Journal of Statistical Planning and Inference 144 (2014) 55–62

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Robust static designs for approximately specified nonlinear regression models

Jamil Hasan Karami [a], Douglas P. Wiens [b,*]

[a] Department of Statistics Biostatistics & Informatics, University of Dhaka, Dhaka 1000, Bangladesh
[b] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1

ARTICLE INFO

Available online 13 September 2012

Keywords:
Approximate response
Bayesian optimality
Genetic algorithm
Minimax
Optimal design

ABSTRACT

We outline the construction of robust, static designs for nonlinear regression models. The designs are robust in that they afford protection from increases in the mean squared error resulting from misspecifications of the model fitted by the experimenter. This robustness is obtained through a combination of minimax and Bayesian procedures. We first maximize (over a neighborhood of the fitted response function) and then average (with respect to a prior on the parameters) the sum (over the design space) of the mean squared errors of the predictions. This average maximum loss is then minimized over the class of designs. Averaging with respect to a prior means that there is no remaining dependence on unknown parameters, thus allowing for static, rather than sequential, design construction. The minimization over the class of designs is carried out by implementing a genetic algorithm. Several examples are discussed.

1. Introduction

Suppose that an investigator is planning an experimental study, the data from which will be analyzed by nonlinear regression. Thus, a specific parametric model will be fitted, and from this fit various inferences will be made. The design in such studies is typically chosen with the intention of minimizing some loss function such as the variance, or mean squared error, of the parameter estimates or the predicted values. A problem that arises immediately is that, by the nature of nonlinear models, these common measures depend on the unknown values of the parameters.

One way in which this problem might be handled is by constructing a 'locally optimal' design—one that is optimal only at a particular value $\boldsymbol{\theta}_0$ of the, often vector-valued, parameter. The designer hopes that the optimality will extend, at least approximately, to a neighborhood of $\boldsymbol{\theta}_0$. This parameter might arise from the experimenter's prior knowledge, or perhaps as an estimate from an earlier experiment.

A somewhat more robust approach allows for uncertainty about the parameter values, but not the form of the response function—one first maximizes the chosen loss function (or minimizes an efficiency measure such as the determinant of the information matrix) over a neighborhood of a local parameter $\boldsymbol{\theta}_0$ and then optimizes over the class of designs. See King and Wong (2000), Dette and Biedermann (2003), and Dette and Pepelyshev (2008) for examples of this approach.

With or without robustness considerations, a more practical method – if time allows – is to design sequentially. In this approach, evidently first proposed by Box and Hunter (1965a,b) parameter estimates are recalculated as the data accrue, and subsequent design points are computed in such a way as to optimize a performance measure, evaluated at the current

* Corresponding author.
E-mail addresses: karami@univdhaka.edu (J.H. Karami), doug.wiens@ualberta.ca (D.P. Wiens).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.09.001

estimates. Sinha and Wiens (2002) implement a robust variation of this method. They entertain a class of nonlinear models forming a neighborhood of that which the experimenter will fit, maximize the loss over this neighborhood, and then sequentially construct minimizing designs. We emphasize the 'full' nature of the neighborhoods over which they maximize – in contrast to the work referred to in the previous paragraph, the entire form of the response, rather than merely the parameter value, is allowed to vary. See also Sinha and Wiens (2003) for the asymptotic theory of this approach.

Yet another manner in which one can deal with this problem is through Bayesian optimality. Here the loss function is averaged, with respect to a 'prior' distribution on the parameters; this averaged value is then minimized. See Dette and Neugebauer (1997) for details. We note that it is not necessary to adopt any particular position on Bayesianism in order to take this approach; the introduction of a prior can be viewed merely as a convenient manner in which to eliminate the parameters from the loss function, thus allowing for static, i.e. nonsequential, design construction.

In this article we combine Bayesian optimality with robustness considerations as in Sinha and Wiens (2002), and go on to exhibit static designs with attractive robustness properties. Thus, in Section 2, we introduce a class of nonlinear models forming a neighborhood of that which the experimenter intends to fit. For each member of this class the 'true' parameter $\boldsymbol{\theta}_0$ is defined as that which minimizes an L_2-distance between the true regression response $E[Y|\mathbf{x}]$ and the fitted response $f(\mathbf{x}|\boldsymbol{\theta})$, as $\mathbf{x}$ ranges over a finite 'design space' $\mathcal{S}$. The loss is evaluated at $\boldsymbol{\theta}_0$ and maximized as

$$d(\mathbf{x}|\boldsymbol{\theta}_0) \stackrel{def}{=} E[Y|\mathbf{x}] - f(\mathbf{x}|\boldsymbol{\theta}_0), \tag{1}$$

ranges over an $L_2(\mathcal{S})$-neighborhood of *0*; the resulting maximum is then averaged with respect to a prior $p(\boldsymbol{\theta}_0)$.

In Section 3 we describe a genetic algorithm used to minimize the average maximized loss. Examples and further computational issues are presented and discussed in Section 4. Computing code, written in R, to duplicate these examples is available from us. Derivations are in the Appendix.

The designs constructed here are indexed by a parameter $v \in [0,1]$ representing the relative emphasis that the experimenter places on those errors, in the MSE of the fitted response, which are due to bias versus those due to random variation. When designing with $v = 0$ the experimenter is implicitly stating that he has complete faith in the fitted model $E[Y|\mathbf{x}] = f(\mathbf{x}|\boldsymbol{\theta}_0)$ and that only the parameter $\boldsymbol{\theta}_0$ is in doubt; the resulting designs are Bayesian optimal in the usual sense. It will be seen in the examples of Section 4 that as v increases the designs become more diffuse, hence more robust against errors in the specification of the functional form of the response. As $v \to 1$ the interpretation is that the experimenter has no faith in the accuracy of his specified model and so seeks a uniform exploration of the design space. There is of course a dependence of the design on the form of the prior, but this appears to be slight in comparison with the dependence on v.

Our designs are by construction 'exact', and so there is no difficulty in implementing them. As for their comparative performance—there is no shortage in the literature of simulation studies comparing robust with nonrobust designs; these all tend to give the same, unsurprising message which need not be dwelt upon here. The robust designs give more protection, against model misspecifications, than do the nonrobust designs tailored for efficiency at one particular model, and the latter tend to be more efficient when the fitted model is indeed exactly correct. See for example the case studies in Fang and Wiens (2000) and in Li and Wiens (2011).

2. Approximately specified nonlinear models

In this section we describe our scenario for design in an approximately specified nonlinear regression framework. A 'design' is a specification $(n_1/n, \ldots, n_N/n)$, for integers $\{n_i\}$ summing to n. In implementing the design the experimenter makes n_i observations $\{Y_{ij}\}_{j=1}^{n_i}$ at 'location' $\mathbf{x}_i$—a q-dimensional vector of covariates chosen from a design space $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^{N}$. We note that N, while finite, may be arbitrarily large, and that $n_i \geq 0$—there is no requirement that observations be made at every design point. After sampling, a known – with some reservations – regression response function $f(\mathbf{x}|\cdot)$ will be fitted, viz. $\hat{Y}(\mathbf{x}) = f(\mathbf{x}|\hat{\boldsymbol{\theta}})$. The approximate nature of the model – that $E[Y|\mathbf{x}] \approx f(\mathbf{x}|\boldsymbol{\theta}_0)$ for some p-dimensional $\boldsymbol{\theta}_0$ – is characterized by first defining

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} (E[Y|\mathbf{x}_i] - f(\mathbf{x}_i|\boldsymbol{\theta}))^2. \tag{2}$$

Then with $d(\mathbf{x}|\boldsymbol{\theta}_0)$ given by (1) the probability model is

$$Y(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}_0) + d(\mathbf{x}|\boldsymbol{\theta}_0) + \varepsilon,$$

with random errors ε. We assume independent, homoscedastic errors, with variance σ^2.

Assume that $f(\mathbf{x}|\cdot)$ is differentiable with respect to $\boldsymbol{\theta}$, define $\mathbf{Z}(\boldsymbol{\theta})$ to be the $N \times p$ matrix with i^{th} row

$$\mathbf{z}^T(\mathbf{x}_i|\boldsymbol{\theta}) = \frac{\partial f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

and assume that $\mathbf{Z}(\boldsymbol{\theta}_0)$ has full column rank. A consequence of (2) is that, if $\mathbf{d}(\boldsymbol{\theta}) = (d(\mathbf{x}_1|\boldsymbol{\theta}), \ldots, d(\mathbf{x}_N|\boldsymbol{\theta}))^T$, then

$$\mathbf{Z}^T(\boldsymbol{\theta}_0)\mathbf{d}(\boldsymbol{\theta}_0) = \mathbf{0}_{p\times 1}. \tag{3}$$

In order that the bias in the estimates, due to model misspecification, remain of the same order asymptotically as errors due to random variation, we impose the 'contiguity' requirement

$$\|\mathbf{d}(\boldsymbol{\theta}_0)\| \le \frac{\tau}{\sqrt{n}}, \tag{4}$$

for a nonnegative constant τ.

It is our intention to evaluate the asymptotic mean squared error (*MSE*) of the predicted values, to maximize this *MSE* over a class of models defined by (3) and (4), and to then average this maximized loss with respect to a prior on $\boldsymbol{\theta}_0$. For the first of these steps we assume that the estimate $\hat{\boldsymbol{\theta}}_n$ is computed by Least Squares, although our results extend to Ordinary M-estimation with only minor changes. As a measure of loss we use the average, over $\mathcal{S}$, mean squared error of the predicted values (AMSE)

$$\text{AMSE} = \frac{1}{N}\sum_{i=1}^{N} E[\{f(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_n) - E[Y|\mathbf{x}_i]\}^2]. \tag{5}$$

The following result is proven in the Appendix, with further details in Karami (2011). We use the definitions $\xi_i = n_i/n$, $\mathbf{D}_\xi = \text{diag}(\xi_1,\ldots,\xi_N)$ and

$$\mathbf{R}_\xi(\boldsymbol{\theta}_0) = \mathbf{Z}(\boldsymbol{\theta}_0)(\mathbf{Z}^T(\boldsymbol{\theta}_0)\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta}_0))^{-1}\mathbf{Z}^T(\boldsymbol{\theta}_0).$$

Theorem 1. *An asymptotic, first order approximation to* (5)*, maximized over the neighborhood given by vectors* $\mathbf{d}$ *satisfying* (3) *and* (4)*, is* $((\tau^2+\sigma^2)/nN)\mathcal{L}_v(\boldsymbol{\theta}_0|\xi)$*, where*

$$\mathcal{L}_v(\boldsymbol{\theta}_0|\xi) = (1-v)\text{tr}[\mathbf{R}_\xi(\boldsymbol{\theta}_0)] + v ch_{\max}[\mathbf{R}_\xi(\boldsymbol{\theta}_0)\mathbf{D}_\xi^2(\boldsymbol{\theta}_0)\mathbf{R}_\xi(\boldsymbol{\theta}_0)], \tag{6}$$

$ch_{\max}[\cdot]$ *denotes the maximum eigenvalue, and* $v = \tau^2/(\tau^2+\sigma^2)$.

Now let $p(\boldsymbol{\theta})$ be a density on the parameter space Θ, and define

$$\mathcal{L}_v(\xi) = \int_\Theta \mathcal{L}_v(\boldsymbol{\theta}|\xi)p(\boldsymbol{\theta})\,d\boldsymbol{\theta}. \tag{7}$$

In the next section we discuss the construction of optimally robust designs, i.e. vectors $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_N)^T$ minimizing $\mathcal{L}_v(\xi)$. Note that it is not necessary for the experimenter to furnish values σ^2 or τ^2—he can merely interpret v as expressing the relative importance, to him, of errors due to variance rather than to bias. The limiting cases $v=0$ and $v=1$ correspond to 'pure variance' and 'pure bias' problems, respectively.

3. Minimizing the maximized loss

We minimize $\mathcal{L}_v(\xi)$, at (7), via a genetic algorithm (GA). Such algorithms have been developed using notions of evolutionary theory: we generate 'populations' of designs that evolve over 'generations'. In the evolutionary process 'fit parents' – pairs of designs with small values of $\mathcal{L}_v(\xi)$ – are combined, to produce 'children' via stochastic processes of 'crossover' and 'mutation'. For background material see Karami (2011), Welsh and Wiens (in press) and Mandal et al. (2007).

Explicitly, the GA we have used is as follows. The values of the various tuning parameters will be given in Section 4.

1. Start by randomly generating a first generation of n_g designs. For this, each design is identified with a multinomial vector of length N, with sum n, with each of the N locations being equally likely to be selected at each draw.
2. For the current generation of designs, compute the loss $\mathcal{R}_k = \mathcal{L}_v(\xi_k)$ for each design ξ_k, $k=1,\ldots,n_g$, and the corresponding 'fitness levels'

$$\text{fitness}_k = \frac{1}{(\mathcal{R}_k - .99\mathcal{R}_{\min})^2}, \quad k=1,\ldots,n_g,$$

where $\mathcal{R}_{\min}$ is the minimum value of the loss in the current generation. Scale the fitness levels $\{\text{fitness}_k\}_{k=1}^{n_g}$ to form a probability distribution

$$\psi_k = \frac{\text{fitness}_k}{\sum_{j=1}^{n_g}\text{fitness}_j}, \quad k=1,\ldots,n_g.$$

3. Form a new generation of n_g designs to replace the current generation in the following way.
   (a) Include the fittest $N_{elite} = n_g P_{elite}$ of the current generation; they are an 'elite' group that survives through to the next generation. The remaining $n_g - N_{elite}$ members are formed by crossover and mutation.

(b) Crossover proceeds as follows:
- Choose two members of the current generation to be parents, with probability proportional to their fitness level: If $\zeta_1,\zeta_2 \sim$ independent Uniform(0,1), then choose to be parents the current generation members i_1^* and i_2^*, where

$$i_1^* = \min\left\{ i : \sum_{j=1}^{i} \psi_j \geq \zeta_1 \right\} \quad \text{and} \quad i_2^* = \min\left\{ i : \sum_{j=1}^{i} \psi_j \geq \zeta_2 \right\}.$$

  This is the so-called *roulette-wheel* selection method, common in GA. It ensures that the most fit members of the current population are the most likely to be chosen as parents. The same parent can be chosen twice without posing difficulties for the algorithm.
- With probability $1-P_{crossover}$, the child is identical to the fittest parent.
- With probability $P_{crossover}$, the parents both contribute towards the child, in the following manner. Each member of the current generation can be represented by its vector $n\boldsymbol{\xi}$ of allocations. The two vectors of allocations arising from the parents are averaged, and then any fractional allocations are rounded down. This results in a vector with integer elements, with sum s possibly less than n. If $s < n$ then $n-s$ design points are randomly chosen from $\mathcal{S}$ (with replacement) and added to the design. The child formed in this way is added to the new generation.

(c) Mutation is applied independently to each child – regardless of how the child is formed – as follows. With probability $P_{mutation}$, k elements of the vector $\boldsymbol{\xi}$ defining the child are randomly chosen, and replaced by a multinomial vector of length k, with the same sum as the elements being replaced. The value of k is chosen by the user; we typically use $2 \leq k \leq 6$. With probability $1-P_{mutation}$ we do nothing.

4. Step 3 is carried out until the next generation has been formed. Then its fitness levels are computed and the process is repeated from Step 2. The loss is guaranteed to decrease (weakly) in each generation, because of the inclusion of the elite members. We run the algorithm until the best design has not changed in G consecutive generations.

There is considerable arbitrariness here; in particular there is no canonical crossover method in GA. The crossover method we employ ensures that any design point appearing with a high frequency in both parents continues to do so in the child, reflecting the feeling that such a point might be contributing to the fitness of the parents. Similarly, points absent in both parents are not acquired by the child, except by mutation.

4. Examples; computational issues

In all examples described here we took $P_{crossover} = .95$, $P_{elite} = .1$. The integration was carried out by Simpson's Rule, using a 101-point quadrature for one-dimensional integration, and a 51-point quadrature on each axis for two-dimensional integration. In some cases we found it helpful to run the GA several times, each time 'seeding' the best design found to date into the initial generation, where it would join the elite group until it was improved upon.

An anonymous referee has suggested that we make mention of the methods in Dette et al. (2008), as a way, different from the method outlined above, of obtaining 'initial' designs. In Dette et al. (2008), an algorithm is presented for the construction of approximate, D-optimal designs for linear regression models. Other optimality criteria are mentioned, but none coincide with that being adopted here. Given this, and our focus on *exact* designs for *nonlinear* models, we feel that the additional complexity imposed by the suggestion being made would be, at best, superfluous. We add that the current algorithm is already sufficiently quick. Even when carried out in R, the construction of the designs in Fig. 1 each took only about 7 min of cpu time.

Example 1. Here we take the approximate response $f(x|\theta) = e^{-\theta x}$, with x taking on $N=25$ equally spaced values spanning [0,10]. In (7) we take $p(\theta)$ to be the uniform density on [0,1]. We use generations of size $n_g = 40$, and vary $P_{mutation}$ linearly from 0 to .5 as the number of generations since the last occurrence of an improved design varies from 0 to $G=200$. Specifically, when the currently best design has not changed in g generations, we take $P_{mutation} = .5g/200$. The computations use a design size of $n=70$. See Fig. 1, where we plot the minimax designs for $v = 0,.5,1$. When $v = 0$ the loss arises entirely from variance, which is minimized by an extreme placement of the design points: 43 at $x=2.08$, 10 at $x=2.5$ and 17 at $x=10$. This design is plotted in (b) of Fig. 1; in (a) we plot, for purposes of comparison, the D-optimal design when $v = 0$. In the notation of Theorem 1 this design minimizes (7), with the loss in the integrand replaced by

$$\mathcal{L}_0^D = -\log(\det(\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta}))),$$

the design places 20 observations at $x=1.67$, 46 at $x=2.08$ and 4 at $x=10$. When $v = 1$ the loss arises entirely from bias, and as expected the design allows for a near uniform exploration of the design space – the exact frequencies, as x increases, are $\{0,3,\ldots,3,2,2\}$. For $v = .5$ the design is

x:	.00	.42	.83	1.25	1.67	2.08	2.50	2.92	3.33	3.75	4.17	4.58	5.00
$n\xi$:	0	0	0	8	10	9	8	6	4	3	2	2	1
x:	5.42	5.83	6.25	6.67	7.08	7.50	7.92	8.33	8.75	9.17	9.58	10.0	
$n\xi$:	1	1	0	1	1	1	1	2	2	2	2	3.	

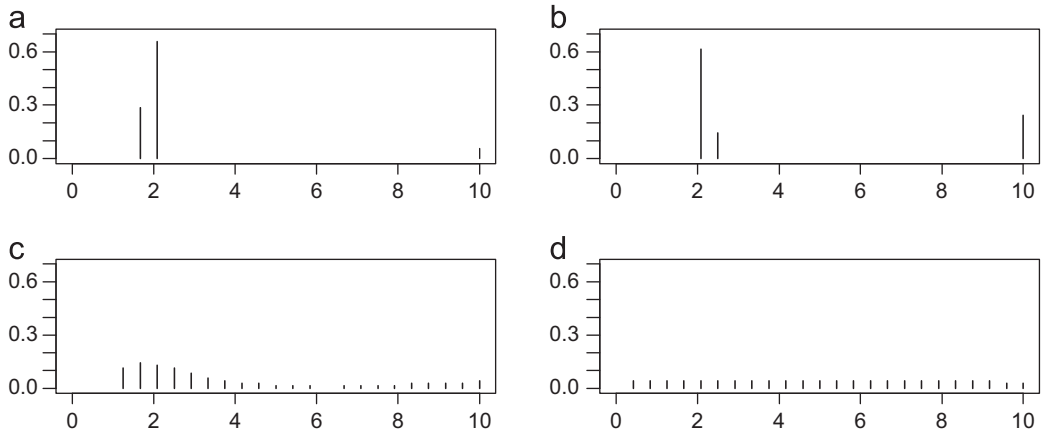

Fig. 1. Minimax designs for Example 1. (a) D-optimal design and (b–d) minimax AMSE designs. Values of the maximum AMSE are: (b) $\mathcal{L}_0(\xi)=17.763$, (c) $\mathcal{L}_{.5}(\xi)=9.985$, and (d) $\mathcal{L}_1(\xi)=1.004$. (a) $v=0$; D-optimal design, (b) $v=0$, (c) $v=.5$ and (d) $v=1$.

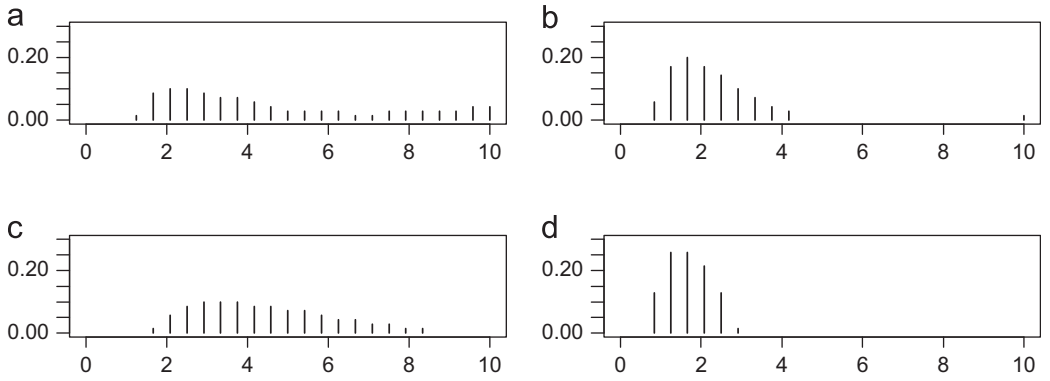

Fig. 2. Designs for the exponential regression model of Example 1; varying parameters of the Beta prior. (a) $\mathcal{L}_{.5}(\xi)=11.380$, (b) $\mathcal{L}_{.5}(\xi)=6.755$, (c) $\mathcal{L}_{.5}(\xi)=10.842$, and (d) $\mathcal{L}_{.5}(\xi)=4.858$. (a) (α,β)=(1,2), (b) (β,α)=(2,1), (c) (α,β)=(2,5) and (d) (α,β)=(5,2).

In Fig. 2 we plot the results when $v=.5$ but the prior is Beta(α,β) for various choices of (α,β). Recall that the uniform density is Beta(1,1). In all cases in this example the mode of the design is near $1/E[\theta]=(\alpha+\beta)/\alpha$. Other aspects of the shape of the designs seem to vary in a more subtle manner with the prior, which should thus be chosen with some care.

Example 2. Bates and Watts (1988) describe an experiment, carried out in 1798 by Count Rumford, an American-born employee of the Bavarian government known for his work on the nature of heat. In this experiment an object was allowed to cool from an initial temperature of 130°F to an ambient temperature of 60 °F, following which a model based on Newton's law of cooling was fitted, using $f(x|\theta)=60+70e^{-\theta x}$. The covariate values, representing 'time', were $x=\{4,5,7,12,14,16,20,24,28,31,34,37.5,41\}$. Here we redesign this experiment, using the same values of x and $n=20$. We view $f(x|\theta)$ as only an approximation, with other environmental factors accounting for errors in this response, and take $v=.5$. The tuning parameters of the GA are as in Example 1, and $p(\theta)$ is again the uniform density on [0,1]. See Fig. 3.

Example 3. Our methods are of course also valid for linear models, in which case $\mathcal{L}_v(\boldsymbol{\theta}|\xi)$ at (6) does not depend upon $\boldsymbol{\theta}$ and there is no need for integration. Fang and Wiens (2000) obtained minimax designs, for the same situations as are entertained in the current work, for approximate cubic regression on $x\in[-1,1]$. The minimization was carried out by simulated annealing, resulting in the design of Fig. 4(a) for $n=20$, $v=1/11$ and a design space consisting of 40 equally spaced points. The loss was $\mathcal{L}_{1/11}(\xi)=116.52$. Using the GA described here ($n_g=30$, $G=5000$) – but modified to yield only symmetric designs – we obtained the improved design of Fig. 4(b), with $\mathcal{L}_{1/11}(\xi)=113.09$.

Example 4. Here we take an approximate, two-parameter Michaelis–Menten model, with $f(x|\boldsymbol{\theta})=\theta_1 x/(\theta_2+x)$. The design space is $x=0(.1)1$, with 11 points, and $n=20$. Our development is based on the 'Puromycin' experiment from Bates and Watts (1988), where estimates $\hat{\boldsymbol{\theta}}=(195.8,.0484)\approx(200,.05)$ were obtained by linearizing $1/f$ and carrying out a

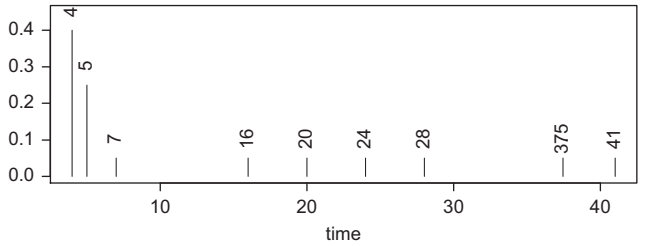

Fig. 3. Minimax design for the Rumford experiment of Example 2; $\mathcal{L}_{.5}(\xi)=3.423$.

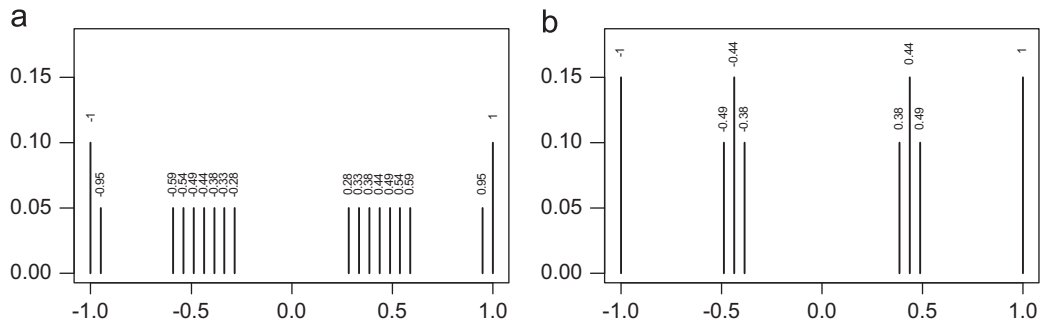

Fig. 4. Design for approximate cubic regression, as in Example 3. (a) Fang/Wiens design, $\mathcal{L}=116.52$ and (b) current improvement, $\mathcal{L}=113.09$.

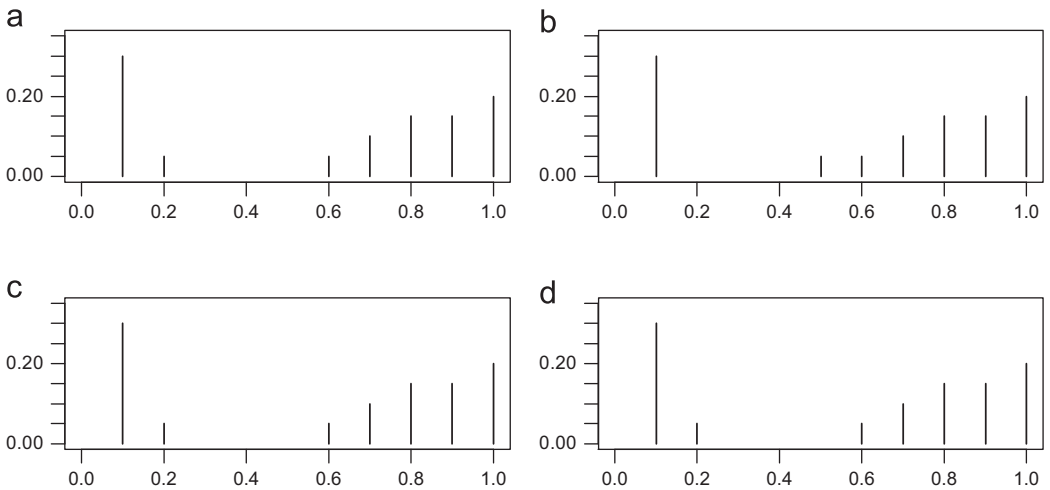

Fig. 5. Designs for the approximate Michaelis–Menten model of Example 4 and various Beta priors. Losses are (a) $\mathcal{L}_{.5}(\xi)=8.52$, (b) $\mathcal{L}_{.5}(\xi)=8.46$, (c) $\mathcal{L}_{.5}(\xi)=8.57$, and (d) $\mathcal{L}_{.5}(\xi)=8.51$.(a) $(\alpha,\beta)=(1,1)$, (b) $(\beta,\alpha)=(2,4)$, (c) $(\alpha,\beta)=(4,2)$ and (d) $(\alpha,\beta)=(20,20)$.

preliminary linear regression. Thus, we introduce variables $T_1,T_2\in[0,1]$, defined by

$$\theta_1=200(T_1+.5)\in[100,300],\quad \theta_2=\frac{T_2+.5}{20}\in[.025,.075].$$

We assign a Beta(α,β) density $p(t|\alpha,\beta)$ to each of T_1,T_2, and thus replace (7) by

$$\mathcal{L}_v(\xi)=\int_0^1\int_0^1 \mathcal{L}_v(\boldsymbol{\theta}(t_1,t_2)|\xi)p(t_1|\alpha,\beta)p(t_2|\alpha,\beta)\,dt_2\,dt_1.$$

We take $v=.5$, $n_g=20$ and obtain the designs shown in Fig. 5, for the combinations $(\alpha,\beta)=(1,1)$, (2,4), (4,2) and (20,20). Recall that the mean and variance of the Beta(α,β) density are $\alpha/(\alpha+\beta)$ and $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$, respectively, so that the choice $(\alpha,\beta)=(20,20)$ concentrates the mass near $\hat{\boldsymbol{\theta}}$. Nonetheless, the message in this example seems to be that the minimax designs are very robust to the shape of the prior—only that at Fig. 5(b) differs slightly from the others.

5. Summary and concluding remarks

We have described a method of obtaining designs for use with nonlinear regression models, when the functional form of the model is in doubt. Our method prescribes that one first work in a neighborhood of the parametric model thought to be a reasonable approximation to the true response. Within this neighborhood the loss (average *MSE* of the predictions) is maximized, and the parameters are integrated out with respect to a prior distribution. The result of this process depends only on the design, which is then chosen to minimize the maximized and integrated loss. A genetic algorithm for carrying out the minimization has been provided and the R code has been made available.

A virtue of the method presented here is that it avoids the need for sequential sampling. The designs have been shown to be quite responsive to the shape of the prior (see in particular Example 1). The genetic algorithm has been shown to perform very efficiently, and to be quite insensitive to the choice of the various tuning constants. The designs are of course robust to alternative model forms, having been constructed with exactly this in mind. They are also intuitively sensible. When the experimenter specifies a great deal of faith in his fitted model the designs are very close to those obtained by more classical principles of design for exactly known models; as this degree of faith is reduced the designs encourage a more complete exploration of the design space.

Acknowledgments

This research has been supported by the Natural Sciences and Engineering Research Council of Canada. The work has benefited from the incisive comments of two anonymous reviewers.

Appendix A. Derivations

Proof of Theorem 1. In terms of

$$\mathbf{b}(\theta_0)=\mathbf{Z}^T(\theta_0)\mathbf{D}_\xi\mathbf{d}(\theta_0):p\times 1,\ \mathbf{M}(\theta_0)=\mathbf{Z}^T(\theta_0)\mathbf{D}_\xi\mathbf{Z}(\theta_0):p\times p, \tag{A.1}$$

the asymptotic theory of Gallant (1987) gives us that $\sqrt{n}(\hat{\theta}_n-\theta_0)$ is asymptotically normally distributed, with asymptotic mean $\sqrt{n}\mathbf{M}^{-1}(\theta_0)\mathbf{b}(\theta_0)$ and asymptotic covariance matrix $\sigma^2\mathbf{M}^{-1}(\theta_0)$. We first approximate AMSE by taking a first order expansion of $f(\mathbf{x}_i|\hat{\theta})$

$$\begin{aligned}\text{AMSE}&=\frac{1}{N}\sum_{i=1}^{N}E[\{(f(\mathbf{x}_i|\hat{\theta}_n)-f(\mathbf{x}_i|\theta_0))+d(\mathbf{x}_i|\theta_0)\}^2]\\&\approx\frac{1}{N}\sum_{i=1}^{N}E[\{(\mathbf{z}^T(\mathbf{x}_i|\theta_0)(\hat{\theta}_n-\theta_0))+d(\mathbf{x}_i|\theta_0)\}^2]\\&=\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}^T(\mathbf{x}_i|\theta_0)E[(\hat{\theta}_n-\theta_0)(\hat{\theta}_n-\theta_0)^T]\mathbf{z}(\mathbf{x}_i|\theta_0)+\frac{1}{N}\|\mathbf{d}(\theta_0)\|^2.\end{aligned}$$

In the last line – whose value we now define as $L(\theta_0|d)$ – the cross-product has vanished by virtue of (3). Upon inserting the asymptotic mean and covariance matrix, and (A.1), this loss becomes

$$L(\theta_0|d)=\frac{1}{nN}\sum_{i=1}^{N}\{\mathbf{z}^T(\mathbf{x}_i|\theta_0)[n\mathbf{M}^{-1}(\theta_0)\mathbf{b}(\theta_0)\mathbf{b}^T(\theta_0)\mathbf{M}^{-1}(\theta_0)+\sigma^2\mathbf{M}^{-1}(\theta_0)]\mathbf{z}(\mathbf{x}_i|\theta_0)\}+\frac{1}{N}\|\mathbf{d}(\theta_0)\|^2 \tag{A.2a}$$

$$\begin{aligned}L(\theta_0|d)=&\frac{1}{nN}[n\mathbf{d}^T(\theta_0)\mathbf{D}_\xi\mathbf{Z}(\theta_0)\mathbf{M}^{-1}(\theta_0)\mathbf{Z}^T(\theta_0)\mathbf{Z}(\theta_0)\mathbf{M}^{-1}(\theta_0)\mathbf{Z}^T(\theta_0)\mathbf{D}_\xi\mathbf{d}(\theta_0)]\\&+\frac{\sigma^2}{nN}\operatorname{tr}[\mathbf{Z}(\theta_0)\mathbf{M}^{-1}(\theta_0)\mathbf{Z}^T(\theta_0)]+\frac{1}{N}\|\mathbf{d}(\theta_0)\|^2.\end{aligned} \tag{A.2b}$$

We now maximize over $\mathbf{d}(\theta_0)$; for this it is convenient to first change the basis. Let $\mathbf{U}(\theta_0)$ be an $N\times p$ matrix, whose columns form an orthogonal basis for the column space of $\mathbf{Z}(\theta_0)$, and augment these columns by $\tilde{\mathbf{U}}(\theta_0):N\times N-p$ in such a way that $[\mathbf{U}(\theta_0)\vdots\tilde{\mathbf{U}}(\theta_0)]:N\times N$ is orthogonal. Then the columns of $\tilde{\mathbf{U}}(\theta_0)$ form an orthogonal basis for the orthogonal complement to the column space of $\mathbf{Z}(\theta_0)$. Although our final results do not depend on these matrices, we remark that $\mathbf{U}(\theta_0)$ is the 'Q' in the QR decomposition of $\mathbf{Z}(\theta_0)$, and $\tilde{\mathbf{U}}(\theta_0)$ is typically offered as a by-product (the 'completion') of this computation. Condition (3) now says that $\mathbf{d}(\theta_0)$ lies in the column space of $\tilde{\mathbf{U}}(\theta_0)$, i.e. $\mathbf{d}(\theta_0)=\tilde{\mathbf{U}}(\theta_0)\mathbf{c}(\theta_0)$, for some $\mathbf{c}(\theta_0)$ with the same norm as $\mathbf{d}(\theta_0)$. Now (A.2a) becomes

$$\frac{\tau^2}{nN}\left[\left(\frac{\sqrt{n}\mathbf{c}(\theta_0)}{\tau}\right)^T\tilde{\mathbf{U}}^T(\theta_0)\mathbf{D}_\xi\mathbf{Z}(\theta_0)\mathbf{M}^{-1}(\theta_0)\mathbf{Z}^T(\theta_0)\mathbf{Z}(\theta_0)\mathbf{M}^{-1}(\theta_0)\mathbf{Z}^T(\theta_0\mathbf{D}_\xi\tilde{\mathbf{U}}(\theta_0)\left(\frac{\sqrt{n}\mathbf{c}(\theta_0)}{\tau}\right)\right],$$

whose maximum over vectors $\sqrt{n}\mathbf{c}(\boldsymbol{\theta}_0)/\tau$ of norm ≤ 1 is

$$\frac{\tau^2}{nN} ch_{\max}[\tilde{\mathbf{U}}^T(\boldsymbol{\theta}_0)\mathbf{D}_\xi \mathbf{Z}(\boldsymbol{\theta}_0)\mathbf{M}^{-1}(\boldsymbol{\theta}_0)\mathbf{Z}^T(\boldsymbol{\theta}_0)\mathbf{Z}(\boldsymbol{\theta}_0)\mathbf{M}^{-1}(\boldsymbol{\theta}_0)\mathbf{Z}^T(\boldsymbol{\theta}_0)\mathbf{D}_\xi \tilde{\mathbf{U}}(\boldsymbol{\theta}_0)],$$

attained at an eigenvector of unit norm. Any such vector also maximizes $\|\mathbf{d}(\boldsymbol{\theta}_0)\|^2$ at (A.2b), and so

$$\max_d L(\boldsymbol{\theta}_0|d) = \frac{1}{nN}\left\{\begin{array}{c}\tau^2 ch_{\max}[\tilde{\mathbf{U}}^T(\boldsymbol{\theta}_0)\mathbf{D}_\xi \mathbf{Z}(\boldsymbol{\theta}_0)\mathbf{M}^{-1}(\boldsymbol{\theta}_0)\mathbf{Z}^T(\boldsymbol{\theta}_0)\mathbf{Z}(\boldsymbol{\theta}_0)\mathbf{M}^{-1}(\boldsymbol{\theta}_0)\mathbf{Z}^T(\boldsymbol{\theta}_0)\mathbf{D}_\xi \tilde{\mathbf{U}}(\boldsymbol{\theta}_0)] \\ +\sigma^2 \operatorname{tr}[\mathbf{Z}(\boldsymbol{\theta}_0)\mathbf{M}^{-1}(\boldsymbol{\theta}_0)\mathbf{Z}^T(\boldsymbol{\theta}_0)] + \tau^2\end{array}\right\}.$$

The nonzero eigenvalues of a product of two matrices do not change if the order of this product is reversed. Repeated applications of this maxim, together with the identity $\tilde{\mathbf{U}}(\boldsymbol{\theta}_0)\tilde{\mathbf{U}}^T(\boldsymbol{\theta}_0) = \mathbf{I}_N - \mathbf{U}(\boldsymbol{\theta}_0)\mathbf{U}^T(\boldsymbol{\theta}_0)$, results in

$$\max_d L(\boldsymbol{\theta}_0|d) = \frac{1}{nN}\left\{\begin{array}{c}\tau^2 ch_{\max}[\mathbf{Z}(\boldsymbol{\theta}_0)\mathbf{M}^{-1}(\boldsymbol{\theta}_0)\mathbf{Z}^T(\boldsymbol{\theta}_0)\mathbf{D}_\xi^2\mathbf{Z}(\boldsymbol{\theta}_0)\mathbf{M}^{-1}(\boldsymbol{\theta}_0)\mathbf{Z}^T(\boldsymbol{\theta}_0)] \\ +\sigma^2 \operatorname{tr}[\mathbf{Z}(\boldsymbol{\theta}_0)\mathbf{M}^{-1}(\boldsymbol{\theta}_0)\mathbf{Z}^T(\boldsymbol{\theta}_0)]\end{array}\right\},$$

from which (6) follows. □

References

Bates, D.M., Watts, D.G., 1988. Nonlinear Regression Analysis and Its Applications. Wiley.

Box, G.E.P., Hunter, W.G., 1965a. The experimental study of physical mechanisms. Technometrics 7, 23–42.

Box, G.E.P., Hunter, W.G., 1965b. Sequential design of experiments for nonlinear models. In: Proceedings of the IBM Scientific Computing Symposium on Statistics, October 21-23, 1963, pp. 113–137.

Dette, H., Neugebauer, H.-M., 1997. Bayesian D-optimal designs for exponential regression models. Journal of Statistical Planning and Inference 60, 331–349.

Dette, H., Biedermann, S., 2003. Robust and efficient designs for the Michaelis–Menten model. Journal of the American Statistical Association 98, 679–686.

Dette, H., Pepelyshev, A., 2008. Efficient experimental designs for sigmoidal growth models. Journal of Statistical Planning and Inference 138, 2–17.

Dette, H., Pepelyshev, A., Zhigljavsky, A., 2008. Improving Updating Rules in Multiplicative Algorithms for Computing D-optimal designs. Computational Statistics and Data Analysis 53, 312–320.

Fang, Z., 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.

Gallant, A.R., 1987. Nonlinear Statistical Models. Wiley.

Karami, J.H., 2011. Designs for Nonlinear Regression With a Prior on the Parameters. Unpublished MSc Thesis. University of Alberta, Department of Mathematical and Statistical Sciences.

King, J., Wong, W.-K., 2000. Minimax D-optimal designs for the logistic model. Biometrics 56, 1263–1267.

Li, P., Wiens, D.P., 2011. Robustness of design for dose–response studies. Journal of the Royal Statistical Society (Series B) 17, 215–238.

Mandal, A., Johnson, K., Wu, J.C.F., Bornemeier, D., 2007. Identifying promising compounds in drug discovery: genetic algorithms and some new statistical techniques. Journal of Chemical Information and Modeling 47, 981–988.

Sinha, S., Wiens, D.P., 2002. Robust sequential designs for nonlinear regression. The Canadian Journal of Statistics 30, 601–618.

Sinha, S., Wiens, D.P., 2003. Asymptotics for robust sequential designs in misspecified regression models. In: Moore, M., Léger, C., Froda, S. (Eds.), Mathematical Statistics and Applications: Festschrift for Constance van Eeden. IMS Lecture Notes—Monograph Series, pp. 233–248.

Welsh, A.H., Wiens, D.P. Robust model-based sampling designs. Statistics and Computing, http://dx.doi.org/10.1007/s11222-012-9339-3, in press.