ASYMPTOTICS FOR ROBUST SEQUENTIAL DESIGNS IN MISSPECIFIED REGRESSION MODELS

Sanjoy Sinha
University of Winnipeg

Douglas P. Wiens
University of Alberta

We revisit a proposal, for robust sequential design in the presence of uncertainty about the regression response, previously made by these authors. We obtain conditions under which a sequence of designs for nonlinear regression models leads to asymptotically normally distributed estimates. The results are illustrated in a simulation study. We conclude that estimates computed after the experiment has been carried out sequentially may, in only moderately sized samples, be safely used to make standard normal-theory inferences, ignoring the dependencies arising from the sequential nature of the sampling. The quality of the normal approximation deteriorates somewhat when the random errors are heteroscedastic.

1. Introduction

In a recent article (Sinha and Wiens 2002, henceforth referred to as SW), we applied notions of robustness of design in the presence of response uncertainty in a nonlinear regression setting. We developed and implemented an algorithm for the sequential selection of design points $\mathbf{x}$, from a specified "design space" $\mathcal{S} \subset \mathbb{R}^q$, at which to observe a random variable Y. This random variable was assumed to follow a regression model with a nonlinear and possibly misspecified response function. The sequential sampling scheme used in SW—and described below—is somewhat involved. Asymptotic normality of the resulting estimates was posited, and tested in a simulation study. In this article we shall fill in some of the theoretical gaps left by SW, and then revisit the aforementioned nonlinear regression application.

We begin by describing in detail the application which motivates the current study. We entertain a sequence of nonlinear regression problems indexed by the sample size n. At the nth stage it is supposed that one samples from a distribution P^n of r.v.s Y_n, with means depending on $\mathbf{x}$ through an unknown parameter vector $\boldsymbol{\theta}_n \in \mathbb{R}^p$ and a nonlinear function of $\mathbf{x}$ and $\boldsymbol{\theta}_n$ ranging over a neighbourhood of a tentative choice f:

$$E[Y_n \mid \mathbf{x}] \approx f(\mathbf{x}; \boldsymbol{\theta}_n). \tag{1.1}$$

For instance the experimenter may fit a Michaelis-Menten response $f(x; \boldsymbol{\theta}) = \theta_0 x/(\theta_1 + x)$ when in fact the true response is exponential: $E[Y \mid \mathbf{x}] = \theta_0(1 - e^{-\theta_1 x})$. (We shall return to this particular example in §3.1 below.)

AMS subject classifications: Primary 62L05, 60F05; Secondary 62J02..

Keywords and phrases: asymptotic normality; biased regression; mixing; nonlinear regression; sequential design..

Given the uncertainty about the response, one naturally asks what meaning one can give to the parameter $\boldsymbol{\theta}_n$. We define this quantity to be the parameter value making (1.1) most accurate in the L_2-sense, *viz.*,

$$\boldsymbol{\theta}_n = \arg\min_{\boldsymbol{\theta}} \int_{\mathcal{S}} \{E[Y_n \mid \mathbf{x}] - f(\mathbf{x};\boldsymbol{\theta})\}^2 d\mathbf{x}. \tag{1.2}$$

The approximation (1.1) is assumed to be sufficiently accurate that the integral in (1.2) exists for all $\boldsymbol{\theta}$ in some open set. To formalize our (shrinking) neighbourhood structure, we define $d_n(\mathbf{x};\boldsymbol{\theta}_n) = E[Y_n \mid \mathbf{x}] - f(\mathbf{x};\boldsymbol{\theta}_n)$. Then the regression response is modelled as

$$E[Y_n \mid \mathbf{x}] = f(\mathbf{x};\boldsymbol{\theta}_n) + d_n(\mathbf{x};\boldsymbol{\theta}_n),$$

and we assume that $d_n(\cdot;\boldsymbol{\theta}_n)$ is a "small" function in a sense made precise in Assumption B5) of §3.

Given data $\{\mathbf{z}_{n,i} = (Y_{n,i}, \mathbf{x}_i); i = 1, \ldots, n\}$, we define

$$g(\mathbf{z};\boldsymbol{\theta}) = \big(Y - f(\mathbf{x};\boldsymbol{\theta})\big)^2,$$

$$H_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_{n,i};\boldsymbol{\theta}).$$

Here and elsewhere we adopt the convention that, if multiple observations are made at a location $\mathbf{x}_i$, then summands involving $\mathbf{x}_i$ are to be repeated an appropriate number of times.

We suppose that the estimate $\widehat{\boldsymbol{\theta}}_n$ is obtained by least squares:

$$\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} H_n(\boldsymbol{\theta}),$$

and aim to show that $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n)$ is asymptotically normally distributed; for this we shall compare $\boldsymbol{\theta}_n$ to

$$\overline{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} \overline{H}_n(\boldsymbol{\theta}),$$

where

$$\overline{H}_n(\boldsymbol{\theta}) = E[H_n(\boldsymbol{\theta})]. \tag{1.3}$$

In Section 2 of this paper we apply work of Domowitz and White (1982) and others to derive conditions under which $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \overline{\boldsymbol{\theta}}_n)$ is asymptotically normally distributed. In Section 3 we discuss these conditions, in the context of the sequential design problem outlined above. This includes the evaluation of the asymptotic mean $\overline{\boldsymbol{\theta}}_n$ as "$\boldsymbol{\theta}_n$ + bias," with the bias expressed explicitly. A simulation study is carried out in §3.1. On the basis of this we conclude that estimates computed after the experiment has been carried out sequentially may, in only moderately sized samples, be safely used to make standard normal-theory inferences, ignoring the dependencies arising from the sequential nature of the sampling. The quality of the normal approximation deteriorates somewhat when the random errors are heteroscedastic.

1.1. Sampling scheme

In a nonlinear experiment, the usual measures of performance of a design depend on the parameters being estimated. A sequential approach is then naturally suggested - design points should be chosen so as to minimize some loss function evaluated at the estimates obtained from observations made at previous design points. Sequential designs in nonlinear experiments have been studied by a number of authors—see Ford, Titterington and Kitsos (1989) for a review. In particular, Chaudhuri and Mykland (1993) considered the problem of choosing optimal designs sequentially. In their approach a static initial design was to be augmented by a fully adaptive sequential design, using parameter estimates based on available data. Assuming that the fitted response was in fact a member of the chosen parametric family, they showed that a sequence of maximum likelihood estimates of the regression parameters was consistent and asymptotically normally distributed, and that the sequence of designs was asymptotically D-optimal, in the sense of maximizing the determinant of the true information matrix.

SW extended the work of Chaudhuri and Mykland (1993) to the case—outlined above—in which the fitted response is possibly of an incorrect form and the variances are possibly heteroscedastic. We proposed a sequential sampling mechanism which minimizes a loss function based on the integrated mean squared error of the estimates of $E[Y \mid \mathbf{x}]$. In a small sample simulation study, we showed that the resulting designs were very successful, relative to some common competitors, in reducing mean squared error due to model misspecification and to heteroscedastic variation.

With $\dot{\mathbf{f}}(\mathbf{x};\boldsymbol{\theta}) = \partial f(\mathbf{x};\boldsymbol{\theta})/\partial\boldsymbol{\theta} : p \times 1$, we anticipated in SW the asymptotic normality result

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n \sim \mathcal{AN}\big(\overline{\mathbf{M}}_n^{-1}\overline{\mathbf{b}}_n, \overline{\mathbf{M}}_n^{-1}\overline{\mathbf{Q}}_n\overline{\mathbf{M}}_n^{-1}\big) \tag{1.4}$$

where $\overline{\mathbf{M}}_n = E[\mathbf{M}_n]$ and $\overline{\mathbf{b}}_n = E[\mathbf{b}_n]$ for

$$\mathbf{M}_n = \sum_{i=1}^{n} \dot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta}_n)\dot{\mathbf{f}}^T(\mathbf{x}_i;\boldsymbol{\theta}_n),$$
$$\mathbf{b}_n = \sum_{i=1}^{n} \dot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta}_n)d_n(\mathbf{x}_i;\boldsymbol{\theta}_n),$$

and

$$\overline{\mathbf{Q}}_n = \mathrm{COV}\bigg[\sum_{i=1}^{n}\big(Y_{n,i} - f(\mathbf{x}_i;\boldsymbol{\theta}_n)\big)\dot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta}_n)\bigg].$$

In our simulations to assess empirically the approach to normality, the expected sums $\overline{\mathbf{M}}_n$ and $\overline{\mathbf{b}}_n$ were estimated by the sample sums $\mathbf{M}_n$ and $\mathbf{b}_n$

respectively, each evaluated at $\widehat{\boldsymbol{\theta}}_n$. The difficulties in estimating matrices of the form $\overline{\mathbf{Q}}_n$ were underscored by White (1984), who proposed an estimator which is consistent under mixing conditions, if the response function is correctly specified. As pointed out by White (1984) and exploited by Wu (1985) however, if the fitted response is correct and as well the measurement errors form a martingale difference sequence then $n^{-1}(\mathbf{Q}_n - \overline{\mathbf{Q}}_n) \xrightarrow{L_1} \mathbf{0}$ for

$$\mathbf{Q}_n = \sum_{i=1}^{n} \dot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta}_n)\sigma^2(\mathbf{x}_i)\dot{\mathbf{f}}^T(\mathbf{x}_i;\boldsymbol{\theta}_n),$$

where $\sigma^2(\mathbf{x})$ is the variance of Y when observed at $\mathbf{x}$. The convergence holds as well for static designs, i.e., predetermined design points. For these reasons SW replaced $\overline{\mathbf{Q}}_n$ by $\mathbf{Q}_n$ (evaluated at $\widehat{\boldsymbol{\theta}}_n$) for the numerical work.

The asymptotic mean squared error matrix corresponding to (1.4) is

$$\mathrm{MSE}_n = \overline{\mathbf{M}}_n^{-1}\big(\overline{\mathbf{Q}}_n + \overline{\mathbf{b}}_n\overline{\mathbf{b}}_n^T\big)\overline{\mathbf{M}}_n^{-1}.$$

A primary purpose of nonlinear regression is typically response estimation or prediction, and so a robust choice of design should focus on the minimization, in some sense, of the average error when $E[Y \mid \mathbf{x}]$ is estimated by $f(\mathbf{x};\widehat{\boldsymbol{\theta}}_n)$. Integrating the first order approximation of this error over the design space yields the asymptotic integrated mean squared error

$$\begin{aligned}\int_{\mathcal{S}} E[(f(\mathbf{x};\widehat{\boldsymbol{\theta}}_n) - E[Y \mid \mathbf{x}])^2]\,d\mathbf{x} \\ &\approx \int_{\mathcal{S}} E[\{(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n)^T\dot{\mathbf{f}}(\mathbf{x};\boldsymbol{\theta}_n) - d_n(\mathbf{x};\boldsymbol{\theta}_n)\}^2]\,d\mathbf{x} \\ &= \mathrm{tr}[\mathrm{MSE}_n \cdot \mathbf{A}_n] + \int_{\mathcal{S}} d_n^2(\mathbf{x};\boldsymbol{\theta}_n)\,d\mathbf{x},\end{aligned}$$

where $\mathbf{A}_n = \int_{\mathcal{S}} \dot{\mathbf{f}}(\mathbf{x};\boldsymbol{\theta}_n)\dot{\mathbf{f}}^T(\mathbf{x};\boldsymbol{\theta}_n)\,d\mathbf{x}$. Minimization of

$$\mathcal{L}_I = \mathrm{tr}[\mathrm{MSE}_n \cdot \mathbf{A}_n]$$

is thus a natural analogue of the classical I-optimality criterion. (We drop the integral $\int_{\mathcal{S}} d_n^2(\mathbf{x};\boldsymbol{\theta}_n)\,d\mathbf{x}$ since it is not affected by the choice of design.)

The sampling mechanism used in SW is as follows:

Step 1 Start with a static design, in which r_0 observations are independently made at each of n_0 locations $\mathbf{x}_1, \ldots, \mathbf{x}_{n_0}$. Put $n = n_0 r_0$ and define $n_1 = 0$.

Step 2 Consider augmenting the existing design by a further r_1 observations at an arbitrary point $\mathbf{x} \in \mathcal{S}$. Let $l(\mathbf{x} \mid \boldsymbol{\theta}_n)$ be the resulting

(estimated) value of $\mathcal{L}_I$, viewed as a function of $\mathbf{x}$ alone. Choose the next location to be

$$\mathbf{x}_{n_0+n_1+1} = \arg\min l(\mathbf{x} \mid \widehat{\boldsymbol{\theta}}_n).$$

Increment n_1 to n_1+1 and n to $n+r_1$ after making r_1 observations at $\mathbf{x}_{n_0+n_1+1}$.

Step 3 Repeat Step 2 until .... Here "..." refers to any practical stopping rule—perhaps defined by the attainment of a desired precision, but more likely by the availability of resources.

A complicating factor is that the determination of $\mathbf{b}_n$ and $\mathbf{Q}_n$ entails as well the computation of estimates of $d_n(\mathbf{x}; \widehat{\boldsymbol{\theta}}_n)$ and $\sigma^2(\mathbf{x})$. These estimates are obtained through a process of smoothing the residuals, as follows. The discrepancies $d_n(\mathbf{x}_i; \boldsymbol{\theta}_n)$ are first estimated by the medians of the residuals $e_{ij} = y_{ij} - f(\mathbf{x}_i; \widehat{\boldsymbol{\theta}})$, $j = 1, \ldots, r_1$. Here $\widehat{\boldsymbol{\theta}}$ denotes the current value of the estimate. These medians are then smoothed to yield final estimates $\widehat{\mathbf{d}}_{n-1} = (\hat{d}(\mathbf{x}_1), \ldots, \hat{d}(\mathbf{x}_{n-1}))^T$ and predictions $\hat{d}(\mathbf{x}_n)$. Then $\sigma^2(\mathbf{x}_i)$ is estimated by the squared median absolute deviation (*mad*) of the e_{ij}, with the *mad* normalized by division by $\Phi^{-1}(.75) = .6745$ for consistency at the normal distribution. Finally, these variance estimates are smoothed to yield predictions $\hat{\sigma}^2(\mathbf{x}_n)$. Smoothing is carried out by fitting a cubic spline in the case of scalar x's. For vector $\mathbf{x}$'s we instead fit a generalized additive model on S-Plus; this in turns employs both cubic spline fitting and loess smoothing.

2. General theory

In this section, we study the asymptotic properties of estimators based on observations which are obtained from a sequential design scheme sharing features of that described in the previous section or, more generally, are dependent in a suitably weak sense. In particular, the response function may be misspecified and the errors may be heteroscedastic.

We first summarize some relevant results from Domowitz and White (1982). To establish the consistency and asymptotic normality of the estimator $\widehat{\boldsymbol{\theta}}_n$, the following assumptions are imposed.

A1 The sequence $\{Y_{n,i}\}$ of responses is generated as

$$Y_{n,i} = E[Y_n \mid \mathbf{x}_i] + \epsilon_{n,i}, i = 1, \ldots, n,$$

where $E[Y_n \mid \mathbf{x}_i]$ are unknown mean functions of the random vector $\mathbf{x}_i$. The vector $\mathbf{z}_{n,i} = (Y_{n,i}, \mathbf{x}_i)$ is finite-dimensional and jointly distributed with distribution function P_i on Ω, a Euclidean space. The elements $\epsilon_{n,i}$ are unobservable random errors.

The experimenter models the mean function, perhaps erroneously, by a function $f(\mathbf{x};\boldsymbol{\theta})$. This approximating function is assumed to satisfy

A2 $f(\mathbf{x};\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ for each $\mathbf{x}$ in S and a measurable function of $\mathbf{x}$ for each $\boldsymbol{\theta}$ in Θ, a compact subset of a finite-dimensional Euclidean space.

Before stating the next assumptions we review the notion of "mixing". Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be two σ-algebras on a probability space $(\Omega, \mathcal{B}, P)$ and define

$$\phi(\mathcal{B}_1,\mathcal{B}_2) = \sup\{|P(B_2 \mid B_1) - P(B_2)| \mid B_1 \in \mathcal{B}_1, B_2 \in \mathcal{B}_2, P(B_1) > 0\},$$
$$\alpha(\mathcal{B}_1,\mathcal{B}_2) = \sup\{|P(B_1B_2) - P(B_1)P(B_2)| \mid B_1 \in \mathcal{B}_1, B_2 \in \mathcal{B}_2\}.$$

Intuitively, the coefficients ϕ and α measure the dependence of the events in $\mathcal{B}_2$ on those in $\mathcal{B}_1$ in terms of how much the probability of the joint occurrence of an event in each σ-algebra differs from the product of the probabilities of each event occurring. The events in $\mathcal{B}_1$ and $\mathcal{B}_2$ are independent if and only if ϕ and α are zero. The function α provides an absolute measure of dependence, while ϕ measures dependence relative to $P(B_1)$.

Definition 2.1 (Mixing). For a sequence of random vectors $\{Y_i\}$ defined on the probability space $(\Omega, \mathcal{B}, P)$, let $\mathcal{B}_a^b$ be the Borel σ-algebra of events generated by $\{Y_a, Y_{a+1}, \ldots, Y_b\}$. Define the mixing coefficients

$$\phi(m) = \sup_n \phi(\mathcal{B}_{-\infty}^n, \mathcal{B}_{n+m}^\infty) \quad \text{and} \quad \alpha(m) = \sup_n \alpha(\mathcal{B}_{-\infty}^n, \mathcal{B}_{n+m}^\infty).$$

A sequence for which $\phi(m) \to 0$ as $m \to \infty$ is termed *uniform* or ϕ-mixing and a sequence for which $\alpha(m) \to 0$ as $m \to \infty$ is termed *strong* or α-mixing.

The coefficients $\phi(m)$ and $\alpha(m)$ measure the dependence between events separated by at least m time periods. Thus if $\phi(m) = 0$ or $\alpha(m) = 0$ for some m, events m periods apart are independent. By allowing $\phi(m)$ or $\alpha(m)$ to approach zero as $m \to \infty$, we allow considerations of situations where events are asymptotically independent. Note that as $\phi(m) \geq \alpha(m)$, ϕ-mixing implies α-mixing. For a real number $r \geq 1$, if

(i) $\phi(m) = O(m^{-\tau})$ for $\tau > r/(2r-1)$,

we say that $\phi(m)$ is of size $r/(2r-1)$ and if

(ii) $\alpha(m) = O(m^{-\tau})$ for $\tau > r/(r-1)$, $r > 1$,

we say that $\alpha(m)$ is of size $r/(r-1)$. This definition gives a precise idea about the memory of a random sequence that can be related to moment conditions expressed in terms of r. As $r \to \infty$ a sequence exhibits more dependence; as $r \to 1$ it exhibits less dependence.

We also require

Definition 2.2 (Uniform Integrability). A family $\{Y_i : i \in I\}$ of integrable random variables is said to be *uniformly integrable* if

$$\limsup\left\{\int_{|Y_i|>K} |Y_i|\, dP : i \in I\right\} = 0 \quad \text{as } K \to \infty.$$

A sufficient condition for $\{Y_i : i \in I\}$ to be uniformly integrable is that $E|Y_i|^{1+\delta} \leq \Delta < \infty$ for some positive constants Δ and δ (see Hoadley 1971). Moreover, if $E|Y_i|^{r+\delta} \leq \Delta < \infty$ for some $r \geq 1$ and $0 < \delta \leq r$, then the family $\{Y_i : i \in I\}$ is said to be *uniformly* $(r+\delta)$*-integrable* (see Domowitz and White, 1982). If a sequence of measurable functions $\{q_n\}$ satisfies $|q_n(\mathbf{z}_{n,i}, \boldsymbol{\theta})| \leq K_n(\mathbf{z}_{n,i})$ for all $\boldsymbol{\theta} \in \Theta$, with $\{K_n\}$ measurable and uniformly $(r+\delta)$-integrable, we say that $\{q_n\}$ is *dominated by uniformly* $(r+\delta)$*-integrable functions.*

Recall now the definition of $\overline{\boldsymbol{\theta}}_n$. To show that $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\overline{\boldsymbol{\theta}}_n$, the following assumptions are required.

A3 The sequence $\{\mathbf{z}_{n,i}\}$ is either ϕ-mixing, with $\phi(m)$ of size $r_1/(2r_1-1)$, $r_1 \geq 1$, or α-mixing, with $\alpha(m)$ of size $r_1/(r_1-1)$, $r_1 > 1$.

A4 $\{g(\mathbf{z}_{n,i}; \boldsymbol{\theta})\}$ is dominated by uniformly $(r_1+\delta)$-integrable functions, $r_1 \geq 1$, $0 < \delta \leq r_1$.

A5 The function $\overline{H}_n(\boldsymbol{\theta})$ defined in (1.3) has an identifiably unique minimizer (in the sense of Definition 2.1 of Domowitz and White, 1982) $\overline{\boldsymbol{\theta}}_n \in \Theta$ for all sufficiently large n.

Assumption A3 restricts the memory of the process $\{\mathbf{z}_{n,i}\}$ in a fashion analogous to the role of ergodicity for a stationary stochastic process. Assumption A4 restricts the moments of the approximation error, and A5 gives an identification condition. Theorem 2.1 below addresses the consistency of the estimator $\widehat{\boldsymbol{\theta}}_n$.

Theorem 2.1 (Domowitz and White, 1982). *Under Assumptions* A1–A5, $\widehat{\boldsymbol{\theta}}_n - \overline{\boldsymbol{\theta}}_n \to \mathbf{0}$ *a.s. as* $n \to \infty$.

The asymptotic normality of $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \overline{\boldsymbol{\theta}}_n)$ is established by applying the Mean Value Theorem to the first order conditions for a minimum of $H_n(\boldsymbol{\theta})$. To establish asymptotic normality, the following assumptions are made.

A6 The functions $g(\mathbf{z}_{n,i}; \boldsymbol{\theta})$ are twice continuously differentiable in $\boldsymbol{\theta}$, uniformly in n and i, a.s.-P.

A7 For $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, $\{(\partial g(\mathbf{z}_{n,i}; \boldsymbol{\theta})/\partial\theta_j)^2\}$, $j = 1, \ldots, p$, are dominated by uniformly r_2-integrable functions, $r_2 > 1$.

A8 Define $\overline{\mathbf{V}}_{a,n} = \mathrm{COV}[n^{-1/2}\sum_{i=a+1}^{a+n} g'(\mathbf{z}_{n,i};\overline{\boldsymbol{\theta}}_n)]$. Assume that there exists a positive definite matrix $\mathbf{V}$ such that $\boldsymbol{\lambda}^T\overline{\mathbf{V}}_{a,n}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\mathbf{V}\boldsymbol{\lambda} \to 0$ as $n \to \infty$, uniformly in a, for any non-zero vector $\boldsymbol{\lambda}$.

A9 For $\boldsymbol{\theta} = (\theta_1 \ldots, \theta_p)^T$, $\{\partial^2 g(\mathbf{z}_{n,i};\boldsymbol{\theta})/\partial\theta_j\partial\theta_k\}$, $j,k = 1,\ldots,p$, are dominated by uniformly $(r_1+\delta)$-integrable functions, $0 < \delta \le r_1$.

A10 The matrix $(\overline{H}_n''(\boldsymbol{\theta}) =)\ \mathbf{A}_n(\theta) = n^{-1}\sum_{i=1}^n E[g''(\mathbf{z}_{n,i};\boldsymbol{\theta})]$ has constant rank p in some open ε-neighbourhood of $\overline{\boldsymbol{\theta}}_n$, for all sufficiently large n, uniformly in n.

Assumption A9 is used to ensure the convergence of the sample Hessian and, together with A7, allows for the calculation of the gradient and Hessian of $\overline{H}_n(\boldsymbol{\theta})$ by interchanging differentiation with expectation. Assumption A10 is used to guarantee that $\overline{\mathbf{A}}_n := \mathbf{A}_n(\overline{\boldsymbol{\theta}}_n)$ is positive definite for sufficiently large n. These assumptions together also ensure the non-singularity of $\widehat{\mathbf{A}}_n := H_n''(\widehat{\boldsymbol{\theta}}_n)$.

The following strengthening of A3 is required.

A3′ Assumption A3 holds, and either $\phi(m)$ is of size $r_2/(r_2-1)$, or $\alpha(m)$ is of size $\max[r_1/(r_1-1), r_2/(r_2-1)]$ $(r_1, r_2 > 1)$.

Theorem 2.2 (Domowitz and White, 1982). *Under Assumptions* A1, A2, A3′–A10, *if* $\overline{\boldsymbol{\theta}}_n$ *is interior to* Θ *we have*

$$\overline{\mathbf{V}}_n^{-1/2}\overline{\mathbf{A}}_n\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \overline{\boldsymbol{\theta}}_n) \xrightarrow{\mathcal{L}} N(\mathbf{0},\mathbf{I})$$

where $\overline{\mathbf{V}}_n = \overline{\mathbf{V}}_{0,n}$.

3. Application to sequential design

To apply the asymptotic theory of the preceding section to the design problem of §1 we make the following assumptions. They are in some cases considerably stronger than the assumptions of §2, but plausible and realistic in a design context. They allow for a straightforward application of Theorems 2.1 and 2.2.

B1 Conditionally (given $\mathbf{x}$) the errors $\varepsilon_n = Y_n - E[Y_n \mid \mathbf{x}]$ have mean 0 and variance $\sigma^2(\mathbf{x})$, with $\sup_{\mathbf{x}\in\mathcal{S}}\sigma^2(\mathbf{x}) < \infty$. For some $\delta > 0$ the sequence $\{E[|\varepsilon_n|^{2+\delta}]\}_{n=1}^{\infty}$ is bounded.

B2 The function $f(\mathbf{x};\boldsymbol{\theta})$ is a twice continuously differentiable function of $\boldsymbol{\theta}$ for each $\mathbf{x} \in \mathcal{S}$, and a measurable function of $\mathbf{x}$ for each $\boldsymbol{\theta} \in \Theta$. This function as well as the gradient $\dot{\mathbf{f}}(\mathbf{x};\boldsymbol{\theta})$ and the Hessian $\ddot{\mathbf{f}}(\mathbf{x};\theta)$ are bounded for $\mathbf{x} \in \mathcal{S}$ and $\boldsymbol{\theta} \in \Theta$.

Assumptions B1 and B2 imply A1, A2, A4, A6, A7 and A9.

B3 The minimum eigenvalues of $n^{-1}\overline{\mathbf{M}}_n$ are bounded away from 0.

Assumptions A5 and A10 follow from B3 and the fact, shown below, that

$$\frac{1}{n}\overline{\mathbf{M}}_n - \frac{1}{2}\overline{\mathbf{A}}_n \to \mathbf{0}, \tag{3.1}$$

so that $\overline{H}_n(\boldsymbol{\theta})$ is locally convex.

Define

$$\overline{\mathbf{Q}}_{a,n} = \mathrm{COV}\Big[\sum_{i=a+1}^{a+n} \big(Y_{n,i} - f(\mathbf{x}_i;\boldsymbol{\theta}_n)\big)\dot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta}_n)\Big]$$

and note that $\overline{\mathbf{Q}}_{0,n} = \overline{\mathbf{Q}}_n$. Assume that

B4 There exists a positive definite matrix $\overline{\mathbf{Q}}$ such that $n^{-1}\overline{\mathbf{Q}}_{a,n} \to \overline{\mathbf{Q}}$, uniformly in a.

We will show that

$$\frac{1}{4}\overline{\mathbf{V}}_{a,n} - \frac{1}{n}\overline{\mathbf{Q}}_{a,n} \to \mathbf{0}, \tag{3.2}$$

uniformly in a. This together with B4 implies A8.

As exploited in a similar context by Jaeckel (1971), in order that standard error and bias be of the same order of magnitude asymptotically the true and fitted models should approach each other at a rate $n^{-1/2}$. We assume

B5 $\tau_n = \sup_{\mathbf{x}\in\mathcal{S}} |d_n(\mathbf{x};\boldsymbol{\theta}_n)|$ is $O(n^{-1/2})$ as $n \to \infty$.

We can now verify (1.4).

Theorem 3.1. *Under Assumptions* B1–B5 *and* A3′, *if* $\overline{\boldsymbol{\theta}}_n$ *is interior to* Θ *we have:*

(i) $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n \to \mathbf{0}$ *a.s. as* $n \to \infty$,

(ii) $\overline{\mathbf{Q}}_n^{-1/2}\overline{\mathbf{M}}_n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n - \overline{\mathbf{M}}_n^{-1}\overline{\mathbf{b}}_n) \xrightarrow{\mathcal{L}} N(\mathbf{0},\mathbf{I})$.

Proof. By the discussion preceding the statement of the theorem, the assumptions of Theorems 2.1 and 2.2 hold.

To show (i) we first calculate

$$\begin{aligned}\frac{1}{2}\overline{H}'_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} E\big[\big(f(\mathbf{x}_i;\boldsymbol{\theta}) - f(\mathbf{x}_i;\boldsymbol{\theta}_n)\big)\dot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta})\big] \\ &\qquad - \frac{1}{n}\sum_{i=1}^{n} E[d_n(\mathbf{x}_i;\boldsymbol{\theta}_n)\dot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta})], \\ \frac{1}{2}\overline{H}''_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} E[\dot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta})\dot{\mathbf{f}}^T(\mathbf{x}_i;\boldsymbol{\theta})] \\ &\qquad + \frac{1}{n}\sum_{i=1}^{n} E\big[\big(f(\mathbf{x}_i;\boldsymbol{\theta}) - f(\mathbf{x}_i;\boldsymbol{\theta}_n)\big)\ddot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta})\big] \\ &\qquad - \frac{1}{n}\sum_{i=1}^{n} E[d_n(\mathbf{x}_i;\boldsymbol{\theta}_n)\ddot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta})].\end{aligned} \tag{3.3}$$

In particular,

$$\frac{1}{2}\overline{H}'_n(\boldsymbol{\theta}_n) = -\frac{1}{n}\overline{\mathbf{b}}_n.$$

By the Mean Value Theorem, there is a mean value $\widetilde{\boldsymbol{\theta}}_n$ for which

$$\mathbf{0} = \tfrac{1}{2}\overline{H}'_n(\overline{\boldsymbol{\theta}}_n) = \tfrac{1}{2}\overline{H}'_n(\boldsymbol{\theta}_n) + [\tfrac{1}{2}\overline{H}''_n(\widetilde{\boldsymbol{\theta}}_n)](\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n),$$

whence

$$\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n = \left[\frac{1}{2}\overline{H}''_n(\widetilde{\boldsymbol{\theta}}_n)\right]^{-1}\left[\frac{1}{n}\overline{\mathbf{b}}_n\right]. \tag{3.4}$$

By B2 and B5, $[-\frac{1}{2}\overline{H}''_n(\widetilde{\boldsymbol{\theta}}_n)]$ is $O(1)$ and $\frac{1}{n}\overline{\mathbf{b}}_n \to \mathbf{0}$, so that

$$\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n \to \mathbf{0} \tag{3.5}$$

and (i) follows from Theorem 2.1.

To show (ii) we must establish (3.1), (3.2) and that

$$\sqrt{n}(\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n - \overline{\mathbf{M}}_n^{-1}\overline{\mathbf{b}}_n) \to \mathbf{0}. \tag{3.6}$$

For (3.1), we calculate

$$\begin{aligned}\frac{1}{n}\overline{\mathbf{M}}_n - \frac{1}{2}\overline{\mathbf{A}}_n &= \frac{1}{n}\sum_{i=1}^{n} E\big[\big(f(\mathbf{x}_i;\boldsymbol{\theta}_n) - f(\mathbf{x}_i;\overline{\boldsymbol{\theta}}_n)\big)\ddot{\mathbf{f}}(\mathbf{x}_i;\overline{\boldsymbol{\theta}}_n)\big] \\ &\qquad + \frac{1}{n}\sum_{i=1}^{n} E[\dot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta}_n)\dot{\mathbf{f}}^T(\mathbf{x}_i;\boldsymbol{\theta}_n) - \dot{\mathbf{f}}(\mathbf{x}_i;\overline{\boldsymbol{\theta}}_n)\dot{\mathbf{f}}^T(\mathbf{x}_i;\overline{\boldsymbol{\theta}}_n)] \\ &\qquad + \frac{1}{n}\sum_{i=1}^{n} E[d_n(\mathbf{x}_i;\boldsymbol{\theta}_n)\ddot{\mathbf{f}}(\mathbf{x}_i;\overline{\boldsymbol{\theta}}_n)].\end{aligned} \tag{3.7}$$

For some mean value $\widetilde{\boldsymbol{\theta}}_n$ the first sum is

$$\left\{\frac{1}{n}\sum_{i=1}^{n} E[\ddot{\mathbf{f}}(\mathbf{x}_i;\overline{\boldsymbol{\theta}}_n)\otimes\dot{\mathbf{f}}^T(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}_n)]\right\}\big(\mathbf{I}_p\otimes(\overline{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n)\big),$$

and so converges to $\mathbf{0}$ by (3.5) and B2. A further application of the Mean Value Theorem shows that the second sum also converges to $\mathbf{0}$. The third sum is bounded in norm by

$$\frac{\tau_n}{n}\sum_{i=1}^{n} E[\|\ddot{\mathbf{f}}(\mathbf{x}_i;\overline{\boldsymbol{\theta}}_n)\|] = \tau_n\cdot O(1),$$

and thus converges to $\mathbf{0}$.

To establish (3.6), we use (3.4) to write

$$\sqrt{n}(\overline{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n-\overline{\mathbf{M}}_n^{-1}\overline{\mathbf{b}}_n) = \left(\left[\frac{1}{2}\overline{H}_n''(\widetilde{\boldsymbol{\theta}}_n)\right]^{-1}-\left[\frac{1}{n}\overline{\mathbf{M}}_n\right]^{-1}\right)\frac{1}{\sqrt{n}}\overline{\mathbf{b}}_n.$$

Since $n^{-1/2}\overline{\mathbf{b}}_n$ is $O(1)$ by B5, it suffices to establish that $\frac{1}{2}\overline{H}_n''($ $widetilde\boldsymbol{\theta}_n) = n^{-1}\overline{\mathbf{M}}_n + o(1)$. But

$$\begin{aligned}\frac{1}{2}\overline{H}_n''(\widetilde{\boldsymbol{\theta}}_n)-\frac{1}{n}\overline{\mathbf{M}}_n &= \frac{1}{n}\sum_{i=1}^{n} E\big[\big(f(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}_n)-f(\mathbf{x}_i;\boldsymbol{\theta}_n)\big)\ddot{\mathbf{f}}(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}_n)\big]\\ &\quad+\frac{1}{n}\sum_{i=1}^{n} E[\dot{\mathbf{f}}(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}_n)\dot{\mathbf{f}}^T(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}_n)-\dot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta}_n)\dot{\mathbf{f}}^T(\mathbf{x}_i;\boldsymbol{\theta}_n)]\\ &\quad-\frac{1}{n}\sum_{i=1}^{n} E[d_n(\mathbf{x}_i;\boldsymbol{\theta}_n)\ddot{\mathbf{f}}(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}_n)],\end{aligned}$$

so that the verification of this last point is entirely analogous to that of (3.1).

To verify (3.2) we will show that

$$\alpha_n := \boldsymbol{\lambda}^T\left[\frac{1}{4}\overline{\mathbf{V}}_{a,n}-\frac{1}{n}\overline{\mathbf{Q}}_{a,n}\right]\boldsymbol{\lambda}\to 0, \tag{3.8}$$

uniformly in a, for any vector $\boldsymbol{\lambda}$. For this, write $\mathbf{q}_{n,i}(\boldsymbol{\theta}) = \big(Y_{n,i}-f(\mathbf{x}_i;\boldsymbol{\theta})\big)\times\dot{\mathbf{f}}(\mathbf{x}_i;\boldsymbol{\theta})$, $S_{n,i}(\boldsymbol{\theta}) = \boldsymbol{\lambda}^T\mathbf{q}_{n,i}(\boldsymbol{\theta})$ and again use the Mean Value Theorem to obtain

$$\begin{aligned}\alpha_n &= \frac{1}{n}\left\{\mathrm{VAR}\left[\sum_{i=a+1}^{a+n} S_{n,i}(\overline{\boldsymbol{\theta}}_n)\right]-\mathrm{VAR}\left[\sum_{i=a+1}^{a+n} S_{n,i}(\boldsymbol{\theta}_n)\right]\right\} \qquad (3.9)\\ &= \frac{1}{n}\,\mathrm{VAR}\left[(\overline{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n)^T\sum_{i=a+1}^{a+n}\dot{S}_{n,i}(\widetilde{\boldsymbol{\theta}})\right]\\ &\qquad+\frac{2}{n}\,\mathrm{COV}\left[\sum_{i=a+1}^{a+n} S_{n,i}(\boldsymbol{\theta}),(\overline{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n)^T\sum_{i=a+1}^{a+n}\dot{S}_{n,i}(\widetilde{\boldsymbol{\theta}}_n)\right]\end{aligned}$$

for mean values $\widetilde{\boldsymbol{\theta}}_n$. Since

$$\frac{1}{n}\,\mathrm{VAR}\Big[\sum_{i=a+1}^{a+n} S_{n,i}(\boldsymbol{\theta})\Big] = \boldsymbol{\lambda}^T\Big[\frac{1}{n}\overline{\mathbf{Q}}_{a,n}\Big]\boldsymbol{\lambda}$$

is $O(1)$ by B4, (3.8) will follow if the first term in (3.9) tends to 0. To handle this first term note that, by (3.6), $\sqrt{n}(\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n)$ is $O(1)$, so that it suffices to establish that

$$\frac{1}{n^2}\,\mathrm{COV}\Big[\sum_{i=a+1}^{a+n} \dot{S}_{n,i}(\widetilde{\boldsymbol{\theta}}_n)\Big] \to \mathbf{0}.$$

Equivalently, for any vector $\boldsymbol{\beta}$ and with $T_{n,i} := \boldsymbol{\beta}^T\dot{S}_{n,i}(\widetilde{\boldsymbol{\theta}}_n)$,

$$\boldsymbol{\beta}^T\Big\{\frac{1}{n^2}\,\mathrm{COV}\Big[\sum_{i=a+1}^{a+n} \dot{S}_{n,i}(\widetilde{\boldsymbol{\theta}}_n)\Big]\Big\}\boldsymbol{\beta} = \frac{1}{n^2}\,\mathrm{VAR}\Big[\sum_{i=1}^{n} T_{n,i+a}\Big] \to 0,$$

uniformly in a. This in turn follows from

$$\frac{1}{n^2}\sum_{i,j=1}^{n} |\,\mathrm{CORR}[T_{n,i+a}, T_{n,j+a}]| \to 0,$$

which is a consequence of the mixing conditions together with Corollary 6.16 of White (1984). □

Assumption B1 is very mild, as is B2. The latter is easily seen to hold for the types of response functions (exponential, Michaelis-Menten) considered in SW. Assumptions B3 and B4 are difficult to verify theoretically, but empirical evidence for them can be garnered from the behaviour of the sample values. Assumption B5 of course restricts the range of alternatives against which robustness can be expected.

The mixing conditions A3′ are the most contentious issue. As White (1984) points out, such conditions are impossible to verify empirically. What can be assessed empirically however is the extent to which (1.4) holds in simulation studies, with $\overline{\mathbf{M}}_n$, $\overline{\mathbf{Q}}_n$ and $\overline{\mathbf{b}}_n$ replaced by sample estimates.

3.1. Simulation Study

We have carried out an empirical investigation of the adequacy of the normal approximation result of Theorem 3.1. Define $\widehat{\mathbf{M}}_n$ and $\widehat{\mathbf{Q}}_n$ to be the estimates of $\overline{\mathbf{M}}_n$ and $\overline{\mathbf{Q}}_n$ computed as described in §1.1 and define

$$\begin{aligned}\widehat{\mathbf{C}}_n &= \widehat{\mathbf{M}}_n^{-1}\widehat{\mathbf{Q}}_n\widehat{\mathbf{M}}_n^{-1},\\ \mu_n(\mathbf{x};\boldsymbol{\theta}_0) &= \mathbf{z}^T(\mathbf{x};\boldsymbol{\theta}_0)\overline{\mathbf{M}}_n^{-1}\overline{\mathbf{b}}_n,\\ s_n^2(\mathbf{x}) &= \mathbf{z}^T(\mathbf{x};\widehat{\boldsymbol{\theta}}_n)\widehat{\mathbf{C}}_n\mathbf{z}(\mathbf{x};\widehat{\boldsymbol{\theta}}_n).\end{aligned}$$

Under the conditions of Theorem 3.1,

$$t^* = \frac{f(\mathbf{x}; \widehat{\boldsymbol{\theta}}_n) - f(\mathbf{x}; \boldsymbol{\theta}_0) - \mu_n(\mathbf{x}; \boldsymbol{\theta}_0)}{s_n(\mathbf{x})}$$

is approximately distributed as Student's t_{n-p} for $n = n_0 r_0 + n_1 r_1$.

Figures 1–4 give Q-Q plots of the t^* statistic for two 2-parameter cases. The design space is $\mathcal{S} = [.5, 5]$; we investigate both $x = 3$ (interpolation) and $x = 6$ (extrapolation). In Figure 1 we consider, as a benchmark, the case in which the fitted straight line response is exactly correct and the errors are homoscedastic. In Figures 2–4 we fit a Michaelis-Menten response function, with the true response and error structures being (Michaelis-Menten, homoscedastic), (exponential, homoscedastic), (exponential, heteroscedastic) respectively. When the errors are heteroscedastic the variance function is $\sigma^2(x) = 1 + (x - .5)^2$. When the true response is Michaelis-Menten the parameters are $(\theta_0, \theta_1) = (50, .5)$. In the exponential case the true parameters are $(\theta_0, \theta_1) = (10, .5)$; in this case the parameters of the closest fitting

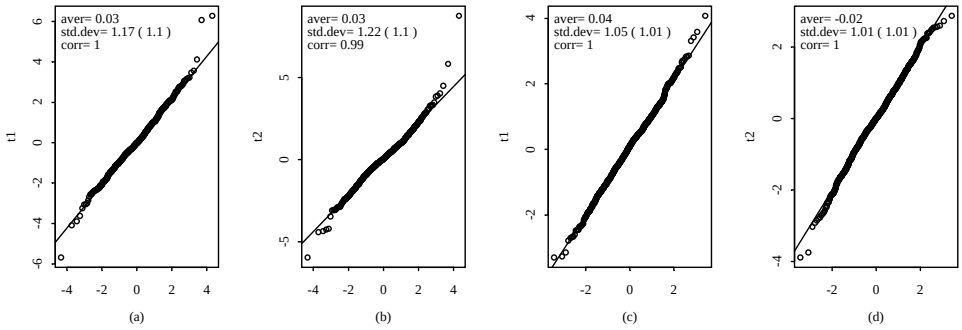

Figure 1. Q-Q plots for t-approximations to the Studentized distribution of $f(x; \widehat{\boldsymbol{\theta}}_N)$. Fitted response is linear, true response is linear, errors are homoscedastic. Values of (n_0, x) are (a) $(0, 3)$, (b) $(0, 6)$, (c) $(15, 3)$, (d) $(15, 6)$.

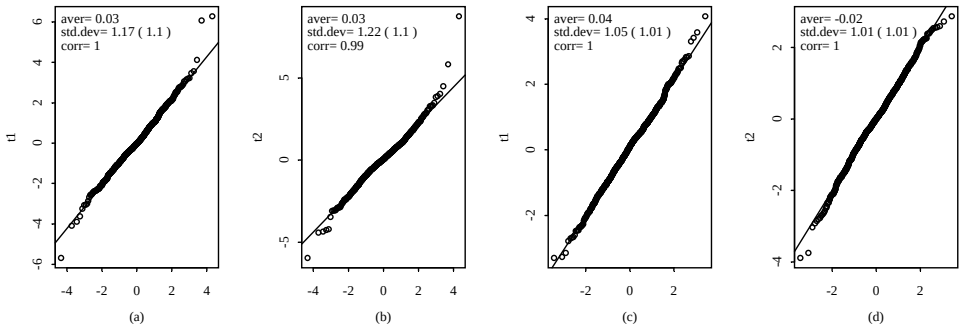

Figure 2. Q-Q plots for t-approximations to the Studentized distribution of $f(x; \widehat{\boldsymbol{\theta}}_N)$. Fitted response is Michaelis-Menten, true response is Michaelis-Menten, errors are homoscedastic. Values of (n_0, x) are (a) $(0, 3)$, (b) $(0, 6)$, (c) $(15, 3)$, (d) $(15, 6)$.

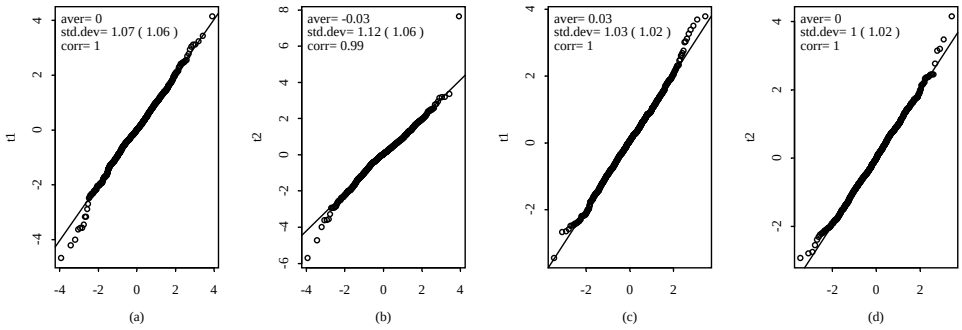

Figure 3. Q-Q plots for t-approximations to the Studentized distribution of $f(x;\widehat{\boldsymbol{\theta}}_N)$. Fitted response is Michaelis-Menten, true response is exponential, errors are homoscedastic. Values of (n_0, x) are (a) $(0, 3)$, (b) $(0, 6)$, (c) $(15, 3)$, (d) $(15, 6)$.

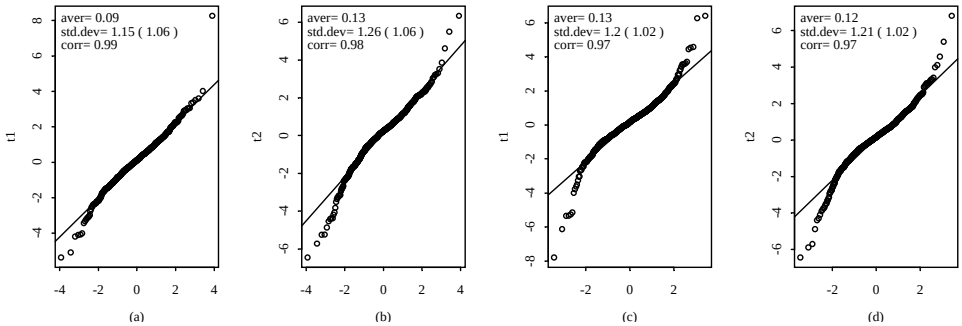

Figure 4. Q-Q plots for t-approximations to the Studentized distribution of $f(x;\widehat{\boldsymbol{\theta}}_N)$. Fitted response is Michaelis-Menten, true response is exponential, errors are heteroscedastic. Values of (n_0, x) are (a) $(0, 3)$, (b) $(0, 6)$, (c) $(15, 3)$, (d) $(15, 6)$.

Michaelis-Menten response, in the sense of (1.2)), are $(\theta_0, \theta_1) = (13.94, 2.45)$. For Figure 1 we took $r_0 = 2$, $r_1 = 4$ and $n_0 = 7$, for Figures 2–4 $r_0 = 2$, $r_1 = 3$ and $n_0 = 10$. In each case we assess the quality of the distributional approximation after the addition of 0 and $n_1 = 15$ sequentially chosen locations. Thus $n = 14$ and 74 for the two situations in Figure 1, while $n = 20$ and 65 for the two situations in each of Figures 2–4. The figures are based on 1000 simulations. The displays on the plots are the empirical means and standard deviations of t^*, with the 'target' theoretical values in parentheses. The correlations between the observed and theoretical quantiles are also given.

As revealed in Figures 1–3, the theoretical and empirical distributions are in very close agreement when the errors are homoscedastic, even when the fitted and true responses disagree. As in Figure 4, heteroscedasticity results in t^* statistics which are somewhat more varied than the t_{n-2} law predicts. See SW for confidence interval coverages, as well as comparisons with some competing (equispaced, locally D-optimal) design strategies.

REFERENCES

Chaudhuri, P., and Mykland, P. A. (1993). Nonlinear experiments: optimal design and inference based on likelihood. *J. Amer. Statist. Assoc.* 88, 538–546.

Domowitz, I., and White, H. (1982). Misspecified models with dependent observations. *J. Econometrics* 20, 35–58.

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

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Ann. Math. Statist.* 42, 1977–1991.

Jaeckel, L.A. (1971). Robust estimates of location: Symmetry and asymmetric contamination. *Ann. Math. Statist.* 42, 1020-1034.

Sinha, S. K., and Wiens, D. P. (2002). Robust, Sequential Designs for Nonlinear Regression. To appear *Canad. J. Statist.*.

White, H. (1984), *Asymptotic Theory for Econometricians*, Academic Press, London.

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

Sanjoy Sinha
Dept. of Mathematics and Statistics
University of Winnipeg
Winnipeg MB R3B 2E9
Canada
s.sinha@uwinnipeg.ca

Douglas P. Wiens
Dept. of Mathematical and Statistical Sciences
University of Alberta
Edmonton AB T6G 2G1
Canada
wiens@stat.ualberta.ca