1}^{n} \{Y_i - \alpha - \beta(X_i - x)\}^2 K_{b_n}(x - X_i).$$

By Section 2.4 in Li & Racine (2007), Equation (2) holds with $\rho(x) = C_K^2 \mu''(x)^2/4$ and $W(x) = D_K \sigma^2(x)$, where $\sigma^2(x) = \mathrm{var}(Y|X = x)$ is the conditional variance function. By Theorem 1, the optimal design density is $f^*(x) = \sigma(x)/\int_0^1 \sigma(x)\mathrm{d}x$. To estimate $\sigma^2(x)$, we can apply the local linear estimate to the squared residuals:

$$\left[\hat{\sigma}^2(x), \hat{\gamma}(x)\right] = \underset{(\alpha,\gamma)}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\{Y_i - \hat{\mu}(X_i)\}^2 - \alpha - \gamma(X_i - x)\right]^2 K_{h_n}(x - X_i), \tag{9}$$

where h_n is another bandwidth. For the consistency and other properties of $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$, we refer the reader to Fan & Yao (1998) and Li & Racine (2007) for more details.

We can build different models by specifying different structures for $e = Y - \mu(X)$. For example, if $e = \sigma(X)\varepsilon$ for an unknown function $\sigma(x) \geq 0$, $x \in [0, 1]$, and independent error ε with $E(\varepsilon^2) = 1$, then we have the conditional heteroscedastic model:

$$Y = \mu(X) + \sigma(X)\varepsilon. \tag{10}$$

3.2. Inference for Conditional Variance Function

Suppose that researchers are interested in the conditional variance function $\sigma^2(x)$ in Equation (10). Then we can use Equation (9) to estimate $\sigma^2(x)$. By Fan & Yao (1998), $\sigma^2(x)$ can be estimated as well as if $\mu(x)$ were known and Equation (2) holds with $\rho(x) = C_K^2 \sigma''(x)^2/4$ and $W(x) = D_K \sigma^4(x)\mathrm{var}(\varepsilon^2)$. By Theorem 1, the optimal design density is $f^*(x) = \sigma^2(x)/\int_0^1 \sigma^2(x)\mathrm{d}x$, which

can be estimated by replacing $\sigma^2(x)$ with $\hat{\sigma}^2(x)$. We see that, for model (10), the optimal design density for $\mu(x)$ is proportional to $\sigma(x)$, whereas the optimal design density for $\sigma^2(x)$ is proportional to $\sigma^2(x)$.

3.3. Inference for Conditional Distribution Function

Denote by $F(y|x) = \mathbb{P}(Y \leq y|X = x)$ the conditional distribution function of Y given $X = x$. Clearly, $F(y|x)$ offers more information than the conditional mean regression function. To estimate $F(y|x)$, consider the local linear estimator

$$\left[\hat{F}(y|x), \hat{\beta}(x, y)\right] = \underset{(\alpha,\beta)}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\mathbf{1}_{Y_i \leq y} - \alpha - \beta(X_i - x)\right]^2 K_{b_n}(x - X_i). \tag{11}$$

By Section 6.1 in Li & Racine (2007), Equation (2) holds with

$$\rho(x, y) = \frac{1}{4} C_K^2 F''(y|x)^2 \quad \text{and} \quad W(x, y) = D_K[F(y|x) - F^2(y|x)].$$

For double-integrated mean squared error, as in Equation (3), we have the leading term

$$\int\int \text{MISE}\{\hat{F}(y|x)\}\mathrm{d}x\mathrm{d}y \asymp b_n^4 \int\int \rho(x, y)\mathrm{d}x\mathrm{d}y + \frac{1}{nb_n} \int\int \frac{W(x, y)}{f(x)}\mathrm{d}x\mathrm{d}y.$$

By Theorem 1, the optimal design density is

$$f^*(x) = c^{-1}\left[\int W(x, y)\mathrm{d}y\right]^{1/2}, \quad \text{where} \quad c = \int \left[\int W(x, y)\mathrm{d}y\right]^{1/2} \mathrm{d}x.$$

We can estimate $f^*(x)$ by replacing $W(x, y)$ with $\hat{W}(x, y) = \hat{F}(y|x) - \hat{F}^2(y|x)$.

3.4. Inference for Quantile Regression

Quantile regression has become an active area of research over the past three decades; see Koenker (2005) for an extensive exposition. Unlike the ordinary regression that studies the conditional mean function of Y given X as in Section 3.1, quantile regression studies the conditional quantile function of Y given X and hence is robust against outliers. As a measure of how response Y depends on covariate X, the conditional quantile can offer a full picture of the local structure of Y and X by specifying different quantiles.

For $\tau \in (0, 1)$, denote by $\mu_\tau(x)$ the conditional τ-th quantile of Y given $X = x$. By Yu & Jones (1998), we can estimate $\mu_\tau(x)$ by the local linear quantile regression

$$\left[\hat{\mu}_\tau(x), \hat{\beta}(x)\right] = \underset{(\mu,\beta)}{\operatorname{argmin}} \sum_{i=1}^{n} L_\tau\{Y_i - \mu - \beta(x - X_i)\}K_{b_n}(x - X_i), \tag{12}$$

where $L_\tau(t) = |t| + (2\tau - 1)t$ is the check function. Equation (2) holds with

$$\rho(x) = \frac{1}{4} C_K^2 \sigma''(x)^2 \quad \text{and} \quad W(x) = \frac{\tau(1 - \tau)D_K}{f_{Y|X}^2\{\mu_\tau(x)|x\}},$$

where $f_{Y|X}(y|x)$ is the conditional density function of Y given $X = x$. Therefore, by Theorem 1, the optimal design density is

$$f^*(x) = \frac{1}{cf_{Y|X}\{\mu_\tau(x)|x\}}, \quad \text{where} \quad c = \int_0^1 \frac{1}{f_{Y|X}\{\mu_\tau(x)|x\}}\mathrm{d}x. \tag{13}$$

To estimate $f^*(x)$, we can estimate the conditional density $f_{Y|X}(y|x)$ by the nonparametric conditional density estimator

$$\hat{f}_{Y|X}(y|x) = \frac{(nb_xb_y)^{-1}\sum_{i=1}^n K_{b_x}(x - X_i)K_{b_y}(y - Y_i)}{(nb_x)^{-1}\sum_{i=1}^n K_{b_x}(x - X_i)}, \tag{14}$$

where b_x and b_y are two bandwidths. We then estimate $f^*(x)$ by plugging $\hat{\mu}_\tau(\cdot)$ and $\hat{f}_{Y|X}(\cdot|\cdot)$ into Equation (13). See Section 4.4 for a discussion on the selection of b_x and b_y.

3.5. Inference for Varying Coefficient Models

Since the introduction by Cleveland, Grosse, & Shyu (1991), varying coefficient models have received considerable attention in many scientific areas, including economics, epidemiology, medical science and ecology among others; see Fan & Zhang (2008) for a survey. Let $X \in [0, 1]$ be random design point with density $f(x)$, $\mathbf{Z} \in \mathbb{R}^p$ a column random vector independent of X, and $\boldsymbol{\alpha}(\cdot)$ a p dimensional vector of functions on $[0, 1]$. Given independent samples $(\mathbf{Z}_i, X_i, Y_i)$ from

$$Y = \mathbf{Z}^T\boldsymbol{\alpha}(X) + \sigma(X)\varepsilon,$$

we are interested in the p dimensional vector $\alpha(\cdot)$ of functional coefficients. For a given $x \in [0, 1]$, consider the local linear estimator $\hat{\boldsymbol{\alpha}}(x)$ of $\boldsymbol{\alpha}(x)$:

$$\left[\hat{\boldsymbol{\alpha}}(x), \hat{\boldsymbol{\beta}}(x)\right] = \underset{(\boldsymbol{\alpha},\boldsymbol{\beta})}{\operatorname{argmin}} \sum_{i=1}^n \left\{Y_i - Z_i^T\boldsymbol{\alpha} - \mathbf{Z}_i^T\boldsymbol{\beta}(X_i - x)\right\}^2 K_{b_n}(x - X_i).$$

By Theorem 1 in Fan & Zhang (2008), the asymptotic bias of $\hat{\alpha}(x)$ does not depend on the design density and the asymptotic covariance matrix is

$$\operatorname{cov}\{\hat{\alpha}(x)\} \asymp \frac{D_K\{\mathbb{E}(\mathbf{Z}\mathbf{Z}^T)\}^{-1}}{nb_n}\frac{\sigma^2(x)}{f(x)}.$$

Thus, by Theorem 1, the optimal design density is $\sigma(x)/\int_0^1 \sigma(x)\mathrm{d}x$. To estimate $\sigma(x)$, one can apply local linear regression to $\{Y_i - \mathbf{Z}_i^T\hat{\boldsymbol{\alpha}}(X_i)\}^2$; see Fan & Zhang (2008).

3.6. Inference for Semiparametric Varying Coefficient Partially Linear Models

The varying coefficient partially linear model (Fan & Huang, 2005) is a very useful semiparametric model that models the key covariates linearly and models the rest of the covariates nonparametrically. The model assumes the form

$$Y = \mathbf{Z}^T\boldsymbol{\alpha}(X) + \mathbf{U}^T\boldsymbol{\beta} + \sigma(X)\varepsilon, \tag{15}$$

where Y is the response variable, $\mathbf{Z} \in \mathbb{R}^p$ and $\mathbf{U} \in \mathbb{R}^q$ are random covariate vectors, $X \in [0, 1]$ is the design point and $\boldsymbol{\alpha}(\cdot)$ is a p dimensional vector of nonparametric functions on $[0, 1]$. Note that

the varying coefficient model in Section 3.5 and the partially linear model (Green & Silverman, 1994) are two special cases of (15). Let $\hat{\boldsymbol{\beta}}$ be a $\sqrt{n}$-consistent estimate of $\boldsymbol{\beta}$. Then $\boldsymbol{\alpha}(\cdot)$ can be estimated as well as if $\boldsymbol{\beta}$ were known using the varying coefficient models in Section 3.5, and the optimal design density is $f^*(x) = \sigma(x)/\int_0^1 \sigma(x)\mathrm{d}x$. Here, $\sigma^2(x)$ can be estimated by applying local linear regression to the squared residuals $[Y_i - \mathbf{Z}_i^T\hat{\boldsymbol{\alpha}}(X_i) - \mathbf{U}_i^T\hat{\boldsymbol{\beta}}]^2$. See Fan & Huang (2005) for methods to obtain $\hat{\boldsymbol{\beta}}$.

4. SIMULATION STUDY

In this section we examine the performance of the proposed sequential design procedure through Monte Carlo studies. For an estimate $\hat{m}(x)$ of $m(x)$, its MISE is computed as the average of 1,000 realizations of $\int_0^1[\hat{m}(x) - m(x)]^2\mathrm{d}x$. Recall Definition 1. We denote by $\mathrm{REL}(\hat{f}^*)$ and $\mathrm{REL}(f_U)$ the relative efficiency loss of the sequential design and the uniform design in comparison to the optimal design f^* in Equation (5). Here, $\mathrm{REL}(\hat{f}^*)$ can be interpreted as the percentage of efficiency loss by replacing the unknown optimal design density Equation (5) with its sequential estimator in Equation (7).

4.1. Conditional Mean Function

Consider $\mu(\cdot)$ from the model

$$Y = \mu(X) + \sigma(X)\varepsilon, \quad \sigma(x) = \sqrt{0.01 + x^r}, \tag{16}$$

where ε is the standard normal error and $X \in [0, 1]$ is the design point. We use $\mu(x) = \sin(2\pi x)$. To see the effect of $\sigma(\cdot)$, we consider five choices of $r = 1, 2, \ldots, 5$. The local linear fit is implemented using the command `locpoly` in `R` package `KernSmooth`, with the optimal plug-in bandwidth selection using command `dpill`. To estimate $\sigma^2(\cdot)$, we use (9) with the plug-in bandwidth based on $(X_i, [Y_i - \hat{\mu}(X_i)]^2)$, $i = 1, \ldots, n$. The result is summarized in Table 2 for sample size $n = 600$ and 1,200, with the last two columns being the REL for the uniform design. We see that the empirical results agree with the theoretical derivation in Table 1. Clearly, the uniform design suffers from substantial efficiency loss for almost all choices of r except $r = 1$, and this loss widens as r increases. In contrast, the sequential design has a performance comparable to that of the optimal design. In Equation (16), we have also tried other choices of $\sigma(\cdot)$, including

$$\sigma(x) = 0.5 + 0.4\cos(2\pi x), \quad \mathbf{1}_{x\le 0.5} + c_1\mathbf{1}_{x>0.5}, \quad \phi\{(x - 0.5)/c_2\},$$

for various values of $c_1, c_2 > 0$, where ϕ is the standard normal density. Our conclusion is that, as $\sigma(\cdot)$ becomes further from the constant function, the performance of the sequential design improves in comparison to the uniform design.

In Table 2, we note that in some cases the sequential design performs even better than the optimal design. As explained in Section 2.3, this is because the optimal design contains few observations at regions where $\sigma(\cdot)$ is very small, leading to poor estimation. The latter phenomenon becomes even more remarkable if we drop the small number 0.01 and use $\sigma(x) = \sqrt{x^r}$ in Equation (16). For example, if $r = 3$, the estimate from the optimal design occasionally exhibits a volatile pattern near zero, producing unusually large values of the MISE. Figure 1 presents boxplots for 1,000 realizations of MISEs from the optimal, sequential and uniform designs, respectively. For better presentation, the seven largest MISEs are dropped from the optimal design. There are relatively more outliers in the boxplots for the optimal and uniform designs than the sequential design. For the optimal design, these outliers are caused by sparse observations near zero. For the uniform design, due to a high noise level at $x \approx 1$, the uniform sampling does not sample

TABLE 2: REL($\hat{f}^*$) and REL(f_U) for $\mu(x)$ in Equation (16): units are percentages.

	REL($\hat{f}^*$) ($n = 600$)				REL($\hat{f}^*$) ($n = 1{,}200$)					REL(f_U) (n=)	
$r\backslash\ell_n$	6	4	3	2	8	6	4	3	2	600	1,200
1	0.4	−0.4	0.3	2.5	1.5	−0.5	0.0	−0.8	3.7	11.5	10.6
2	1.6	2.4	4.0	7.7	−5.3	−4.2	−0.4	−0.4	4.2	23.0	19.9
3	2.1	1.3	1.6	9.1	1.4	−0.5	−0.5	1.9	8.3	28.1	27.9
4	0.9	0.6	3.1	8.8	1.8	0.0	1.8	4.6	6.7	33.1	31.1
5	0.4	2.2	3.7	6.7	−0.7	1.4	0.0	2.8	11.4	34.9	34.6

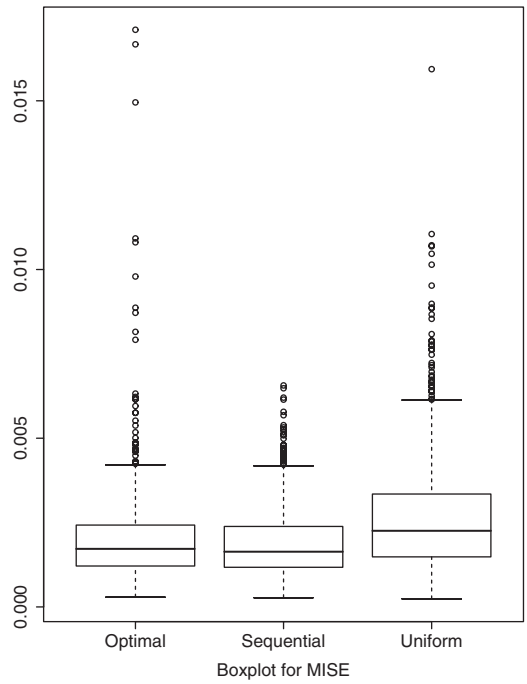

FIGURE 1: Boxplots for MISEs for estimates of $\mu(\cdot)$ based on $n = 1{,}200$ observations from $Y = \mu(X) + \sqrt{X^3}\varepsilon$ with true $\mu(x) = \sin(2\pi x)$ using optimal design (left), sequential design (middle) with $\ell_n = 4$ batches and uniform design (right). Boxplots for sequential and uniform designs are based on 1,000 realizations of MISEs, whereas, for better visualization, boxplot for optimal design is based on all 1,000 minus the 7 largest realizations.

enough observations from that region. In contrast, the sequential design performs much better by adapting to the unknown model dynamics.

4.2. Conditional Variance Function

Consider $\sigma^2(\cdot)$ from the model

$$Y = \sigma(X)\varepsilon, \quad \sigma(x) = \sqrt{0.01 + x^r}, \tag{17}$$

TABLE 3: REL($\hat{f}^*$) and REL(f_U) for $\sigma^2(x)$ in Equation (17): units are percentages.

	REL($\hat{f}^*$) ($n = 600$)				REL($\hat{f}^*$) ($n = 1{,}200$)					REL(f_U) (n=)	
$r\backslash\ell_n$	6	4	3	2	8	6	4	3	2	600	1,200
0.5	−6.9	−5.6	−3.8	−8.9	2.6	0.0	0.8	−0.6	0.2	3.0	8.3
1.0	−0.7	7.9	4.4	6.3	−5.4	−3.5	−3.1	−5.0	6.9	23.7	26.8
1.5	−5.3	−3.8	−0.2	7.6	−2.8	0.5	2.7	2.2	11.3	30.6	32.3
2.0	−7.2	−8.9	−2.1	4.9	−7.0	−4.6	6.1	5.4	9.4	36.0	36.2
2.5	3.1	1.4	8.3	20.7	0.7	0.0	2.7	11.5	20.1	45.5	51.7

where we use the same setting for X and ε as in Equation (16). Again, we apply local linear regression to (X, Y^2) and report the results in Table 3 for sample size $n = 600, 1{,}200$ and $r = 0.5, 1.0, \ldots, 2.5$. Again, the sequential design performs much better than the uniform design and is comparable to the optimal design.

4.3. Conditional Distribution Function

Consider the conditional distribution function $F_{Y|X}(y|x)$ for model (16). Let Φ be the standard normal distribution function. Then $F_{Y|X}(y|x) = \Phi\{[y - \mu(x)]/\sigma(x)\}$. We use Equation (11) to estimate $F_{Y|X}(y|x)$. The MISE is computed as a double-integral over $x \in [0, 1]$, $y \in [q_{0.1}, q_{0.9}]$, and $\mu(x) - 2\sigma(x) \le y \le \mu(x) + 2\sigma(x)$, where $q_{0.1}$ and $q_{0.9}$ are the 10th and 90th percentiles of Y. For computational reasons, the latter integral is approximated using $51 \times 31 = 1{,}581$ uniformly spaced grid points on the two coordinates. For each fixed grid point y, we can use the command `dpill` to choose the optimal bandwidth, denoted by $b_n(y)$. The final optimal bandwidth is taken as the average of those $b_n(y)$'s. Another bandwidth selection method is the more computationally expensive cross-validation method; see Li & Racine (2007). The result of the simulation study is summarized in Table 4. We see that the sequential design is comparable to the optimal design and much better than the uniform design.

4.4. Conditional Quantile Function

Consider the conditional τ-th quantile of Y given $X = x$, denoted by $\mu_\tau(x)$, for model (16). To choose the bandwidth b_n in Equation (12), we follow Yu & Jones (1998) and use $b_n = \{\tau(1 - \tau)/[\phi(\Phi^{-1}(\tau))]^2\}^{1/5} b_n^{\mathrm{LS}}$, where b_n^{LS} is the optimal local linear least-squares plug-in bandwidth

TABLE 4: REL($\hat{f}^*$) and REL(f_U) for $F_{Y|X}$ in Equation (16): units are percentages.

	REL($\hat{f}^*$) ($n = 600$)				REL($\hat{f}^*$) ($n = 1{,}200$)					REL(f_U) (n=)	
$r\backslash\ell_n$	6	4	3	2	8	6	4	3	2	600	1,200
1	−0.4	1.4	0.8	0.9	−2.1	−1.8	−1.1	−1.8	−1.1	14.8	12.5
2	−0.2	−0.2	0.8	1.1	−1.9	−1.1	−1.5	−0.7	0.4	18.0	16.5
3	1.2	2.1	1.6	3.6	0.8	0.4	1.2	1.2	2.7	22.6	20.6
4	3.2	2.0	1.0	2.9	3.4	0.4	3.0	2.6	5.4	27.4	25.9
5	2.3	2.8	3.5	4.7	2.2	1.3	1.7	2.6	4.2	28.7	26.7

TABLE 5: REL($\hat{f}^*$) and REL(f_U) for $\mu_{0.5}(x)$ in Equation (16): units are percentages.

	REL($\hat{f}^*$) ($n = 600$)				REL($\hat{f}^*$) ($n = 1{,}200$)					REL(f_U) (n=)	
$r\backslash\ell_n$	6	4	3	2	8	6	4	3	2	600	1,200
1	1.9	−6.1	2.7	4.7	3.0	0.5	3.9	4.3	4.3	12.7	14.3
2	17.2	14.1	14.7	14.3	5.6	1.8	9.0	7.4	7.4	33.2	25.5
3	13.1	5.3	12.7	16.3	4.2	5.9	10.7	9.5	9.5	33.2	35.3
4	12.4	15.4	17.4	21.6	9.2	2.4	11.9	10.4	10.4	39.3	36.6
5	11.7	12.8	8.6	21.7	13.5	12.0	15.5	10.2	10.2	44.3	42.4

using the command `dpill`. To estimate the optimal design density f^* in Equation (13), we adopt the likelihood cross-validation method (Li & Racine, 2007) to choose b_x and b_y in Equation (14) as follows:

$$\underset{b_x,b_y}{\operatorname{argmax}} \sum_{i=1}^{n} \log\left[\hat{f}^{(-i)}(Y_i|X_i)\right],$$

where $\hat{f}^{(-i)}$ is the estimator using all but the point (X_i, Y_i). We use only 300 realizations because it is computationally expensive to implement the quantile regression estimation and the cross-validation bandwidth selection. We summarize the result for $\tau = 0.5$ (the conditional median) in Table 5. The sequential design significantly outperforms the uniform design for all cases considered, but the performance in Table 5 is not as impressive as those in Tables 2–4. The latter phenomenon can be attributed to the difficulty in estimating the two-dimensional conditional density function in Equation (14). With the larger sample size $n = 1{,}200$, the conditional density estimator is more accurate, leading to better performance.

4.5. Discussions on Sample Size and Block Length

The proposed sequential design method also shows encouraging performance for smaller sample sizes. For example, for $n = 200$, the RELs for the uniform design for $\mu(x)$ in Equation (16) are 1.95%, 18.8%, 15.4%, 17.4%, 31.3% for $r = 1, 2, 3, 4, 5$, respectively, compared to 0.50%, 8.45%, −0.88%, 1.57%, 11.4% for the sequential design with $\ell_n = 2$. With a smaller sample size, the sequential design performs less impressively compared to the cases with larger sample sizes. A similar phenomenon has also been observed in conditional variance, conditional distribution and conditional quantile cases.

For the number of blocks ℓ_n, the asymptotic theory shows that the optimal ℓ_n is proportional to $n^{2/5}$. Our empirical studies also indicate that a larger ℓ_n is preferred as the sample size increases. Roughly speaking, ℓ_n would double when the sample size increases from n to $6n$. For $n = 200$, it seems that $\ell_n = 2$ works the best; for $\ell_n = 4, 5, 8$, due to the small block size, the local linear estimation procedure is not stable and occasionally produces extreme outputs. Therefore, for smaller sample sizes $n \le 200$, we recommend using two blocks; for sample size λn with $\lambda > 1$, we can take $\ell_n \approx 2\lambda^{2/5}$.

ACKNOWLEDGEMENTS

We are grateful to the associate editor and two anonymous referees for their comments that have significantly improved this paper. We also thank Amanda Applegate for help on improving the

presentation. Zhao's research was partially supported by a National Institute on Drug Abuse grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse or the National Institutes of Health.

BIBLIOGRAPHY

Abdelbasit, K. M. & Plackett, R. L. (1983). Experimental design for binary data. *Journal of the American Statistical Association*, 78, 90–98.

Biedermann, S. & Dette, H. (2001). Minimax optimal designs for nonparametric regression: A further optimality property of the uniform distribution. In *MODA 6: Advances in Model-Oriented Design and Analysis*, Atkinson, A. C., Hackl, P. & Müller, W. G., editors. Physica-Verlag, Heidelberg, pp. 13–20.

Chaudhuri, P. & Mykland, P. A. (1993). Nonlinear experiments: Optimal design and inference based on likelihood. *Journal of the American Statistical Association*, 88, 538–546.

Cheng, M. Y., Hall, P., & Titterington, D. M. (1998). Optimal design for curve estimation by local linear smoothing. *Bernoulli*, 4, 3–14.

Cleveland, W. S., Grosse, E., & Shyu, W. M. (1991). Local regression models. In *Statistical Models in S*, Chambers, J. M. & Hastie, T. J., editors. Wadsworth & Brooks, Pacific Grove, pp. 309–376.

Dette, H., Melas, V. B., & Pepelyshev, A. (2004). Optimal designs for a class of nonlinear regression models. *Annals of Statistics*, 32, 2142–2167.

Draper, N. R. & Hunter, W. G. (1967). The use of prior distributions in the design of experiments for parameter estimation in nonlinear situations. *Biometrika*, 54, 147–153.

Efromovich, S. (2008). Optimal sequential design in a controlled non-parametric regression. *Scandinavian Journal of Statistics*, 35, 266–285.

Fan, J. & Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11, 1031–1057.

Fan, J. & Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85, 645–660.

Fan, J. & Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface*, 1, 179–195.

Faraway, J. J. (1990). Sequential design for the nonparametric regression of curves and surfaces. In *Computer Science and Statistics: Proceedings of the 22nd Annual Symposium on the Interface*, Michigan State University, MI, 1990, pp. 104–110.

Ford, I. & Silvey, S. D. (1980). A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika*, 67, 381–388.

Ford, I., Titterington, D. M., & Kitsos, C. P. (1989). Recent advances in nonlinear experimental design. *Technometrics*, 31, 49–60.

Ford, I., Titterington, D. M., & Wu, C. F. J. (1985). Inference and sequential design. *Biometrika*, 72, 545–551.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London.

Hohmann, G. & Jung, W. (1975). On sequential and nonsequential D-optimal experiment design. *Biometrische Zeitshrift*, 17, 329–336.

Kiefer, J. (1959). Optimum experimental designs (with discussion). *Journal of the Royal Statistical Society, Series B*, 21, 272–319.

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press, New York.

Li, Q. & Racine, J. (2007). *Nonparametric Econometrics*, Princeton University Press, Princeton, New Jersey.

Müller, H. G. (1984). Optimal designs for nonparametric kernel regression. *Statistics & Probability Letters*, 2, 285–290.

Müller, W. G. (1996). Optimal design for local fitting. *Journal of Statistical Planning and Inference*, 55, 389–397.

Park, D. & Faraway, J. J. (1998). Sequential design for response curve estimation. *Journal of Nonparametric Statistics*, 9, 155–164.

Wu, C. F. J. (1985). Asymptotic inference from sequential design in a nonlinear situation. *Biometrika*, 72, 553–558.

Yu, K. & Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, 93, 228–237.

APPENDIX

Proof of Theorem 1. In Equation (4), it is easy to see that the optimal bandwidth is $b_n^* = n^{-1/5}\{\int_0^1 W(x)/f(x)\mathrm{d}x/[4\int_0^1 \rho(x)\mathrm{d}x]\}^{1/5}$. Using the latter optimal bandwidth, we can get

$$\mathrm{MISE}^*\{\hat{m}|f\} = (4^{-4/5} + 4^{1/5})n^{-4/5}\left\{\int_0^1 \rho(x)\mathrm{d}x\right\}^{1/5}\left\{\int_0^1 \frac{W(x)}{f(x)}\mathrm{d}x\right\}^{4/5}. \tag{18}$$

So, it suffices to find $f(x)$ to minimize $\int_0^1 W(x)/f(x)\mathrm{d}x$. Recall the Cauchy-Schwarz inequality $\int g^2 \int h^2 \geq (\int gh)^2$ for any square-integrable functions g and h. Thus,

$$\int_0^1 \frac{W(x)}{f(x)}\mathrm{d}x = \int_0^1 \frac{W(x)}{f(x)}\mathrm{d}x \int_0^1 f(x)\mathrm{d}x \geq \left\{\int_0^1 \sqrt{W(x)}\mathrm{d}x\right\}^2.$$

Here, under the constraint $\int_0^1 f(x)\mathrm{d}x = 1$, the equality holds if and only if, for some ω,

$$\frac{W(x)}{f(x)} = \omega f(x) \quad \text{or equivalently} \quad f(x) = \frac{\sqrt{W(x)}}{\int_0^1 \sqrt{W(x)}\mathrm{d}x},$$

and $\omega = \{\int_0^1 W^{1/2}(x)\mathrm{d}x\}^2$, completing the proof. ■

Proof of Proposition 1. The assertion Equation (6) easily follows by replacing f in Equation (18) with the optimal density f^* and the uniform density f_U. ■

Proof of Theorems 2–3. We prove only Theorem 3 because Theorem 2 follows similarly. Write $f^*(x) = \sqrt{W(x)}/c$, where $c = \int_0^1 \sqrt{W(x)}\mathrm{d}x$. Define

$$\rho_n = \frac{U_n - c^2}{U_n} \qquad \text{and} \qquad U_n = \int_0^1 \frac{W(x)}{\lambda_n + (1-\lambda_n)f^*(x)}\mathrm{d}x.$$

Then

$$\frac{U_n - c^2}{c} = \int_0^1 \frac{W(x)}{c\lambda_n + (1-\lambda_n)\sqrt{W(x)}}\mathrm{d}x - \int_0^1 \sqrt{W(x)}\mathrm{d}x = \frac{\tau_n}{c}\int_0^1 \frac{\sqrt{W(x)} - c}{\tau_n/\sqrt{W(x)} + 1}\mathrm{d}x,$$

where $\tau_n = c\lambda_n/(1-\lambda_n) \to 0$. By the Taylor expansion $(1+z)^{-1} = 1 - z + O(z^2)$ as $z \to 0$,

$$\int_0^1 \frac{\sqrt{W(x)} - c}{\tau_n/\sqrt{W(x)} + 1} \mathrm{d}x = \int_0^1 \left\{\sqrt{W(x)} - c\right\}\left\{1 - \tau_n/\sqrt{W(x)} + O(\tau_n^2)\right\}$$

$$= \tau_n \left[c \int_0^1 \frac{1}{\sqrt{W(x)}} \mathrm{d}x - 1 \right] + O(\tau_n^2).$$

As $\lambda_n \to 0$, we have $U_n \to c^2$, $\tau_n \to 0$, $\rho_n \to 0$ and $\tau_n^2/\lambda_n^2 \to c^2$. Therefore,

$$\frac{\rho_n}{\lambda_n^2} = \frac{U_n - c^2}{c\tau_n^2} \frac{c}{U_n} \frac{\tau_n^2}{\lambda_n^2} \to c \int_0^1 \frac{1}{\sqrt{W(x)}} \mathrm{d}x - 1.$$

Note the expansion $(1-z)^b = 1 - bz + O(z^2)$ for $z \to 0$ and $b \in \mathbb{R}$. By Equation (18),

$$\mathrm{REL}(\tilde{f}) = 1 - \frac{\mathrm{MISE}^*\{\hat{m}\,|\,f^*\}}{\mathrm{MISE}^*\{\hat{m}\,|\,\tilde{f}\}} = 1 - (1-\rho_n)^{4/5} = \frac{4}{5}\rho_n + O(\rho_n^2),$$

which produces the desired result. ■

Received 24 January 2011
Accepted 17 November 2011