Journal of the
Royal Statistical Society

SERIES B
Statistical
Methodology

J. R. Statist. Soc. B (2009)
71, *Part* 4, *pp.* 805–829

Robust discrimination designs

Douglas P. Wiens

University of Alberta, Edmonton, Canada

[Received February 2008. Revised January 2009]

Summary. We study the construction of experimental designs, the purpose of which is to aid in the discrimination between two possibly non-linear regression models, each of which might be only approximately specified. A rough description of our approach is that we impose neighbourhood structures on each regression response and determine the members of these neighbourhoods which are least favourable in the sense of minimizing the Kullback–Leibler divergence. Designs are obtained which maximize this minimum divergence. Both static and sequential approaches are studied. We then consider sequential designs whose purpose is initially to discriminate, but which move their emphasis towards efficient estimation or prediction as one model becomes favoured over the other.

Keywords: Discrimination; *D*–*T*-optimality; Integrated Kullback–Leibler optimality; Kullback–Leibler; Kullback–Leibler optimality; Maximin; Michaelis–Menten model; Neyman–Pearson test; Non-linear regression; Robustness; Sequential designs; Simulated annealing; *T*-optimal design

1. Introduction

Consider the following scenario. An investigator is designing an experiment, the purpose of which is to enable him to distinguish between two models of the data. Each model incorporates a response variable which is dependent on various covariates through a possibly non-linear regression response and a possibly non-normal stochastic component. A complicating factor is that, in each model, the specified response function is readily acknowledged to be at best only approximately correct.

Specifically, we suppose that the experimenter is faced with two possibilities. Under the first, his data arise from a density $f_0(y|\mathbf{x},\mu_0,\varphi_0)$; under the other the density is $f_1(y|\mathbf{x},\mu_1,\varphi_1)$. In either case, a random variable Y is observed, together with d-dimensional covariates $\mathbf{x}$, chosen by design. Under model j the mean conditional response is

$$\mu_j(\mathbf{x})=\int y\, f_j(y|\mathbf{x},\mu_j,\varphi_j)\,\mathrm{d}y.$$

The remaining term φ_j represents a, possibly vector-valued, nuisance parameter. Henceforth, φ_j will not be explicitly mentioned if there is no possibility of confusion.

Our approach can be motivated by casting the problem as a problem of hypothesis testing. Given a finite design space $\mathcal{S}=\{\mathbf{x}_i\}_{i=1}^N\subset\mathbb{R}^d$ with $n_i\geqslant 0$ observations $\{y_{ij}\}_{j=1}^{n_i}$ made at $\mathbf{x}_i$, and, if the parameters are completely specified under each hypothesis, the Neyman–Pearson test of $H_0: f_0(y|\mathbf{x},\mu_0)$ *versus* $H_1: f_1(y|\mathbf{x},\mu_1)$ rejects for large values of $R=\Sigma_{i,j}R_{ij}$, where

$$R_{ij}=2\log\left\{\frac{f_1(y_{ij}|\mathbf{x}_i,\mu_1)}{f_0(y_{ij}|\mathbf{x}_i,\mu_0)}\right\}.$$

Address for correspondence: Douglas P. Wiens, Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, T6G 2G1, Canada.
E-mail: doug.wiens@ualberta.ca

1369–7412/09/71805

Under the large sample approximation to the distribution of R given in theorem 1 of this paper, the power of the test is maximized by the design maximizing

$$E_{H_1}[R]=2n\int \mathcal{I}\{\mu_0(\mathbf{x}),\mu_1(\mathbf{x})\}\,\xi(\mathrm{d}\mathbf{x}), \tag{1}$$

where $n=\Sigma_{i=1}^N n_i$, ξ is the design measure placing mass $\xi_i=n_i/n$ at $\mathbf{x}_i$ and

$$\mathcal{I}\{\mu_0(\mathbf{x}),\mu_1(\mathbf{x})\}=\int_{-\infty}^{\infty} f_1(y|\mathbf{x},\mu_1)\log\left\{\frac{f_1(y|\mathbf{x},\mu_1)}{f_0(y|\mathbf{x},\mu_0)}\right\}\mathrm{d}y$$

is the Kullback–Leibler (KL) divergence, measuring the information which is lost when f_0 is used to approximate f_1.

To be precise, and for ease of reference in our numerical simulations below, we give a statement of the asymptotic distribution of R under conditions which are satisfied in these simulations. We assume the usual regularity conditions for likelihood estimation—these are stated precisely in Wiens (2009)—and we consider contiguous alternatives.

Theorem 1 (Wiens, 2009). Suppose that the densities f_0 and f_1 are the same, i.e. $f_j(y|\mathbf{x},\mu_j,\varphi_j) = f(y|\mathbf{x},\mu_j,\varphi)$ for a density f, and that $\mu_1(\mathbf{x}_i)=\mu_0(\mathbf{x}_i)+n^{-1/2}\Delta_i$, $i=1,\ldots,N$. Define

$$D=\int_{\mathcal{S}}\mathcal{I}\{\mu_0(\mathbf{x}),\mu_1(\mathbf{x})\}\,\xi(d\mathbf{x})=\sum_{i=1}^{N}\mathcal{I}\{\mu_0(\mathbf{x}_i),\mu_1(\mathbf{x}_i)\}\,\xi_i.$$

Then we have the following results.

(a) Under this sequence of contiguous alternatives D is $O(n^{-1})$ and R is asymptotically normally distributed under each hypothesis: under H_0,

$$\frac{R+2nD}{\surd(8nD)}\xrightarrow{\mathrm{L}}N(0,1);$$

under H_1,

$$\frac{R-2nD}{\surd(8nD)}\xrightarrow{\mathrm{L}}N(0,1).$$

Thus a test with asymptotic size α rejects for $R>z_\alpha\surd(8nD)-2nD$ and has asymptotic power

$$\beta=\Phi\{\surd(2nD)-z_\alpha\}, \tag{2}$$

where Φ is the $N(0,1)$ distribution function and $z_\alpha=\Phi^{-1}(1-\alpha)$.

(b) If f is the normal density with variance σ^2 then $2nD=\Sigma_{i=1}^N\Delta_i^2\xi_i/\sigma^2$ and these asymptotic distributions are exact, for all n.

The means $\mu_j(\mathbf{x})$ are generally only partially known. We assume that the experimenter models $\mu_j(\mathbf{x})$ parametrically as $\eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$, with the form of η_j specified (perhaps erroneously) but $\boldsymbol{\theta}_j$ unknown. Then if, for instance, both densities are normal, with common variance parameter $\varphi=\sigma^2$, expression (1) becomes

$$(2\sigma^2)^{-1}\sum_i n_i\{\eta_1(\mathbf{x}_i|\boldsymbol{\theta}_1)-\eta_0(\mathbf{x}_i|\boldsymbol{\theta}_0)\}^2.$$

For this case Hunter and Reiner (1965) proposed a sequential method to construct the design: after n observations have been made and estimates $\hat{\boldsymbol{\theta}}_j$ computed, the next observation should be made at that point $\mathbf{x}_{\text{new}}$ maximizing $\{\eta_1(\mathbf{x}|\hat{\boldsymbol{\theta}}_1)-\eta_0(\mathbf{x}|\hat{\boldsymbol{\theta}}_0)\}^2$. Fedorov and Pazman (1968) extended this approach to heteroscedastic models.

The construction of static, i.e. non-sequential, designs in non-linear models is more problematic. All such problems must address, in one way or another, the issue that the criterion to be optimized depends on the unknown parameters and that one cannot rely on estimates. Fedorov (1975) suggested the *maximin* procedure of maximizing expression (1) after first minimizing over $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. Atkinson and Fedorov (1975a)—see also Atkinson and Fedorov (1975b)—assumed that model 1 was known to be the correct model, that $\boldsymbol{\theta}_1$ was known, and constructed designs, termed *T-optimal* designs, maximizing

$$\inf_{\boldsymbol{\theta}_0}\left[\int\{\eta_1(\mathbf{x}|\boldsymbol{\theta}_1)-\eta_0(\mathbf{x}|\boldsymbol{\theta}_0)\}^2\,\xi(\mathrm{d}\mathbf{x})\right].$$

In this framework Dette and Titoff (2008) linked the T-optimality problem to a problem of optimal non-linear approximation and went on to obtain results on the number of support points required. Uciński and Bogacka (2005) considered extensions to multiresponse models.

Non-normality imposes further complications. López-Fidalgo *et al.* (2007) studied extensions of these notions to non-normal models, leading to the maximization of

$$\inf_{\boldsymbol{\theta}_0}\left[\int\mathcal{I}\{\eta_0(\mathbf{x}|\boldsymbol{\theta}_0),\eta_1(\mathbf{x}|\boldsymbol{\theta}_1)\}\,\xi(\mathrm{d}\mathbf{x})\right];$$

this criterion is termed *KL-optimality*.

For an interesting discussion of these and other competing methods, see Hill (1978). Note, however, that all these approaches assume that the true model form is specified correctly by one of η_0 and η_1. The dangers that are inherent in such assumptions were elegantly described for regression in general in Box and Draper (1959), page 622, and specifically for non-linear models by Ford *et al.* (1989), page 54.

Applying a method that is highly dependent on a specific model form violates modern notions of robustness. It is our purpose in this paper to propose methods of discrimination design which are robust against model misspecification. The work can be viewed as a natural sequel to Wiens (1991), in which the *uniform* design was shown to have desirable maximin properties (maximizing the minimum power) with respect to lack-of-fit testing in the face of model misspecification (see Biedermann and Dette (2001) and Bischoff and Miller (2006) for extensions), and to Sinha and Wiens (2002), in which robust designs for estimation and prediction in non-linear regression were constructed.

In the next section we formulate the robust discrimination design problem. We describe there our approach to the issue of dependence of the optimality criterion on unknown parameters, via the adoption of a 'working response' with respect to which parameters are defined in the two models. In Section 3 optimally robust designs are derived under an assumption of normality. In Section 4 this assumption is dropped and designs are derived in some non-normal situations. In Section 5 we consider sequential design strategies for discrimination and in Section 6 propose construction methods which also take into account the requirements of efficient parameter estimation and response prediction, as the true nature of the model becomes more apparent to the experimenter. Derivations and longer mathematical arguments are in Appendix A. The MATLAB code that was used to analyse the data can be obtained from `http:www.stat.ualberta.ca/~wiens/`.

2. A robustification of the design problem

In the formulation that was outlined in Section 1, an immediate difficulty is that if $\eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$ only approximates the mean response then the meaning of the parameter $\boldsymbol{\theta}_j$ becomes unclear. To address this we shall adopt a working response $E[Y|\mathbf{x}]$; given such a response we *define*

$$\boldsymbol{\theta}_j = \arg\min_{\boldsymbol{\theta}} \left[\sum_{S} \{E[Y|\mathbf{x}_i] - \eta_j(\mathbf{x}_i|\boldsymbol{\theta})\}^2 \right], \tag{3}$$

i.e. $\boldsymbol{\theta}_j$ is to provide the closest agreement, in this L^2-sense, between the working response and that in model j. We assume that the minimizers $\boldsymbol{\theta}_j$ are unique.

A working response can be obtained in various ways. In some cases—see for instance example 1—we take $E[Y|\mathbf{x}] = \eta_1(\mathbf{x}|\boldsymbol{\theta}_1)$ for a particular value $\boldsymbol{\theta}_1$. Then $\boldsymbol{\theta}_0$ is obtained as at definition (3), with $E[Y|\mathbf{x}_i] = \eta_1(\mathbf{x}_i|\boldsymbol{\theta}_1)$. In other cases we compute both $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, from definition (3), relative to a working response $E[Y|\mathbf{x}]$. A possible approach is to average $\eta_1(x|\boldsymbol{\theta})$ (for instance) over a plausible class of parameters, which is specified by a prior $p(\boldsymbol{\theta})$, obtaining

$$E[Y|\mathbf{x}] = \int \eta_1(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}.$$

This is in the flavour of the Bayesian solution of Atkinson and Fedorov (1975a). In our examples we take a simpler approach and adopt a linear response approximating $\eta_0(\mathbf{x}|\boldsymbol{\theta}_0)$—see example 2.

Note that the working reponse is introduced only as a means of defining parameters; it is not assumed that this is the 'true' response. Indeed, under each hypothesis the true response is assumed only to be a function $\mu_j(\cdot)$ in a neighbourhood of $\eta_j(\cdot|\boldsymbol{\theta}_j)$. Although the optimal static robust designs depend on the working response, our experience has been that this dependence is slight and, in any event, one is free to study the classes of designs which are obtained as the working response varies and thereby to discern the quantitative nature of the solutions.

Remark 1. If both models are linear, with model 0 nested within model 1, then the discrimination problem can be reduced to that of testing that a subvector of the parameter vector is $\mathbf{0}$. The criterion depends only on the remaining parameters, and there are numerous ways of eliminating these, including but not limited to the methods of this paper. The corresponding design problem has been well studied, for instance by Atkinson and Cox (1974), Pukelsheim and Rosenberger (1993) and Dette and Kwiecien (2004). Robust designs can be obtained by using methods as in Wiens (1992) and Fang and Wiens (2000).

Set $\delta_j(\mathbf{x}) = E[Y|\mathbf{x}] - \eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$. Then with

$$\begin{aligned}
\boldsymbol{\delta}_j &= (\delta_j(\mathbf{x}_1), \ldots, \delta_j(\mathbf{x}_N))^{\mathrm{T}}, N \times 1, \\
\boldsymbol{\eta}_j = \boldsymbol{\eta}_j(\boldsymbol{\theta}_j) &= (\eta_j(\mathbf{x}_1|\boldsymbol{\theta}_j), \ldots, \eta_j(\mathbf{x}_N|\boldsymbol{\theta}_j))^{\mathrm{T}}, N \times 1, \\
\mathbf{U}_j = \mathbf{U}_j(\boldsymbol{\theta}_j) = \frac{\partial \boldsymbol{\eta}_j(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}} &= (\dot{\eta}_j(\mathbf{x}_1|\boldsymbol{\theta}_j), \ldots, \dot{\eta}_j(\mathbf{x}_N|\boldsymbol{\theta}_j))^{\mathrm{T}}, N \times p_j,
\end{aligned}$$

we have from definition (3) that $\mathbf{U}_j^{\mathrm{T}} \boldsymbol{\delta}_j = \mathbf{0}$. Define (convex) neighbourhoods of the $\eta_j(\cdot)$ by

$$\mathcal{M}_j = \{\boldsymbol{\eta}_j + \boldsymbol{\delta}_j | \mathbf{U}_j^{\mathrm{T}} \boldsymbol{\delta}_j = \mathbf{0}, \|\boldsymbol{\delta}_j\| \leqslant \tau_j\}.$$

The radii τ_j are to be chosen to ensure that

$$\mathcal{M}_0 \cap \mathcal{M}_1 = \emptyset. \tag{4}$$

For this, note that if $\mathcal{M}_0 \cap \mathcal{M}_1 \neq \emptyset$ then there are $\boldsymbol{\delta}_j$ for which, with $\boldsymbol{\eta}_d =^{\mathrm{def}} \boldsymbol{\eta}_1 - \boldsymbol{\eta}_0$, we have that $\boldsymbol{\eta}_d + \boldsymbol{\delta}_1 - \boldsymbol{\delta}_0 = \mathbf{0}$. Then $\|\boldsymbol{\eta}_d\|^2 = \|\boldsymbol{\delta}_0 - \boldsymbol{\delta}_1\|^2 \leqslant (\|\boldsymbol{\delta}_0\| + \|\boldsymbol{\delta}_1\|)^2 \leqslant (\tau_0 + \tau_1)^2$. Thus we assume that

$$0 \leqslant \tau_0 + \tau_1 < \|\boldsymbol{\eta}_d\|, \tag{5}$$

ensuring condition (4).

Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)^{\mathrm{T}}$, with ξ_i as in Section 1. The *robust KL-optimal design* problem is, in this notation, that of determining

$$\boldsymbol{\xi}^* = \arg\max_{\boldsymbol{\xi}} \min_{\delta_0, \delta_1} \left[\sum_{i=1}^{N} \mathcal{I}\{\mu_0(\mathbf{x}_i), \mu_1(\mathbf{x}_i)\}\, \xi_i \right], \tag{6}$$

with

$$\mu_j(\mathbf{x}) = \eta_j(\mathbf{x}|\boldsymbol{\theta}_j) + \delta_j(\mathbf{x}), \qquad \text{subject to } \mathbf{U}_j^{\mathrm{T}} \boldsymbol{\delta}_j = \mathbf{0} \text{ and } \|\boldsymbol{\delta}_j\| \leqslant \tau_j \text{ for } j = 0, 1.$$

In some cases it is convenient to address the orthogonality requirements $\mathbf{U}_j^{\mathrm{T}} \boldsymbol{\delta}_j = \mathbf{0}$ directly. These hold if and only if each $\boldsymbol{\delta}_j$ lies in the orthogonal complement to the column space of $\mathbf{U}_j$. Let $\mathbf{V}_j$, $N \times (N - p_j)$, be a matrix whose columns form an orthonormal basis for this orthogonal complement. (Numerically, this is typically furnished as a by-product of the QR-decomposition of $\mathbf{U}_j$.) Then $\boldsymbol{\delta}_j = \mathbf{V}_j \mathbf{c}_j$ for some $\mathbf{c}_j \in \mathbb{R}^{N-p_j}$ and $\|\boldsymbol{\delta}_j\| = \|\mathbf{c}_j\|$. With $\mathbf{v}_{j,i}^{\mathrm{T}}$ denoting the ith row of $\mathbf{V}_j$ problem (6) can now be phrased as that of determining

$$\boldsymbol{\xi}^* = \arg\max_{\boldsymbol{\xi}} \min_{\mathbf{c}_0, \mathbf{c}_1} \left[\sum_{i=1}^{N} \mathcal{I}\{\mu_0(\mathbf{x}_i), \mu_1(\mathbf{x}_i)\}\, \xi_i \right], \tag{7}$$

with

$$\mu_j(\mathbf{x}_i) = \eta_j(\mathbf{x}_i|\boldsymbol{\theta}_j) + \mathbf{v}_{j,i}^{\mathrm{T}} \mathbf{c}_j, \qquad \text{subject to } \|\mathbf{c}_j\| \leqslant \tau_j \text{ for } j = 0, 1.$$

The minimization problem (7) is of dimension $2N - p_0 - p_1$, but it can sometimes—see the next section—be carried out analytically. When this is not so—see Section 4—a different type of reduction may be more convenient. With μ_0 and μ_1 as in problem (6), define

$$\mathcal{D}(\boldsymbol{\xi}|\boldsymbol{\delta}_0, \boldsymbol{\delta}_1) = \sum_{i=1}^{N} \mathcal{I}\{\mu_0(\mathbf{x}_i), \mu_1(\mathbf{x}_i)\}\, \xi_i. \tag{8}$$

For the minimization step it is clearly sufficient to restrict the sum in the definition of $\mathcal{D}(\boldsymbol{\xi}|\boldsymbol{\delta}_0, \boldsymbol{\delta}_1)$ to the set $\{\xi_i > 0\}$. Let $s \leqslant n$ be the cardinality of this set. Assume that the elements of $\mathcal{S}$ have been relabelled in such a way that these s design points are the first. Partition $\boldsymbol{\xi}$, $\boldsymbol{\delta}_j$ and $\mathbf{V}_j$ compatibly as

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\xi}^{(1)},\ s \times 1 \\ \boldsymbol{\xi}^{(2)},\ (N-s) \times 1 \end{pmatrix},$$

$$\boldsymbol{\delta}_j = \begin{pmatrix} \boldsymbol{\delta}_j^{(1)},\ s \times 1 \\ \boldsymbol{\delta}_j^{(2)},\ (N-s) \times 1 \end{pmatrix},$$

$$\mathbf{V}_j = \begin{pmatrix} \mathbf{V}_j^{(1)},\ s \times (N - p_j) \\ \mathbf{V}_j^{(2)},\ (N-s) \times (N - p_j) \end{pmatrix}.$$

Proposition 1. With notation as above, suppose that each $\mathbf{V}_j^{(1)}$ has full rank $s \leqslant N - p_j$. Then the problem of minimizing $\mathcal{D}(\boldsymbol{\xi}|\boldsymbol{\delta}_0, \boldsymbol{\delta}_1)$ over $(\boldsymbol{\delta}_0, \boldsymbol{\delta}_1)$ is equivalent to that of determining

$$(\boldsymbol{\delta}_0^*, \boldsymbol{\delta}_1^*) = \arg\min_{\boldsymbol{\delta}_0^{(1)}, \boldsymbol{\delta}_1^{(1)}} \{\mathcal{D}'(\boldsymbol{\xi}^{(1)}|\boldsymbol{\delta}_0^{(1)}, \boldsymbol{\delta}_1^{(1)})\} \tag{9}$$

with

$$\mathcal{D}'(\boldsymbol{\xi}^{(1)}|\boldsymbol{\delta}_0^{(1)},\boldsymbol{\delta}_1^{(1)})=\sum_{i=1}^{s}\mathcal{I}\{\eta_0(\mathbf{x}_i|\boldsymbol{\theta}_0)+\delta_0^{(1)}(\mathbf{x}_i),\eta_1(\mathbf{x}_i|\boldsymbol{\theta}_1)+\delta_1^{(1)}(\mathbf{x}_i)\}\,\xi_i^{(1)},$$

$$\text{subject to } \boldsymbol{\delta}_j^{(1)^{\mathrm{T}}}(\mathbf{V}_j^{(1)}\mathbf{V}_j^{(1)^{\mathrm{T}}})^{-1}\boldsymbol{\delta}_j^{(1)}\leqslant\tau_j^2 \text{ for } j=0,1.$$

Remark 2. Proposition 1 neither gives nor requires any information about the minimizing $\boldsymbol{\delta}_j^{(2)}$ beyond the requirements (which are necessary to satisfy the constraints $\mathbf{U}_j^{\mathrm{T}}\boldsymbol{\delta}_j=\mathbf{0}$) that each $\boldsymbol{\delta}_j$ lie in the column space of $\mathbf{V}_j$. Since $\boldsymbol{\delta}_j^{(1)}=\mathbf{V}_j^{(1)}\mathbf{c}_j^*$ for $\mathbf{c}_j^*=\mathbf{V}_j^{(1)^{\mathrm{T}}}(\mathbf{V}_j^{(1)}\mathbf{V}_j^{(1)^{\mathrm{T}}})^{-1}\boldsymbol{\delta}_j^{(1)}$, we satisfy these requirement by choosing the canonical $\boldsymbol{\delta}_j^{(2)}=\mathbf{V}_j^{(2)}\mathbf{c}_j^*$, for which $\boldsymbol{\delta}_j$ is of minimum norm.

3. Robust *T*-optimality

In this section we suppose that both densities f_0 and f_1 are normal, with common variance σ^2. Then the minimization problem (7) becomes that of determining

$$(\mathbf{c}_0^*,\mathbf{c}_1^*)=\arg\min_{\mathbf{c}_0,\mathbf{c}_1}\{\mathcal{D}(\boldsymbol{\xi}|\mathbf{c}_0,\mathbf{c}_1)\},\qquad \text{subject to } \|\mathbf{c}_j\|\leqslant\tau_j \text{ for } j=0,1, \tag{10}$$

where

$$\mathcal{D}(\boldsymbol{\xi}|\mathbf{c}_0,\mathbf{c}_1)\overset{\text{def}}{=}\frac{1}{2\sigma^2}\sum_{i=1}^{N}[\eta_1(\mathbf{x}_i|\boldsymbol{\theta}_1)+\mathbf{v}_{1,i}^{\mathrm{T}}\mathbf{c}_1-\{\eta_0(\mathbf{x}_i|\boldsymbol{\theta}_0)+\mathbf{v}_{0,i}^{\mathrm{T}}\mathbf{c}_0\}]^2\,\xi_i.$$

Set $p=p_0+p_1$, let $\lambda_0>0$ and $\lambda_1>0$ be Lagrange multipliers and define a matrix $\mathbf{D}_{\boldsymbol{\xi}}=\mathrm{diag}(\xi_1,\ldots,\xi_N)$. Define also

$$\mathbf{V}=(\mathbf{V}_0\vdots\mathbf{V}_1), N\times 2N-p,$$

$$\boldsymbol{\Lambda}=\begin{pmatrix}\lambda_0\mathbf{I}_{N-p_0} & \mathbf{0}\\ \mathbf{0} & \lambda_1\mathbf{I}_{N-p_1}\end{pmatrix}, 2N-p\times 2N-p,$$

$$\mathbf{P}=\mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^{\mathrm{T}}=\frac{\mathbf{V}_0\mathbf{V}_0^{\mathrm{T}}}{\lambda_0}+\frac{\mathbf{V}_1\mathbf{V}_1^{\mathrm{T}}}{\lambda_1}, N\times N,$$

$$\mathbf{c}=\begin{pmatrix}\mathbf{c}_0\\ -\mathbf{c}_1\end{pmatrix}, 2N-p\times 1.$$

Theorem 2. With notation as above, the solution to problem (10) is

$$\mathbf{c}^*=\boldsymbol{\Lambda}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{D}_{\boldsymbol{\xi}}^{1/2}(\mathbf{I}_N+\mathbf{D}_{\boldsymbol{\xi}}^{1/2}\mathbf{P}\mathbf{D}_{\boldsymbol{\xi}}^{1/2})^{-1}\mathbf{D}_{\boldsymbol{\xi}}^{1/2}\boldsymbol{\eta}_d. \tag{11}$$

The minimum divergence is

$$\mathcal{D}(\boldsymbol{\xi}|\lambda_0,\lambda_1)\overset{\text{def}}{=}\mathcal{D}(\boldsymbol{\xi}|\mathbf{c}_0^*,\mathbf{c}_1^*)=\frac{1}{2\sigma^2}\|(\mathbf{I}_N+\mathbf{D}_{\boldsymbol{\xi}}^{1/2}\mathbf{P}\mathbf{D}_{\boldsymbol{\xi}}^{1/2})^{-1}\mathbf{D}_{\boldsymbol{\xi}}^{1/2}\boldsymbol{\eta}_d\|^2. \tag{12}$$

Remark 3. Theorem 2 does not give the values of the multipliers. Our approach is instead to parameterize the designs $\boldsymbol{\xi}_{\lambda_0,\lambda_1}^*=\arg\max\{\mathcal{D}(\boldsymbol{\xi}|\lambda_0,\lambda_1)\}$ by λ_0 and λ_1, and then to calculate $\tau_j=\|\mathbf{c}_j\|$ and to check that assumption (5) holds.

We obtain $\boldsymbol{\xi}_{\lambda_0,\lambda_1}^*$ by simulated annealing. The algorithm is described below. In the description we denote by q the minimum permissible number of design points; this will typically be the number of parameters in the larger of the two models but may be set to a larger value by the experimenter. We also denote by T a ‘temperature’ parameter.

Step 1: choose an initial design. The first time that this step is carried out the initial design consists of n points chosen at random from $\mathcal{S}$. In subsequent runs we 'restart', i.e. if step 2 has already been carried out at least once then in step 1 the initial design is the best design found up to this point. Compute $\mathcal{D}(\boldsymbol{\xi}|\lambda_0, \lambda_1)$ for the initial design.
Step 2: carry out the following, until L new designs have been tested without any improvement.

(a) Choose, at random, one of the points $\mathbf{x}_i \in \mathcal{S}$ for which $n_i > 0$; reduce this n_i by 1 and reassign the mass to one of the N points in $\mathcal{S}$, again chosen at random. This step is carried out q' times, with $q' \in \{1, 2, \ldots, q\}$ chosen at random. Resulting designs for which the number of design points drops below q are rejected immediately. Otherwise, let $\boldsymbol{\xi}'$ be the resulting design. Compute $\mathcal{D}(\boldsymbol{\xi}'|\lambda_0, \lambda_1)$.
(b) If $\nabla\mathcal{D} = \mathcal{D}(\boldsymbol{\xi}'|\lambda_0, \lambda_1) - \mathcal{D}(\boldsymbol{\xi}|\lambda_0, \lambda_1) > 0$ then the new design $\boldsymbol{\xi}'$ is accepted (and relabelled $\boldsymbol{\xi}$). Otherwise, it is accepted with probability $\exp(\nabla\mathcal{D}/T)$.

Step 3: lower the temperature, $T \leftarrow 0.95T$; repeat steps 1 and 2. Continue lowering the temperature until the fraction of improved designs found, at a fixed temperature, drops below $1/L$. In Fig. 2 later, a sequence of runs at a fixed temperature is termed a 'stage'. We initialize the temperature T at the lowest value T_0 for which about half of the new states are improvements on the current states.
Step 4: check assumption (5).

Remark 4. The annealing algorithm requires a great many evaluations of $\mathcal{D}$, each of which would seem to require the inversion of the, typically huge, matrix $\mathbf{I}_N + \mathbf{D}_{\xi}^{1/2}\mathbf{P}\mathbf{D}_{\xi}^{1/2}$. Fortunately, a substantial numerical simplification is possible. Suppose that $(i) = (i_1, \ldots, i_s)$ is the set of indices of non-zero ξ_i. Standard manipulations give

$$\mathcal{D}(\boldsymbol{\xi}|\lambda_0, \lambda_1) = \frac{1}{2\sigma^2} \|(\mathbf{I}_s + \mathbf{D}_+^{1/2}\mathbf{P}_{(i,i)}\mathbf{D}_+^{1/2})^{-1}\mathbf{D}_+^{1/2}\boldsymbol{\eta}_{d(i)}\|^2$$

where $\mathbf{P}_{(i,i)}$ denotes the $s \times s$ matrix that is formed from rows (i) and columns (i) of $\mathbf{P}$, $\mathbf{D}_+$ is the diagonal matrix with diagonal $(\xi_{i_1}, \ldots, \xi_{i_s})$ and $\boldsymbol{\eta}_{d(i)}$ contains the elements $\eta_{1,i_k} - \eta_{0,i_k}$. This requires only the inversion of a matrix of order $s \leqslant n$. Similarly,

$$\mathbf{c}^* = \begin{pmatrix} \mathbf{V}_{0(i,:)}^{\mathrm{T}}/\lambda_0 \\ \mathbf{V}_{1(i,:)}^{\mathrm{T}}/\lambda_1 \end{pmatrix} \mathbf{D}_+^{1/2}(\mathbf{I}_s + \mathbf{D}_+^{1/2}\mathbf{P}_{(i,i)}\mathbf{D}_+^{1/2})^{-1}\mathbf{D}_+^{1/2}\boldsymbol{\eta}_{d(i)},$$

where $\mathbf{V}_{j(i,:)}$, $s \times (N - p_j)$, consists of rows (i) of $\mathbf{V}_j$. Note also that $\mathbf{P}$ need be computed only once.

3.1. Example 1

Here we consider the Michaelis–Menten and exponential response models

$$\mu_0(x) \approx \eta_0(x|\boldsymbol{\theta}_0) = \frac{V_0 x}{K_0 + x},$$

with $\boldsymbol{\theta}_0 = (V_0, K_0)^{\mathrm{T}}$ and

$$\mu_1(x) \approx \eta_1(x|\boldsymbol{\theta}_1) = V_1\{1 - \exp(-K_1 x)\},$$

with $\boldsymbol{\theta}_1 = (V_1, K_1)^{\mathrm{T}}$.

In our treatment we take, in each case, $x \in \mathcal{S} = 0.1, 0.2, 0.3, \ldots, 5$; thus $N = 50$. We choose designs of size $n = 20$. First consider the case $\lambda_0 = \lambda_1 = \infty$, for which $\mathbf{P} = \mathbf{0}_{N\times N}$ and $\tau_0 = \tau_1 = 0$.

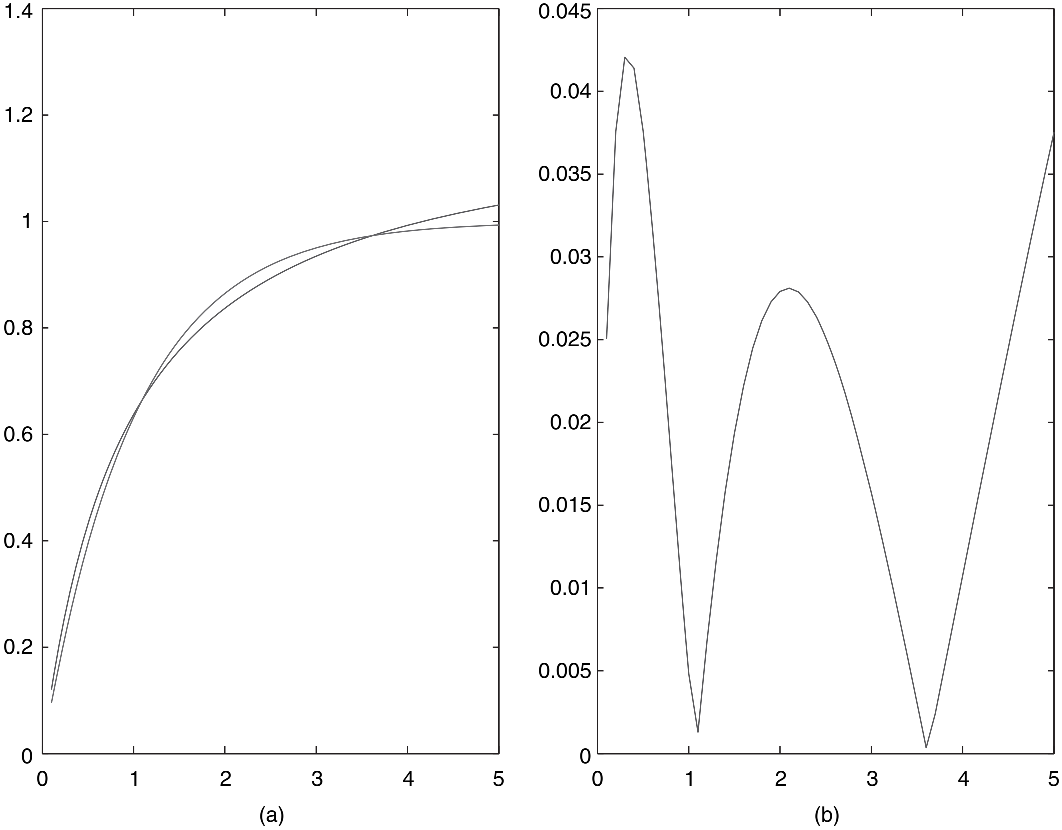


Fig. 1. (a) Least favourable means $\mu_j(x) = \eta_j(x|\boldsymbol{\theta}_j)$ and (b) $|\mu_1(x) - \mu_0(x)|$ for example 1, when $\tau_1 = \tau_0 = 0$

In this case no robustness against model misspecification is sought. We take $\eta_1(x|\boldsymbol{\theta}_1 = (1,1)^{\mathrm{T}})$ as the working response. Then, from definition (3), $\boldsymbol{\theta}_0 = (1.22, 0.91)^{\mathrm{T}}$: Fig. 1.

Although one-point designs make little sense from a practical standpoint, especially when parameters must be estimated, they do furnish an upper bound to the power in the ideal framework of theorem 1, where the parameters are assumed known. If such designs were to be permitted then the optimal design, minimizing problem (10) with $\mathbf{c}_0 = \mathbf{c}_1 = \mathbf{0}$, would place all mass at

$$\arg\max_{x \in \mathcal{S}} |\eta_1(x|\boldsymbol{\theta}_1) - \eta_0(x|\boldsymbol{\theta}_0)| = 0.3.$$

Using this one-point design the powers (2) of level $\alpha = 0.1$ tests against $\mu_1(\cdot)$, evaluated at a range of values of σ^2, are

$$\begin{pmatrix} \sigma^2: & 1 & 0.5 & 0.1 & 0.01 \\ \beta: & 0.14 & 0.15 & 0.25 & 0.73 \end{pmatrix}. \tag{13}$$

These powers necessarily agree with

$$\beta^* = \Phi\{\surd(2nD^*) - z_\alpha\}, \qquad \text{for } D^* = \max_{\mathbf{x} \in \mathcal{S}} [\mathcal{I}\{\eta_0(\mathbf{x}|\boldsymbol{\theta}_0), \eta_1(\mathbf{x}|\boldsymbol{\theta}_1)\},$$

a benchmark with which we shall compare other designs in this and subsequent examples. Suppose now that the experimenter desires at least $q = 6$ distinct design points. In our application of the annealing algorithm the initial temperature is $T_0 = 0.0002$, and $L = 500$.

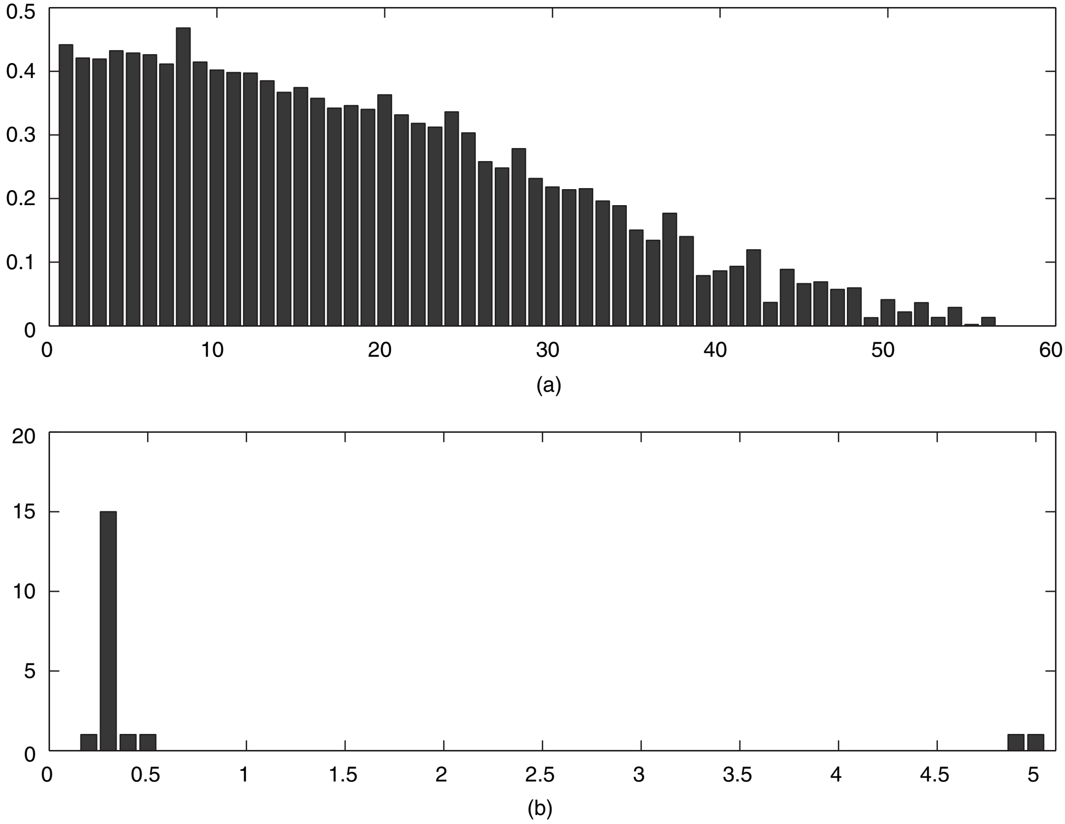


Fig. 2. Output for example 1 when $q = 6$: (a) improvement probability (the fraction of improved designs found in a stage) *versus* stage; (b) T-optimal design when both responses are correctly specified ($\tau_1 = \tau_0 = 0$)

The output is illustrated in Fig. 2, and results in the design with support points and relative frequencies

$$\xi = \begin{pmatrix} 0.2 & 0.3 & 0.4 & 0.5 & 4.9 & 5.0 \\ 0.05 & 0.70 & 0.10 & 0.05 & 0.05 & 0.05 \end{pmatrix}.$$

There is a negligible decline in the powers:

$$\begin{pmatrix} \sigma^2: & 1 & 0.5 & 0.1 & 0.01 \\ \beta: & 0.14 & 0.15 & 0.24 & 0.71 \end{pmatrix}.$$

The other designs of this section use $q = 2$, although the robustness requirements typically result in designs with more than two points of support. For smaller values of λ_0 and λ_1, corresponding to larger values of τ_0 and τ_1, the mass at $x = 0.3$ is reassigned to nearby but distinct points, and to points near $x = 5$. See Fig. 3 for

(a) $\lambda_0 = \lambda_1 = 1$, corresponding to $\tau_0 = 0.0099$ and $\tau_1 = 0.0095$, and
(b) $\lambda_0 = \lambda_1 = 5$, corresponding to $\tau_0 = 0.0033$ and $\tau_1 = 0.0031$.

The designs, and powers (2) of level $\alpha = 0.1$ tests against the least favourable alternatives, evaluated at a range of values of σ^2, are given below. We note the deterioration (from those at expression (13)) in these powers which is caused by the contamination of the responses:

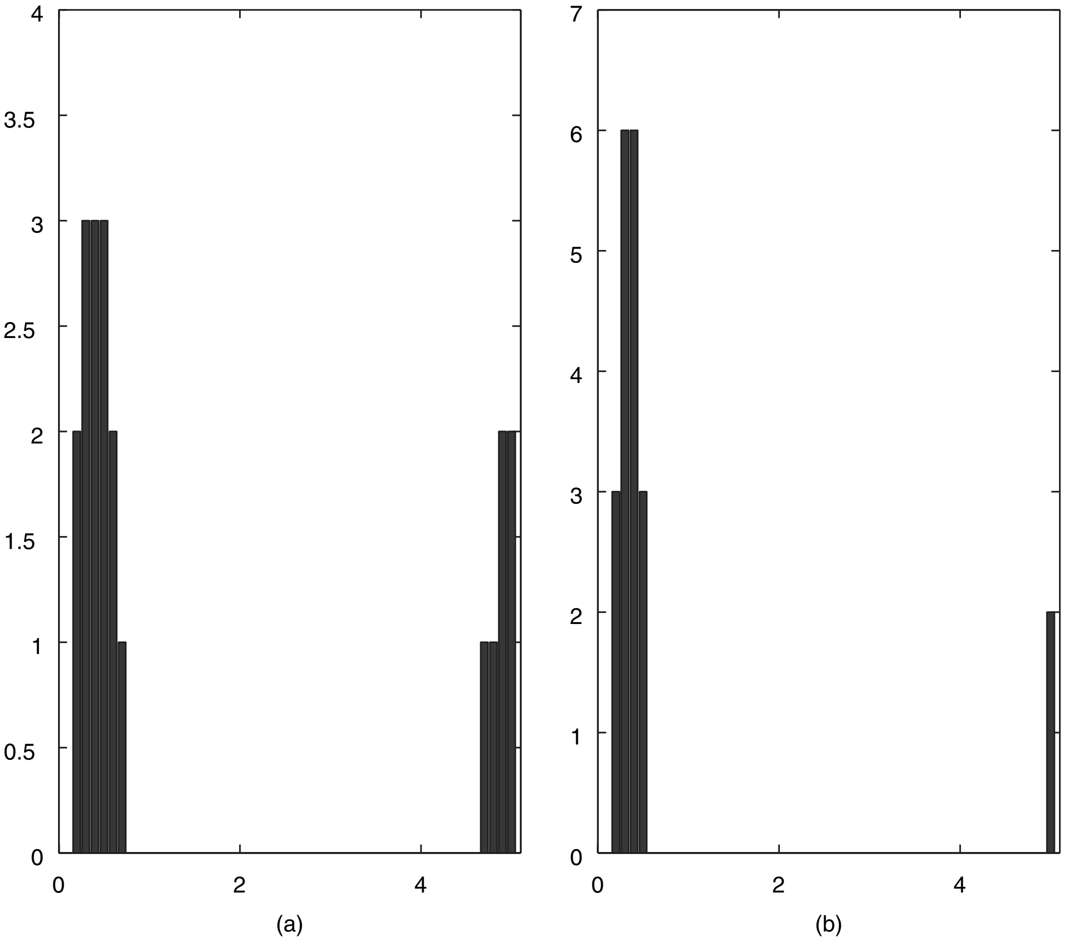


Fig. 3. Robust *T*-optimal designs for example 1: (a) $\lambda_0 = \lambda_1 = 1$; (b) $\lambda_0 = \lambda_1 = 5$

$$\text{(a) } \xi = \begin{pmatrix} 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 4.7 & 4.8 & 4.9 & 5.0 \\ 0.1 & 0.15 & 0.15 & 0.15 & 0.1 & 0.05 & 0.05 & 0.05 & 0.1 & 0.1 \end{pmatrix},$$

$$\begin{pmatrix} \sigma^2: & 1 & 0.5 & 0.1 & 0.01 \\ \beta: & 0.13 & 0.14 & 0.20 & 0.54 \end{pmatrix};$$

$$\text{(b) } \xi = \begin{pmatrix} 0.2 & 0.3 & 0.4 & 0.5 & 5.0 \\ 0.15 & 0.3 & 0.3 & 0.15 & 0.1 \end{pmatrix}, \begin{pmatrix} \sigma^2: & 1 & 0.5 & 0.1 & 0.01 \\ \beta: & 0.13 & 0.14 & 0.22 & 0.65 \end{pmatrix}.$$

3.2. Example 2

Here we continue comparing the same two models as in example 1, but the working response is now a linear response approximating $\eta_0(x|\boldsymbol{\theta})$ when $\boldsymbol{\theta} = (1, 1)^{\mathrm{T}}$, i.e. $E[Y|x] = 0.35 + 0.12x$ (Fig. 4). We then obtain, from equation (3),

$$\boldsymbol{\theta}_0 = (1.02, 1.13)^{\mathrm{T}},$$
$$\boldsymbol{\theta}_1 = (0.85, 0.73)^{\mathrm{T}}$$

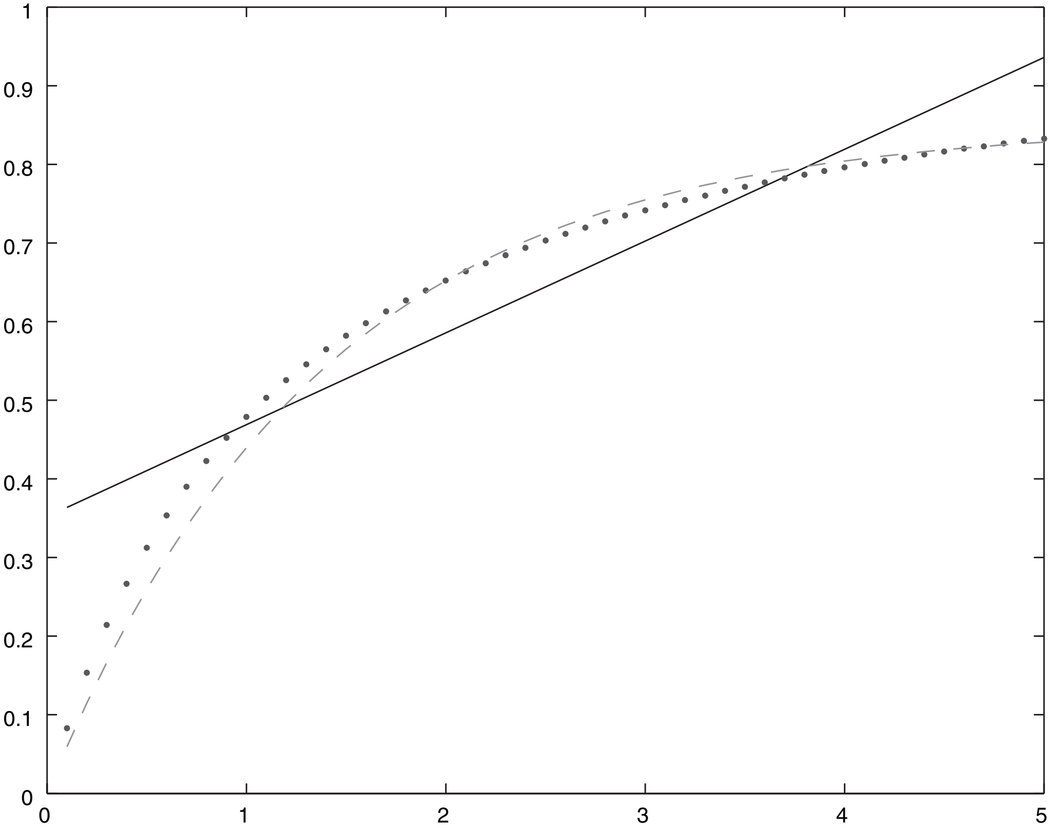


Fig. 4. Response functions for example 2: $E[Y|x] = 0.35 + 0.12x$ with closest Michaelis–Menten response $\eta_0\,(x|\boldsymbol{\theta}_0 = (1.02, 1.13)^{\mathrm{T}})$ $(\cdots\cdots)$ and closest exponential response $\eta_1\,(x|\boldsymbol{\theta}_1 = (0.85, 0.73)^{\mathrm{T}})$ $(-\ -\ -)$

(Fig. 5). With $n=20$ some optimal designs, with the powers against least favourable alternatives ($\alpha=0.1$), are as follows:

$$\lambda_0=\lambda_1=0.02,$$

$$\tau_0=0.0945,$$

$$\tau_1=0.0666,$$

$$\xi(\{0.2, 0.3, 0.4, \ldots, 2.1)\})=0.05,$$

$$\begin{pmatrix} \sigma^2: & 1 & 0.5 & 0.1 & 0.01 \\ \beta: & 0.11 & 0.12 & 0.15 & 0.31 \end{pmatrix};$$

$$\lambda_0=\lambda_1=5,$$

$$\tau_0=0.0039,$$

$$\tau_1=0.0035,$$

$$\xi=\begin{pmatrix} 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 \\ 0.05 & 0.20 & 0.25 & 0.25 & 0.15 & 0.10 \end{pmatrix},$$

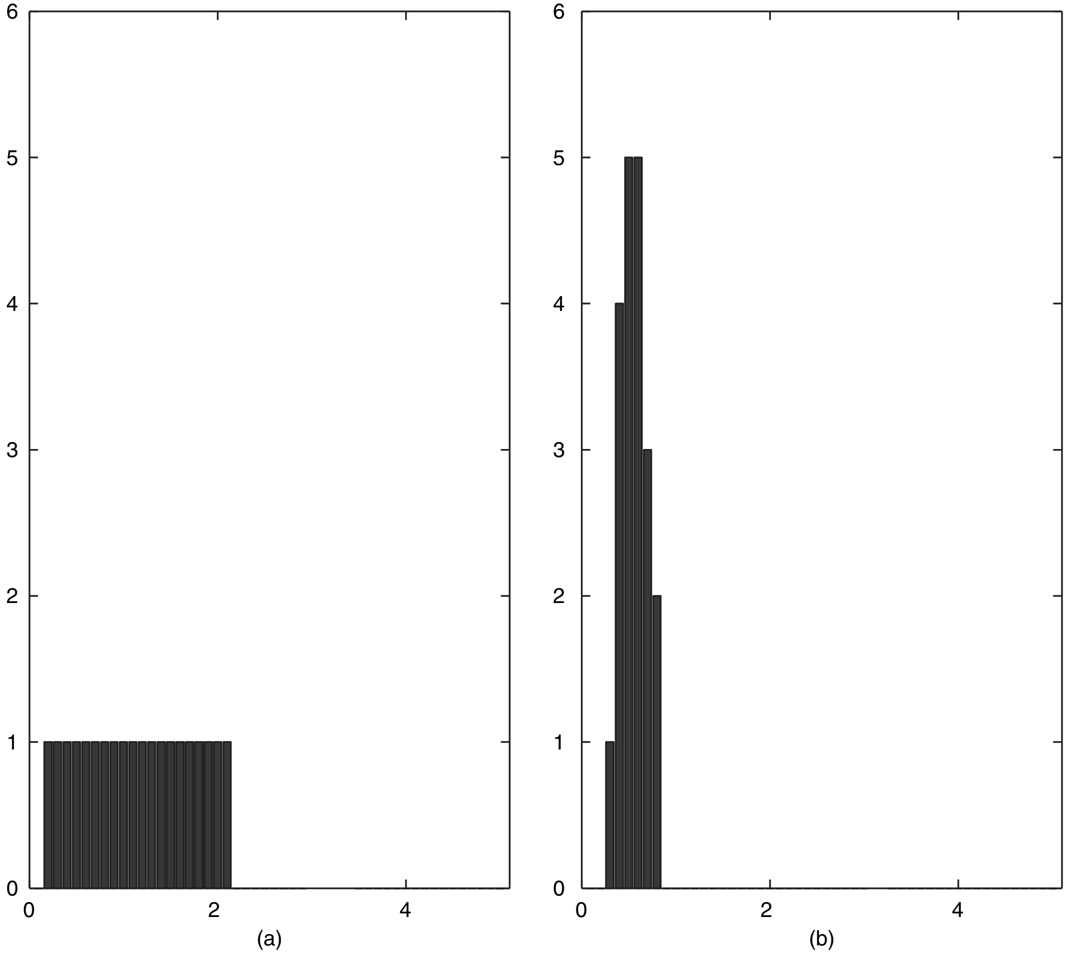


Fig. 5. Robust T-optimal designs for example 2: (a) $\lambda_0 = \lambda_1 = 0.02$; (b) $\lambda_0 = \lambda_1 = 5$

$$\begin{pmatrix} \sigma^2: & 1 & 0.5 & 0.1 & 0.01 \\ \beta: & 0.14 & 0.17 & 0.28 & 0.82 \end{pmatrix}.$$

The robust designs, as τ_0 and τ_1 increase, can be roughly described as being obtained by starting with the designs for smaller values of τ_0 and τ_1 and replacing replicates with unit frequencies at nearby but distinct points.

4. Robust Kullback–Leibler optimality

Here we consider the extension to non-normal target densities. To implement the development of this problem as it was outlined in Section 2, we use the following algorithm.

Step 1: choose an initial n-point design $\xi_{(0)}$.

For $k=0,1,\ldots$ to convergence carry out steps 2 and 3.

Step 2: put $\boldsymbol{\xi}=\boldsymbol{\xi}_{(k)}$ and do one of the following operations.

(a) For $n \leqslant \min(N-p_0, N-p_1)$, minimize $\mathcal{D}'(\boldsymbol{\xi}^{(1)}|\boldsymbol{\delta}_0^{(1)}, \boldsymbol{\delta}_1^{(1)})$, as at problem (9), obtaining minimizers $\boldsymbol{\delta}_j^{(1)}$. Set $\boldsymbol{\delta}_j^{(2)} = \mathbf{V}_j^{(2)}\mathbf{c}_j^*$ (recall remark 2).

(b) For $n > \min(N-p_0, N-p_1)$, minimize $\mathcal{D}(\boldsymbol{\xi}|\boldsymbol{\delta}_0, \boldsymbol{\delta}_1)$, at equation (8), with restrictions as in equation (7), obtaining minimizers $\boldsymbol{\delta}_j^*$.

In either case the minimization is carried out with the aid of MATLAB's constrained minimization routine `fmincon`.

Step 3: update the design. We have implemented two alternatives at this step, each a modification of procedures that were outlined in Cox and Reid (2000), page 178. Let $\boldsymbol{\delta}_0^*$ and $\boldsymbol{\delta}_1^*$ be the vectors of minimizers. In *option 1*, we temporarily drop the restriction that $\xi_i = n_i/n$ for integers n_i. With

$$\psi(\mathbf{x}_i; \boldsymbol{\xi}_{(k)}) \overset{\text{def}}{=} \mathcal{I}\{\eta_0(\mathbf{x}_i|\boldsymbol{\theta}_0) + \delta_0^*(\mathbf{x}_i), \eta_1(\mathbf{x}_i|\boldsymbol{\theta}_1) + \delta_1^*(\mathbf{x}_i)\},$$

our objective is to maximize

$$\mathcal{D}(\boldsymbol{\xi}|\boldsymbol{\delta}_0^*, \boldsymbol{\delta}_1^*) = \sum_{i=1}^{N} \psi(\mathbf{x}_i; \boldsymbol{\xi}_{(k)})\,\xi_i.$$

(a) Compute

$$\boldsymbol{\xi}_{(k+1)} = (1-p_k)\boldsymbol{\xi}_{(k)} + p_k\,\Delta(\mathbf{x}^*),$$

where $\Delta(\mathbf{x}^*)$ is point mass at

$$\mathbf{x}^* = \arg\max\{\psi(\mathbf{x}; \boldsymbol{\xi}_{(k)})\}$$

and $p_k = (k+1)^{-1}$.

(b) If step 3(a) results in a design calling for more than n observations then those frequencies at points of support corresponding to the smallest values of $\psi(\mathbf{x}; \cdot)$ are decreased accordingly.

After convergence is attained, the values $n\xi_i^*$ are rounded down to integers $\lfloor n\xi_i^* \rfloor$. The excess $n - \Sigma\lfloor n\xi_i^* \rfloor$ observations are then assigned to those $\mathbf{x}_i$ at which $\psi(\mathbf{x}_i; \boldsymbol{\xi}^*)$ is largest. *Option 2* is somewhat simpler. With $\psi(\mathbf{x}_i; \boldsymbol{\xi}_{(k)})$ as above, the next design $\boldsymbol{\xi}_{(k+1)}$ is obtained from $\boldsymbol{\xi}_{(k)}$ by decreasing by 1 the number of observations to be made at

$$\mathbf{x}_* = \arg\min\{\psi(\mathbf{x}_i; \boldsymbol{\xi}_{(k)})\},$$

with the minimum taken only over those $\mathbf{x}_i$ for which $\boldsymbol{\xi}_{(k),i} > 0$. Then an additional mass of n^{-1} is assigned to the point $\mathbf{x}^*$ that was defined above.

In our simulations both options often yielded the same or similar designs. Other times the performance of a design that was obtained by using option 1 would deteriorate when the rounding procedure was performed, and then option 2 would yield a superior design. Given that option 2 is also simpler and faster, it is to be preferred.

Some researchers—Atkinson and Fedorov (1975a) and López-Fidalgo *et al.* (2007)—have exploited convex design theory to prove convergence of algorithms, such as that utilizing option 1, to an optimal design $\boldsymbol{\xi}^*$ with the property that it places all mass at those points where $\psi(\mathbf{x}; \boldsymbol{\xi}^*)$ attains its maximum value. In our case the various restrictions that are imposed on the designs render the class of such designs non-convex, so analogous results cannot be expected. The extent to which $\boldsymbol{\xi}^*$ nonetheless approximately attains this property (by using option 2) can be seen in Figs 6 and 7.

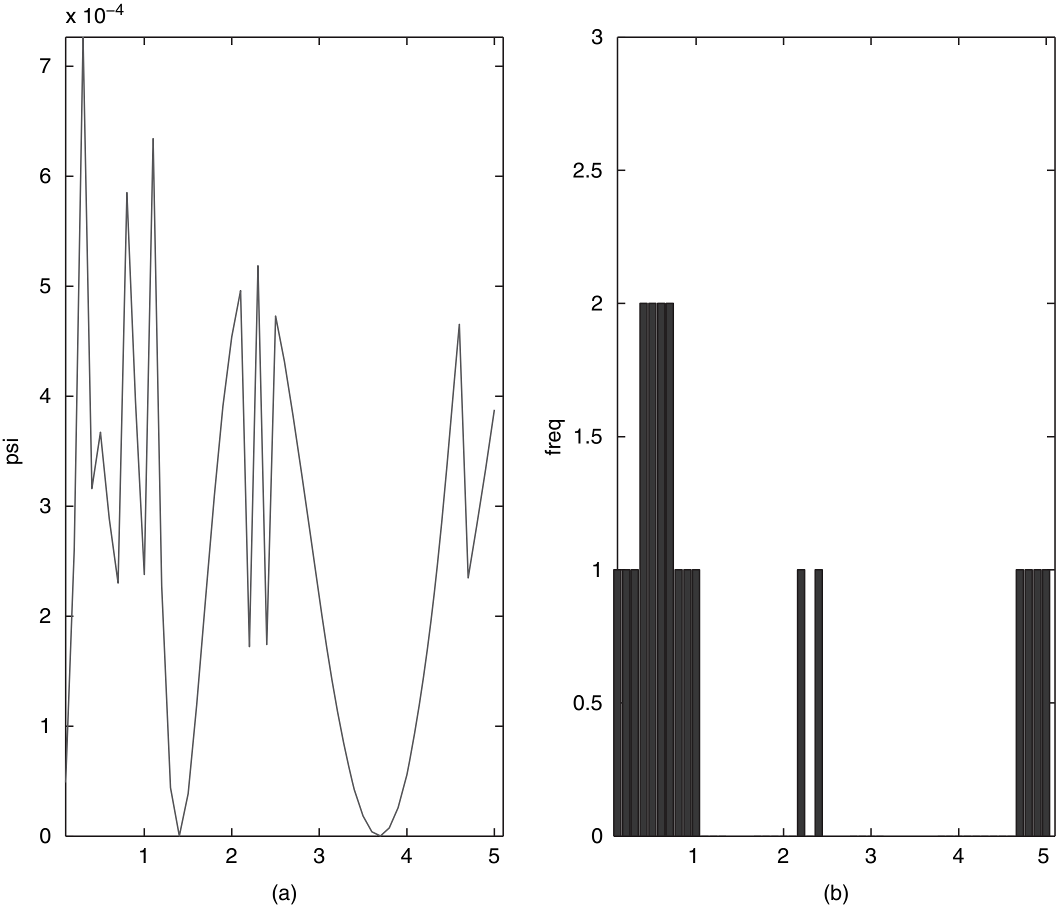


Fig. 6. Design for example 3—$E[Y|x] = \eta_1(x|\boldsymbol{\theta}_1)$ and $\tau_0 = \tau_1 = 0.05$: (a) $\psi(x;\xi^*)$; (b) design frequencies

4.1. Michaelis–Menten and exponential responses; log-normal densities

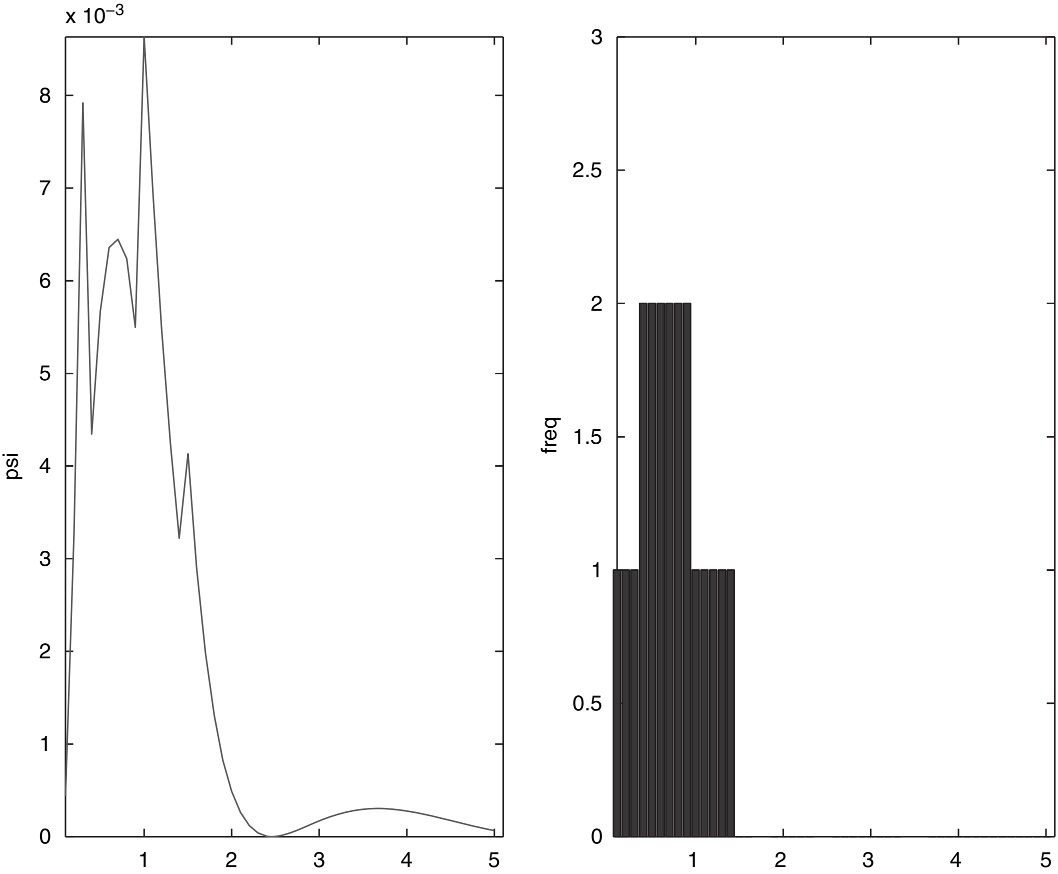
Recall the (approximate) Michaelis–Menten and exponential response models of example 1. Suppose that under each model the observations are log-normal, i.e. $\log Y \sim N\{\alpha_j(x), \sigma_j^2(x)\}$ with

$$E_{\text{Model}j}[Y|x] = \mu_j(x) = \exp(\sigma_j^2/2 + \alpha_j),$$
$$\text{var}_{\text{Model}j}(Y|x) = v_j^2(x) = \mu_j^2(x)\{\exp(\sigma_j^2) - 1\},$$

and so

$$\alpha_j(x) = \log\left[\frac{\mu_j(x)}{\surd\{1 + v_j^2(x)/\mu_j^2(x)\}}\right],$$
$$\sigma_j^2(x) = \log\left\{1 + \frac{v_j^2(x)}{\mu_j^2(x)}\right\}.$$

The density of Y is, in terms of the $N(0,1)$ density $\phi(\cdot)$,



(a) (b)

Fig. 7. Design for example 3—$E[Y|x] = 0.35 + 0.12x$ and $\tau_0 = \tau_1 = 0.05$: (a) $\psi(x; \xi^*)$; (b) design frequencies

$$f_j(y|x_i, \mu_j) = \phi\left\{\frac{\log(y) - \alpha_j}{\sigma_j}\right\} \frac{1}{\sigma_j y} I(y > 0),$$

with

$$\mathcal{I}(\mu_0, \mu_1) = \log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) + \frac{1}{2\sigma_0^2}\{\sigma_1^2 - \sigma_0^2 + (\alpha_1 - \alpha_0)^2\}.$$

At this point we must specify the variance functions $v_j^2(x)$. López-Fidalgo *et al.* (2007) considered several choices. For instance we can assume homoscedastic models with $v_j^2(x) \equiv 1$; in this case we write the divergence as

$$\mathcal{I}_{\text{hom}}(\mu_0, \mu_1) = \log\left[\frac{\log\{1 + \mu_0^{-2}(x)\}}{\log\{1 + \mu_1^{-2}(x)\}}\right] + \frac{\log[\{1 + \mu_1^{-2}(x)\}/\{1 + \mu_0^{-2}(x)\}] + \log^2[\{\mu_1(x)/\mu_0(x)\}\surd\{1 + \mu_0^{-2}(x)\}/\{1 + \mu_1^{-2}(x)\}]}{2\,\log\{1 + \mu_0^{-2}(x)\}}.$$

We might instead assume a constant coefficient of variation:

$$\frac{v_j(x)}{\mu_j(x)} = \surd\{\exp(\sigma_j^2) - 1\} \overset{\text{def}}{=} \text{cv};$$

under this assumption the divergence is

$$\mathcal{I}_{\text{cv}}(\mu_0, \mu_1) = \frac{[\log\{\mu_1(x)\} - \log\{\mu_0(x)\}]^2}{2\log(1+\text{cv}^2)}.$$

4.1.1. Example 3

We consider the divergence $\mathcal{I}_{\text{cv}}(\mu_0, \mu_1)$ and obtain designs, of size $n=20$, for discriminating between the Michaelis–Menten and exponential alternatives of examples 1 and 2. The design space $\mathcal{S}$, working response $E[Y|x]$ and corresponding parameter vectors $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are as in those examples. See Fig. 6 for $\tau_0 = \tau_1 = 0.05$ and $E[Y|x] = \eta_1(x|\boldsymbol{\theta}_1)$. With $E[Y|x]$ as in example 2, see Fig. 7. The powers (2) of level $\alpha = 0.1$ tests against the least favourable alternatives, and the benchmark values (13), evaluated at a range of values of the coefficient of variation, are as follows:

(a) for $E[Y|x] = \eta_1(x|\boldsymbol{\theta}_1)$

$$\begin{pmatrix} \text{cv}^2: & 1 & 0.5 & 0.1 & 0.01 \\ \beta: & 0.12 & 0.12 & 0.15 & 0.31 \\ \beta^*: & 0.49 & 0.64 & 0.98 & 1.00 \end{pmatrix};$$

(b) for $E[Y|x] = 0.35 + 0.12x$

$$\begin{pmatrix} \text{cv}^2: & 1 & 0.5 & 0.1 & 0.01 \\ \beta: & 0.19 & 0.22 & 0.42 & 0.98 \\ \beta^*: & 0.69 & 0.85 & 1.00 & 1.00 \end{pmatrix}.$$

5. Sequential discrimination designs

Here we propose the following procedure. It can be used to construct static designs—one design point at a time—by using as input an assumed response $E[Y|\mathbf{x}]$. We call these *stepwise* designs. Alternatively it can be applied sequentially 'in the field', with the $\boldsymbol{\theta}_j$ replaced by estimates at each stage.

Step 1: choose a small initial design ξ.
Step 2: carry out step 2 of Section 4. For a sequential design the parameters that are required in the evaluation of $\mu_j^*(\mathbf{x}_i) = \eta_j(\mathbf{x}_i|\boldsymbol{\theta}_j) + \delta_j^*(\mathbf{x}_i)$ are replaced by estimates $\hat{\boldsymbol{\theta}}_j$, rather than being defined by equation (3).
Step 3: make the next observation at $\mathbf{x}_{\text{new}} = \arg\max_{\mathbf{x}\in\mathcal{S}}[\mathcal{I}\{\mu_0^*(\mathbf{x}), \mu_1^*(\mathbf{x})\}]$.

Steps 2 and 3 are repeated until an n-point design is obtained.

5.1. Example 4

We consider the divergence $\mathcal{I}_{\text{cv}}(\mu_0, \mu_1)$ and obtain stepwise static designs, of sizes $n=20$, for discriminating between the Michaelis-Menten and exponential alternatives of example 2. The design space $\mathcal{S}$, working response $E[Y|x]$ and corresponding parameter vectors $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are as there. The starting design places one observation at each of $x=0.1$ and $x=1.4$—the extremes of the design of example 3. The remaining points are generated by the algorithm that was described

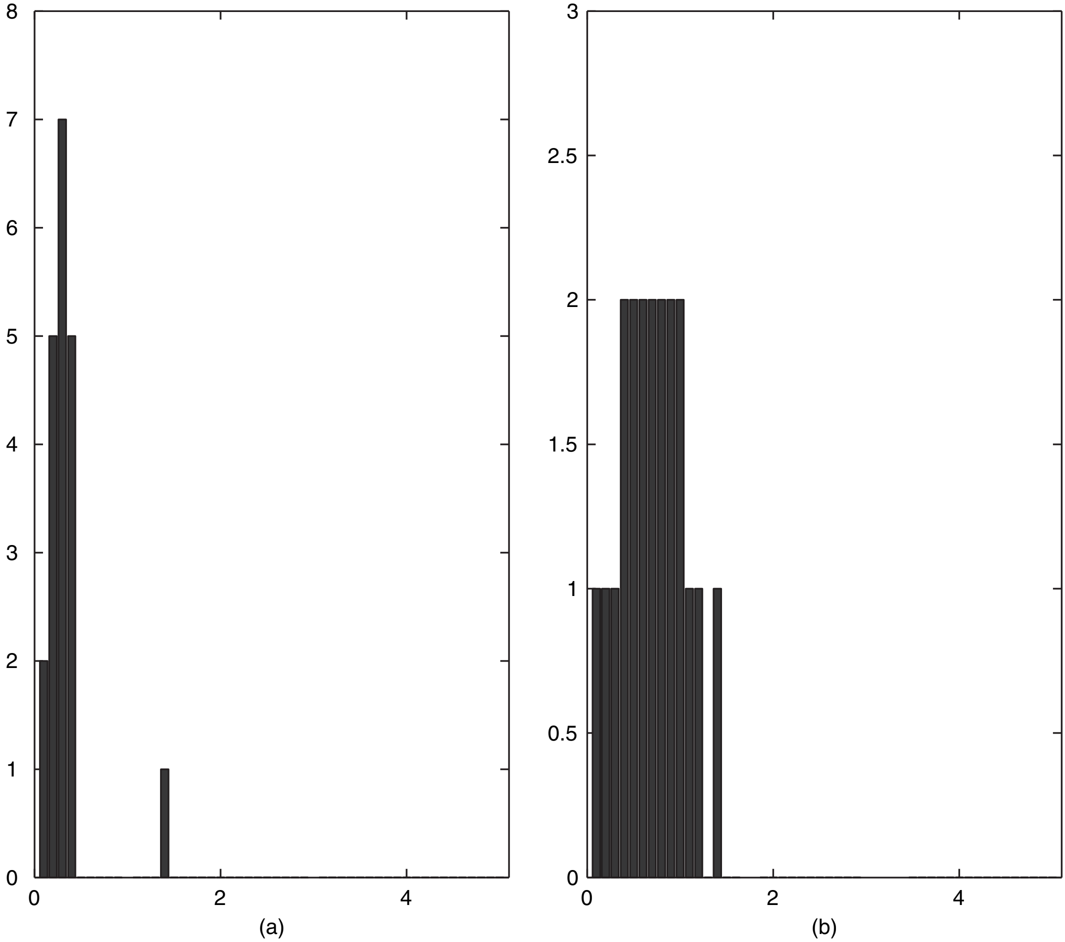


Fig. 8. 'Stepwise' design points and frequencies for example 4: (a) $\tau_0 = \tau_1 = 0.01$; (b) $\tau_0 = \tau_1 = 0.05$

above, resulting in the designs that are illustrated in Fig. 8 for $\tau_0 = \tau_1 = (0.01, 0.05)$ respectively. Compare Fig. 8(b) with Fig. 7—the designs use the same value of $\tau_0 = \tau_1$ and are almost identical. We found this as well by using $\tau_0 = \tau_1 = 0.01$. The powers (2) of level $\alpha = 0.1$ tests against the least favourable alternatives, evaluated at a range of values of the coefficient of variation, are ($\beta^* = (0.69, 0.85, 1.00, 1.00)$):

(a) $\tau_0 = \tau_1 = 0.01$,

$$\begin{pmatrix} \mathrm{cv}^2: & 1 & 0.5 & 0.1 & 0.01 \\ \beta: & 0.40 & 0.52 & 0.93 & 1.00 \end{pmatrix};$$

(b) $\tau_0 = \tau_1 = 0.05$,

$$\begin{pmatrix} \mathrm{cv}^2: & 1 & 0.5 & 0.1 & 0.01 \\ \beta: & 0.19 & 0.22 & 0.42 & 0.98 \end{pmatrix}.$$

5.2. Example 5

Sequential designs, in the same contexts as for example 4 and with $\tau_0 = \tau_1 = 0.01$, were constructed (Fig. 9). Sampling was done from a log-normal population with mean $\mu(x) = 0.35 +$

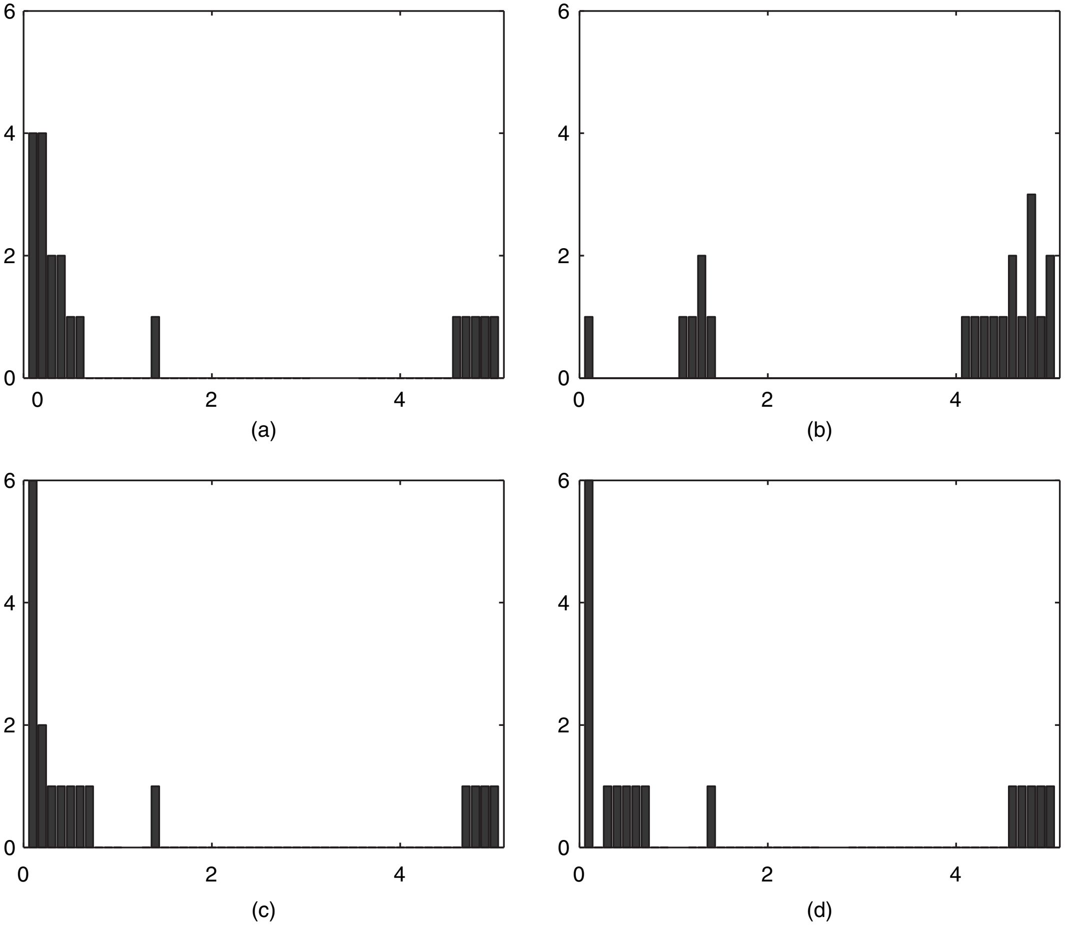


Fig. 9. Sequentially obtained design points and frequencies for example 5, $n=20$ and $\tau_0=\tau_1=0.05$: (a) $\mathrm{cv}^2=1$; (b) $\mathrm{cv}^2=0.5$; (c) $\mathrm{cv}^2=0.1$; (d) $\mathrm{cv}^2=0.01$

$0.12x$ and various coefficients of variation. Of course the results of this example are subject to sampling variation.

6. Robust, sequential integrated Kullback–Leibler optimal designs for discrimination and estimation or prediction

The designs that have been presented so far, although optimal for discrimination, are generally poor for estimation or prediction from the chosen model. Without robustness considerations this was noted by Hill *et al.* (1968), who addressed these concerns by proposing a sequential method to maximize a convex combination of the discriminatory power and a measure of estimation efficiency such as a weighted average of the determinants of the moment matrices in the two models. Atkinson (2008) has proposed the maximization of a convex combination of the logarithm of $\mathcal{D}(\boldsymbol{\xi}|\mathbf{0},\mathbf{0})$, as at problem (10), and the logarithm of the determinant of the moment matrix under model 0. This criterion, combining as it does classical D- and T-optimality criteria, is termed D–T-optimality.

Here we propose a *robust integrated Kullback–Leibler (IKL) optimality* criterion, in which we aim for optimization of the discriminatory power, through maximization of the KL discrepancy (KL optimality) and, simultaneously, minimization of the integrated mean-squared error (IMSE) of the predictions (I-optimality). The underlying idea is that, as the experiment evolves, evidence in favour of one of the two models will accrue. As it does, emphasis should move from model discrimination towards efficient prediction from the model favoured.

We propose a measure $\Delta_n^{(1)}(\mathbf{x};\delta_{n+1,0},\delta_{n+1,1})$ of the increased discriminatory power, due to the addition of $\mathbf{x}$ to a design $\xi^{(n)}$ calling for n observations, and a measure $\Delta_{n,j}^{(2)}(\mathbf{x};\delta_{n+1,j})$ of the drop in IMSE, when predictions are made from model j, due to this addition. We then propose to choose

$$\mathbf{x}_{\text{new}} = \arg\max_{\mathbf{x}\in\mathcal{S}} \min_{\delta_{n+1,0},\delta_{n+1,1}} \{\Delta_n(\mathbf{x};\delta_{n+1,0},\delta_{n+1,1})\}, \tag{14}$$

where Δ_n is the adaptively weighted average

$$\Delta_n = (1-\kappa_n)\Delta_n^{(1)} + \kappa_n \Delta_{n,j_n}^{(2)}.$$

Here

$$\kappa_n = \surd|2\pi(r_n) - 1|,$$

$$j_n = \begin{cases} 0, & \pi(r_n) < \frac{1}{2}, \\ 1, & \pi(r_n) > \frac{1}{2}, \end{cases}$$

r_n is the observed value of the test statistic R after n observations and

$$\pi(r) = \Phi\left\{\frac{r}{\surd(4|r|)}\right\} = \Phi\left\{\frac{\operatorname{sgn}(r)\surd|r|}{2}\right\}.$$

Thus if r_n is negative but large in magnitude, the next design point serves primarily to decrease the IMSE in model 0. If r_n is large and positive then the next observation serves primarily to decrease the IMSE in model 1. If r_n is near 0 then $\kappa_n \approx 0$ and the discrimination problem continues. For intermediary values of r_n the given measure reflects the p-value in a continuous fashion, symmetric in the two hypotheses. In the computation of R the various parameters are replaced by estimates that are based on the current sample. The form of $\pi(r)$ was motivated by theorem 1 together with the observation that $4|r|$ is a crude estimate of $4|E[R]| \sim 8nD$ (which is

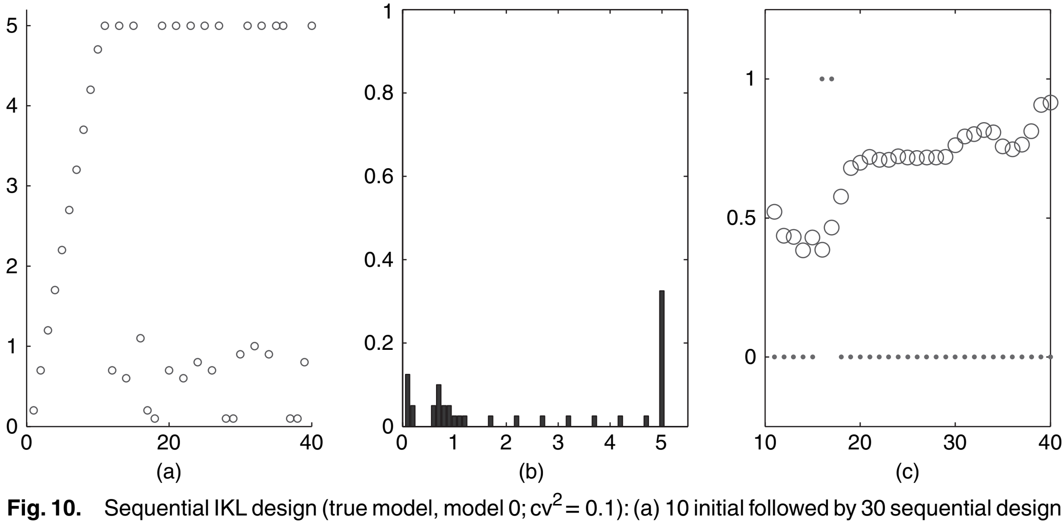


Fig. 10. Sequential IKL design (true model, model 0; cv^2 = 0.1): (a) 10 initial followed by 30 sequential design points; (b) final design; (c) sequential values of κ_n and j_n

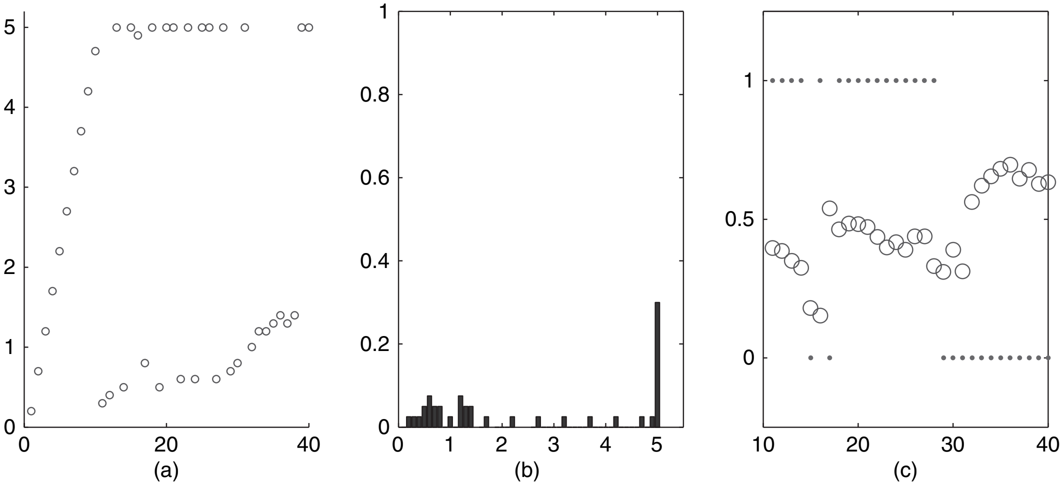


Fig. 11. Sequential IKL design (true model, model 0; $cv^2 = 1$): (a) 10 initial followed by 30 sequential design points; (b) final design; (c) sequential values of κ_n and j_n

less crude for large values of this expectation). The square root in the definition of κ_n is a tuning value chosen after studying the simulations.

The development of the measures $\Delta_n^{(1)}$ and $\Delta_n^{(2)}$ is carried out in Appendix A, where the numerical procedure is outlined as well.

6.1. Example 6

Sequential IKL designs were constructed as described above, using simulated data from log-normal densities with various coefficients of variation, $\mathcal{I} = \mathcal{I}_{\text{cv}}$, and one of the two mean structures of example 1. In each case an initial design with 10 approximately equally spaced design points was used, and a further 30 points were then chosen sequentially (Figs 10–13). In the captions,

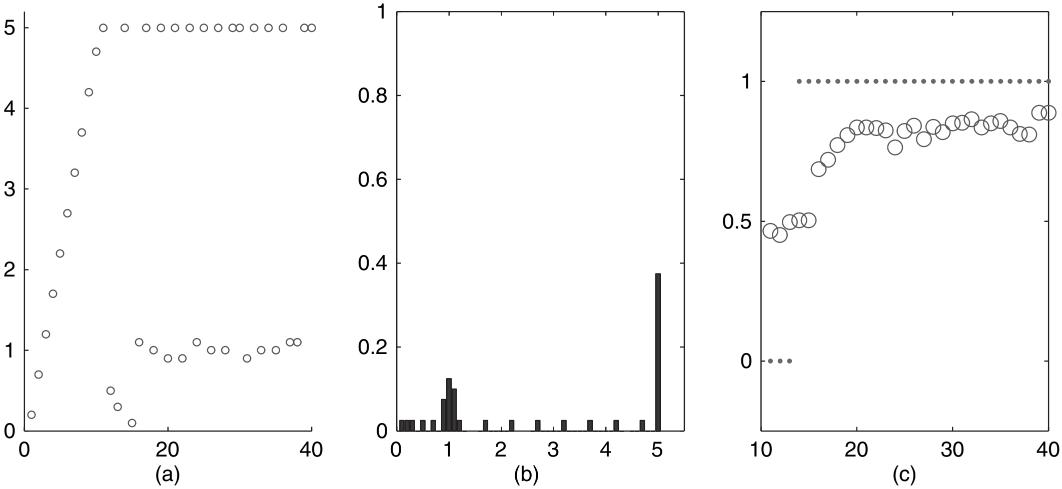


Fig. 12. Sequential IKL design (true model, model 1; $cv^2 = 0.1$): (a) 10 initial followed by 30 sequential design points; (b) final design; (c) sequential values of κ_n and j_n

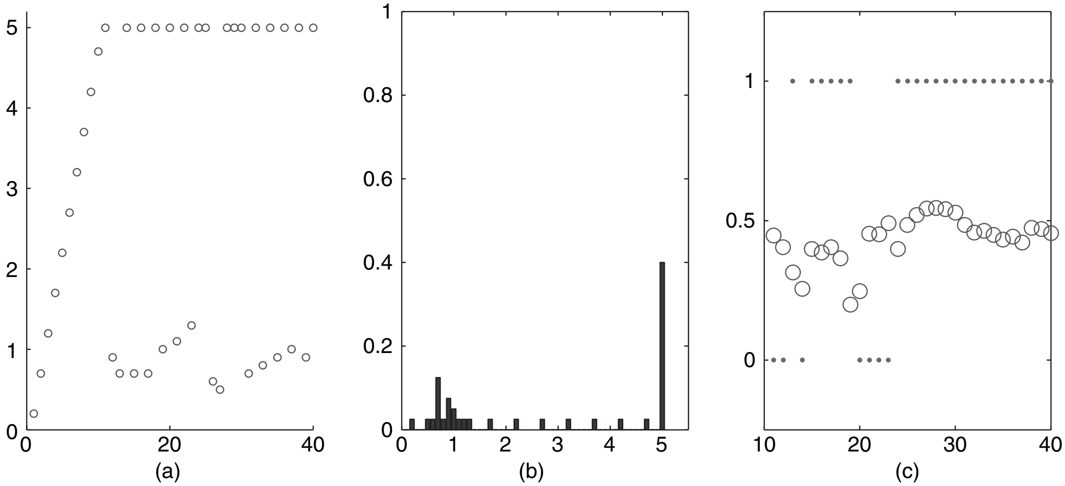


Fig. 13. Sequential IKL design (true model, model 1; $cv^2 = 1$): (a) 10 initial followed by 30 sequential design points; (b) final design; (c) sequential values of κ_n and j_n

'true model = 0' means that sampling was done by using $E[Y|x] = \eta_0\{x|(1,1)^{\mathrm{T}}\}$, and 'true model = 1' that $E[Y|x] = \eta_1\{x|(1,1)^{\mathrm{T}}\}$. In all cases we took $\tau_0 = \tau_1 = 0.01$. The resulting designs generally exhibit the properties that we might hope for and expect. With small coefficients of variation the correct model is identified quite early, after which κ_n approaches 1. Larger coefficients of variation might delay, but do not prevent, this identification. Changes in j_n often seem to be in response to jumps in κ_n. We also note that, although the final designs are quite similar, the paths that are taken differ markedly.

7. Summarizing remarks

We have derived methods of construction for experimental designs, to aid in the discrimination between regression models. The designs do so by maximizing the minimum KL divergence between two neighbourhoods of models. In the case that both models have normal densities, the minimization part of this procedure has been carried out analytically; the maximization part by simulated annealing. For non-normal models both parts require numerical approaches, and appropriate algorithms have been presented.

In the examples we have given static designs for pure discrimination problems. A somewhat arbitrary feature of the development has been the adoption of a 'working response' through which target parameters are defined. To the extent that this arbitrariness is a problem, the difficulty vanishes if we take the sequential approach of Sections 5 and 6. In particular, the IKL designs that were introduced in Section 6, to serve dual purposes of discrimination and estimation or prediction, seem to furnish an attractive combination of adaptability and robustness, while being easily constructed and implemented.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada. The author thanks the reviewers for their insightful comments on an earlier version of this paper.

Appendix A: Derivations

A.1. Proof of proposition 1

The constraints $\mathbf{U}_j^{\mathrm{T}}\boldsymbol{\delta}_j=\mathbf{0}$ lead, as in equation (7), to

$$\begin{pmatrix}\boldsymbol{\delta}_j^{(1)}\\ \boldsymbol{\delta}_j^{(2)}\end{pmatrix}=\begin{pmatrix}\mathbf{V}_j^{(1)}\\ \mathbf{V}_j^{(2)}\end{pmatrix}\mathbf{c}_j,$$

and hence that $\boldsymbol{\delta}_j^{(1)}=\mathbf{V}_j^{(1)}\mathbf{c}_j$ for some $\mathbf{c}_j$ $(N-p_j\times 1)$ with $\|\mathbf{c}_j\|^2\leqslant\tau_j^2$. But, under the condition that $\mathbf{V}_j^{(1)}$ have full rank $s\leqslant N-p_j$, *any* $\boldsymbol{\delta}_j^{(1)}$ lies in the column space of $\mathbf{V}_j^{(1)}$ and we claim that the sets

$$\mathcal{W}_j=\{\boldsymbol{\delta}_j^{(1)}|\boldsymbol{\delta}_j^{(1)}=\mathbf{V}_j^{(1)}\mathbf{c}_j\text{ for some }\mathbf{c}_j\text{ with }\|\mathbf{c}_j\|^2\leqslant\tau_j^2\},$$
$$\mathcal{W}'_j=\{\boldsymbol{\delta}_j^{(1)}|\boldsymbol{\delta}_j^{(1)^{\mathrm{T}}}(\mathbf{V}_j^{(1)}\mathbf{V}_j^{(1)^{\mathrm{T}}})^{-1}\boldsymbol{\delta}_j^{(1)}\leqslant\tau_j^2\}$$

coincide.

To see that $\mathcal{W}_j=\mathcal{W}'_j$, first let $\boldsymbol{\delta}_j^{(1)}\in\mathcal{W}_j$. The set of solutions to $\boldsymbol{\delta}_j^{(1)}=\mathbf{V}_j^{(1)}\mathbf{c}_j$ is, with

$$\mathbf{c}_j^*\overset{\text{def}}{=}\mathbf{V}_j^{(1)^{\mathrm{T}}}(\mathbf{V}_j^{(1)}\mathbf{V}_j^{(1)^{\mathrm{T}}})^{-1}\boldsymbol{\delta}_j^{(1)},$$

given by $\{\mathbf{c}_j=\mathbf{c}_j^*+\mathbf{t}_j|\mathbf{V}_j^{(1)}\mathbf{t}_j=\mathbf{0}\}$. Any such solution entails $\mathbf{t}_j\perp\mathbf{c}_j^*$ and so

$$\|\mathbf{c}_j\|^2=\|\mathbf{c}_j^*\|^2+\|\mathbf{t}_j\|^2,\qquad \|\mathbf{c}_j^*\|^2=\boldsymbol{\delta}_j^{(1)^{\mathrm{T}}}(\mathbf{V}_j^{(1)}\mathbf{V}_j^{(1)^{\mathrm{T}}})^{-1}\boldsymbol{\delta}_j^{(1)}.$$

Thus if $\boldsymbol{\delta}_j^{(1)}\in\mathcal{W}_j$ then $\|\mathbf{c}_j^*\|^2\leqslant\|\mathbf{c}_j\|^2\leqslant\tau_j^2$, i.e. $\boldsymbol{\delta}_j^{(1)}\in\mathcal{W}'_j$. Conversely, let $\boldsymbol{\delta}_j^{(1)}\in\mathcal{W}'_j$. Put $\mathbf{c}_j=\mathbf{c}_j^*+\mathbf{t}_j$ for any $\mathbf{t}_j$ satisfying $\mathbf{V}_j^{(1)}\mathbf{t}_j=\mathbf{0}$, normed so that $\|\mathbf{t}_j\|^2\leqslant\tau_j^2-\|\mathbf{c}_j^*\|^2$. In particular, we can set $\mathbf{c}_j=\mathbf{c}_j^*$. Then $\boldsymbol{\delta}_j^{(1)}=\mathbf{V}_j^{(1)}\mathbf{c}_j$, and $\|\mathbf{c}_j\|^2=\|\mathbf{c}_j^*\|^2+\|\mathbf{t}_j\|^2\leqslant\tau_j^2$, i.e. $\boldsymbol{\delta}_j^{(1)}\in\mathcal{W}_j$.

A.2. Proof of theorem 2

First write

$$2\sigma^2\,\mathcal{D}(\boldsymbol{\xi}|\mathbf{c}_0,\mathbf{c}_1)=2\sigma^2\|\mathbf{D}_{\xi}^{1/2}(\boldsymbol{\eta}_d+\mathbf{V}_1\mathbf{c}_1-\mathbf{V}_0\mathbf{c}_0)\|^2=\|\mathbf{a}+\mathbf{B}_1\mathbf{c}_1-\mathbf{B}_0\mathbf{c}_0\|^2,\tag{15}$$

say, for $\mathbf{a}=\mathbf{D}_{\xi}^{1/2}\boldsymbol{\eta}_d$ and $\mathbf{B}_j=\mathbf{D}_{\xi}^{1/2}\mathbf{V}_j$. The minimum of equation (15) is the distance between two disjoint, closed sets—the translated ellipsoid $\{\mathbf{a}+\mathbf{B}_1\mathbf{c}_1|\|\mathbf{c}_1\|\leqslant\tau_1\}$ and the centred ellipsoid $\{\mathbf{B}_0\mathbf{c}_0|\|\mathbf{c}_0\|\leqslant\tau_0\}$. This minimum is attained on the boundaries of the two sets, i.e. $\|\mathbf{c}_j\|=\tau_j$, for each j.

To minimize equation (15) subject to $\|\mathbf{c}_j\|=\tau_j$, define a function

$$F(\mathbf{c}_0,\mathbf{c}_1;\lambda_1,\lambda_0)=\|\mathbf{a}+\mathbf{B}_1\mathbf{c}_1-\mathbf{B}_0\mathbf{c}_0\|^2+\lambda_0(\|\mathbf{c}_0\|^2-\tau_0^2)+\lambda_1(\|\mathbf{c}_1\|^2-\tau_1^2).$$

For fixed $\lambda_0^*,\lambda_1^*>0$ the function F is convex in $\mathbf{c}_0$ and $\mathbf{c}_1$, and so a critical point $(\mathbf{c}_0^*,\mathbf{c}_1^*,\lambda_0^*,\lambda_1^*)$ will furnish the desired minimum if λ_0^* and λ_1^* are chosen to satisfy the side-conditions. The first-order conditions are these side-conditions together with

$$\frac{\partial F}{\partial\mathbf{c}_0}=\mathbf{0}^{\mathrm{T}},1\times N-p_0,$$
$$\frac{\partial F}{\partial\mathbf{c}_1}=\mathbf{0}^{\mathrm{T}},1\times N-p_1.$$

These last two equations are

$$\begin{pmatrix}\mathbf{B}_0^{\mathrm{T}}\mathbf{B}_0+\lambda_0\mathbf{I}_{N-p_0} & \mathbf{B}_0^{\mathrm{T}}\mathbf{B}_1\\ \mathbf{B}_1^{\mathrm{T}}\mathbf{B}_0 & \mathbf{B}_1^{\mathrm{T}}\mathbf{B}_1+\lambda_1\mathbf{I}_{N-p_1}\end{pmatrix}\begin{pmatrix}\mathbf{c}_0\\ -\mathbf{c}_1\end{pmatrix}=\begin{pmatrix}\mathbf{B}_0^{\mathrm{T}}\\ \mathbf{B}_1^{\mathrm{T}}\end{pmatrix}\mathbf{a}.\tag{16}$$

In the original notation equation (16) is

$$(\mathbf{V}^{\mathrm{T}}\mathbf{D}_{\xi}\mathbf{V}+\boldsymbol{\Lambda})\mathbf{c}=\mathbf{V}^{\mathrm{T}}\mathbf{D}_{\xi}^{1/2}\mathbf{a},$$

and standard manipulations yield equations (11) and (12).

A.3. Development of $\Delta_n^{(1)}$ and $\Delta_n^{(2)}$

In the notation of Section 2, we propose

$$\Delta_n^{(1)}(\mathbf{x};\delta_{n+1,0},\delta_{n+1,1})=\frac{\mathcal{I}\{\eta_0(\mathbf{x}|\hat{\boldsymbol{\theta}}_{n,0})+\delta_{n+1,0}(\mathbf{x}),\eta_1(\mathbf{x}|\hat{\boldsymbol{\theta}}_{n,1})+\delta_{n+1,1}(\mathbf{x})\}}{n},$$

where $\hat{\boldsymbol{\theta}}_{n,j}$ $(j=0,1)$ are the estimates that are computed after implementing $\xi^{(n)}$.

To derive an appropriate measure $\Delta_{n,j_n}^{(2)}(\mathbf{x})$, we first note that, under mild conditions, the regression estimates that are derived from a sequential design have the same asymptotically normal distribution as under independent sampling—see Sinha and Wiens (2003). Using this result, Sinha and Wiens (2002) derived the asymptotic IMSE under a contamination model as used in Section 2 of this paper, obtaining

$$\begin{aligned}\mathrm{IMSE}_{n,j}&=\sum_{i=1}^{N}E_{\mathrm{Model}j}[\{\hat{Y}(\mathbf{x}_i)-E_{\mathrm{Model}j}[Y|\mathbf{x}_i]\}^2]\\&\approx\sum_{i=1}^{N}E[\{(\hat{\boldsymbol{\theta}}_{n,j}-\boldsymbol{\theta}_j)^{\mathrm{T}}\dot{\eta}_j(\mathbf{x}_i|\boldsymbol{\theta}_j)-\delta_{n,j}(\mathbf{x}_i)\}^2]\\&=\mathrm{tr}\{\mathrm{MSE}_n(\boldsymbol{\theta}_j)\cdot\mathbf{A}(\boldsymbol{\theta}_j)\}+\|\boldsymbol{\delta}_{n,j}\|^2.\end{aligned}$$

These $p_j\times p_j$ matrices are

$$\begin{aligned}\mathbf{A}_j(\boldsymbol{\theta}_j)&=\mathbf{U}_j^{\mathrm{T}}(\boldsymbol{\theta}_j)\,\mathbf{U}_j(\boldsymbol{\theta}_j),\\\mathrm{MSE}_n(\boldsymbol{\theta}_j)&=E[(\hat{\boldsymbol{\theta}}_{n,j}-\boldsymbol{\theta}_j)(\hat{\boldsymbol{\theta}}_{n,j}-\boldsymbol{\theta}_j)^{\mathrm{T}}].\end{aligned}$$

The MSE matrix is evaluated by using the usual first-order approximations that are common to non-linear regression—see Gallant (1987), chapter 3—to obtain

$$\begin{aligned}\mathrm{MSE}_n(\boldsymbol{\theta}_j)&=E[(\hat{\boldsymbol{\theta}}_{n,j}-E[\hat{\boldsymbol{\theta}}_{n,j}])(\hat{\boldsymbol{\theta}}_{n,j}-E[\hat{\boldsymbol{\theta}}_{n,j}])^{\mathrm{T}}]+(E[\hat{\boldsymbol{\theta}}_{n,j}]-\boldsymbol{\theta}_j)(E[\hat{\boldsymbol{\theta}}_{n,j}]-\boldsymbol{\theta}_j)^{\mathrm{T}}\\&\approx\sigma_j^2\,\mathbf{B}_n^{-1}(\boldsymbol{\theta}_j)+\mathbf{B}_n^{-1}(\boldsymbol{\theta}_j)\,\mathbf{b}_n(\boldsymbol{\theta}_j)\,\mathbf{b}_n^{\mathrm{T}}(\boldsymbol{\theta}_j)\,\mathbf{B}_n^{-1}(\boldsymbol{\theta}_j),\end{aligned}$$

where σ_j^2 is the variance of Y in model j and

$$\begin{aligned}\mathbf{B}_n(\boldsymbol{\theta}_j)&=\sum_{i=1}^{N}n_i\,\dot{\eta}_j(\mathbf{x}_i|\boldsymbol{\theta}_j)\,\dot{\eta}_j^{\mathrm{T}}(\mathbf{x}_i|\boldsymbol{\theta}_j)=n\,\mathbf{U}_j^{\mathrm{T}}(\boldsymbol{\theta}_j)\mathbf{D}_{\xi^{(n)}}\,\mathbf{U}_j(\boldsymbol{\theta}_j),\\\mathbf{b}_n(\boldsymbol{\theta}_j)&=\sum_{i=1}^{N}n_i\,\dot{\eta}_j(\mathbf{x}_i|\boldsymbol{\theta}_j)\,\delta_{n,j}(\mathbf{x}_i)=n\,\mathbf{U}_j^{\mathrm{T}}(\boldsymbol{\theta}_j)\mathbf{D}_{\xi^{(n)}}\boldsymbol{\delta}_{n,j}.\end{aligned}$$

This gives

$$\mathrm{IMSE}_{n,j}\approx\sigma_j^2\,\mathrm{tr}\{\mathbf{B}_n^{-1}(\boldsymbol{\theta}_j)\,\mathbf{A}(\boldsymbol{\theta}_j)\}+\mathbf{b}_n^{\mathrm{T}}(\boldsymbol{\theta}_j)\,\mathbf{B}_n^{-1}(\boldsymbol{\theta}_j)\,\mathbf{A}(\boldsymbol{\theta}_j)\,\mathbf{B}_n^{-1}(\boldsymbol{\theta}_j)\,\mathbf{b}_n(\boldsymbol{\theta}_j)+\|\boldsymbol{\delta}_{n,j}\|^2.$$

We thus propose to measure the drop in IMSE by

$$\begin{aligned}\Delta_{n,j}^{(2)}(\mathbf{x})=\hat{\sigma}_{n,j}^2&\mathrm{tr}\{\mathbf{B}_n^{-1}(\hat{\boldsymbol{\theta}}_{n,j})\,\mathbf{A}_j(\hat{\boldsymbol{\theta}}_{n,j})\}+\mathbf{b}_n^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_{n,j})\,\mathbf{B}_n^{-1}(\hat{\boldsymbol{\theta}}_{n,j})\,\mathbf{A}_j(\hat{\boldsymbol{\theta}}_{n,j})\,\mathbf{B}_n^{-1}(\hat{\boldsymbol{\theta}}_{n,j})\,\mathbf{b}_n(\hat{\boldsymbol{\theta}}_{n,j})\\&-[\hat{\sigma}_{n,j}^2\,\mathrm{tr}\{\mathbf{B}_{n+1}^{-1}(\hat{\boldsymbol{\theta}}_{n,j})\,\mathbf{A}_j(\hat{\boldsymbol{\theta}}_{n,j})\}+\mathbf{b}_{n+1}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_{n,j})\,\mathbf{B}_{n+1}^{-1}(\hat{\boldsymbol{\theta}}_{n,j})\,\mathbf{A}_j(\hat{\boldsymbol{\theta}}_{n,j})\,\mathbf{B}_{n+1}^{-1}(\hat{\boldsymbol{\theta}}_{n,j})\,\mathbf{b}_{n+1}(\hat{\boldsymbol{\theta}}_{n,j})]-\delta_{n+1,j}^2(\mathbf{x}).\end{aligned}$$

This can be reduced somewhat, for computational purposes. Define terms

$$\begin{aligned}a\;(=a_{n,j}(\mathbf{x}))&=\dot{\eta}_j^{\mathrm{T}}(\mathbf{x}|\hat{\boldsymbol{\theta}}_{n,j})\;\mathbf{B}_n^{-1}(\hat{\boldsymbol{\theta}}_{n,j})\;\mathbf{A}_j(\hat{\boldsymbol{\theta}}_{n,j})\;\mathbf{B}_n^{-1}(\hat{\boldsymbol{\theta}}_{n,j})\;\dot{\eta}_j(\mathbf{x}|\hat{\boldsymbol{\theta}}_{n,j}),\\b\;(=b_{n,j}(\mathbf{x}))&=\dot{\eta}_j^{\mathrm{T}}(\mathbf{x}|\hat{\boldsymbol{\theta}}_{n,j})\;\mathbf{B}_n^{-1}(\hat{\boldsymbol{\theta}}_{n,j})\;\dot{\eta}_j(\mathbf{x}|\hat{\boldsymbol{\theta}}_{n,j}),\\c\;(=c_{n,j}(\mathbf{x}))&=\dot{\eta}_j^{\mathrm{T}}(\mathbf{x}|\hat{\boldsymbol{\theta}}_{n,j})\;\mathbf{B}_n^{-1}(\hat{\boldsymbol{\theta}}_{n,j})\;\mathbf{b}_n(\hat{\boldsymbol{\theta}}_{n,j}),\\d\;(=d_{n,j}(\mathbf{x}))&=\dot{\eta}_j^{\mathrm{T}}(\mathbf{x}|\hat{\boldsymbol{\theta}}_{n,j})\;\mathbf{b}_n^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_{n,j})\;\mathbf{B}_n^{-1}(\hat{\boldsymbol{\theta}}_{n,j})\;\mathbf{A}_j(\hat{\boldsymbol{\theta}}_{n,j})\;\mathbf{B}_n^{-1}(\hat{\boldsymbol{\theta}}_{n,j})\;\mathbf{b}_n(\hat{\boldsymbol{\theta}}_{n,j}).\end{aligned}$$

Then after a calculation we obtain

$$\Delta^{(2)}_{n,j}(\mathbf{x};\delta_{n+1,j}) = \frac{a\hat{\sigma}^2_{n,j} + 2d\{c-\delta_{n+1,j}(\mathbf{x})\}}{1+b} - \frac{a\{c-\delta_{n+1,j}(\mathbf{x})\}^2}{(1+b)^2} - \delta^2_{n+1,j}(\mathbf{x}).$$

Note that only $\mathbf{B}_n(\hat{\boldsymbol{\theta}}_{n,j})$ needs to be inverted.

To carry out the minimization step we write, as at equation (7), $\delta_{n+1,j}(\mathbf{x}_i) = \mathbf{v}^{\mathrm{T}}_{j,i}\mathbf{c}_j$. A strict application of equation (14) then calls for us to determine

$$L_n(\mathbf{x}_i) = \min_{\|\mathbf{c}_0\| \leqslant \tau_0, \|\mathbf{c}_1\| \leqslant \tau_1} \left(\begin{array}{c} (1-\kappa_n)\dfrac{\mathcal{I}\{\eta_0(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_{n,0}) + \mathbf{v}^{\mathrm{T}}_{0,i}\mathbf{c}_0, \eta_1(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_{n,1}) + \mathbf{v}^{\mathrm{T}}_{1,i}\mathbf{c}_1\}}{n} \\ +\kappa_n\left[\dfrac{a_{n,j}(\mathbf{x}_i)\hat{\sigma}^2_{n,j} + 2\,d_{n,j}(\mathbf{x}_i)\{c_{n,j}(\mathbf{x}_i) - \mathbf{v}^{\mathrm{T}}_{j,i}\mathbf{c}_j\}}{1+b_{n,j}(\mathbf{x}_i)} - \dfrac{a_{n,j}(\mathbf{x}_i)\{c_{n,j}(\mathbf{x}_i) - \mathbf{v}^{\mathrm{T}}_{j,i}\mathbf{c}_j\}^2}{\{1+b_{n,j}(\mathbf{x}_i)\}^2} - \|\mathbf{c}_j\|^2\right] \end{array} \right)$$

for $i = 1, \ldots, N$; then the next design point is that for which $L_n(\cdot)$ is largest. In fact, rather than carry out N separate (pairs of) minimizations for each new design point, we have opted for the following, much simpler method of introducing the uncertainty about the models. In this method the minimization is carried out once only, using the initial design. The N values $\delta(\mathbf{x}_i)$ that are so obtained are then randomly permuted, resulting in values $\{\Delta_n(\mathbf{x}_i)\}$, the minimum of which yields the next design point.

References

Atkinson, A. C. (2008) DT-optimum designs for model discrimination and parameter estimation. *J. Statist. Planng Inf.*, **138**, 56–64.

Atkinson, A. C. and Cox, D. R. (1974) Planning experiments for discriminating between models (with discussion). *J. R. Statist. Soc.* B, **36**, 321–348.

Atkinson, A. C. and Fedorov, V. V. (1975a) The design of experiments for discriminating between two rival models. *Biometrika*, **62**, 57–70.

Atkinson, A. C. and Fedorov, V. V. (1975b) Optimal design: experiments for discriminating between several models. *Biometrika*, **62**, 289–303.

Biedermann, S. and Dette, H. (2001) Optimal designs for testing the functional form of a regression via nonparametric estimation techniques. *Statist. Probab. Lett.*, **52**, 215–224.

Bischoff, W. and Miller, F. (2006) Optimal designs which are efficient for lack of fit tests. *Ann. Statist.*, **34**, 2015–2025.

Box, G. E. P. and Draper, N. R. (1959) A basis for the selection of a response surface design. *J. Am. Statist. Ass.*, **54**, 622–654.

Cox, D. R. and Reid, N. (2000) *The Theory of the Design of Experiments*. Boca Raton: Chapman and Hall–CRC.

Dette, H. and Kwiecien, R. (2004) A comparison of sequential and non-sequential designs for discrimination between nested regression models. *Biometrika*, **91**, 165–176.

Dette, H. and Titoff, S. (2008) Optimal discrimination designs. *Ann. Statist.*, to be published.

Fang, Z. and Wiens, D. P. (2000) Integer-valued, minimax robust designs for estimation and extrapolation in heteroscedastic, approximately linear models. *J. Am. Statist. Ass.*, **95**, 807–818.

Fedorov, V. V. (1975) Optimal experimental designs for discriminating two rival regression models. In *A Survey of Statistical Design and Linear Models* (ed. J. N. Srivastava). Amsterdam: North-Holland.

Fedorov, V. V. and Pazman, A. (1968) Design of physical experiments. *Fortsch. Phys.*, **16**, 325–355.

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

Gallant, A. R. (1987) *Nonlinear Statistical Models*. Toronto: Wiley.

Hill, P. D. H. (1978) A review of experimental design procedures for regression model discrimination. *Technometrics*, **20**, 15–21.

Hill, W. J., Hunter, W. G. and Wichern, D. W. (1968) A joint design criterion for the dual problem of model discrimination and parameter estimation. *Technometrics*, **10**, 145–160.

Hunter, W. G. and Reiner, A. M. (1965) Designs for discriminating between two rival models. *Technometrics*, **7**, 307–323.

López-Fidalgo, J., Tommasi, C. and Trandafir, P. C. (2007) An optimal experimental design criterion for discriminating between non-normal models. *J. R. Statist. Soc.* B, **69**, 231–242.

Pukelsheim, F. and Rosenberger, J. L. (1993) Experimental designs for model discrimination. *J. Am. Statist. Ass.*, **88**, 642–649.

Sinha, S. and Wiens, D. P. (2002) Robust sequential designs for nonlinear regression. *Can. J. Statist.*, **30**, 601–618.

Sinha, S. and Wiens, D. P. (2003) Asymptotics for robust sequential designs in misspecified regression models; mathematical statistics and applications. In *Festschrift for Constance van Eeden* (eds M. Moore, C. Léger and S. Froda), pp. 233–248. Hayward: Institute of Mathematical Statistics.

Uciński, D. and Bogacka, B. (2005) T-optimum designs for discrimination between two multiresponse dynamic models. *J. R. Statist. Soc.* B, **67**, 3–18.

Wiens, D. P. (1991) Designs for approximately linear regression: two optimality properties of uniform designs. *Statist. Probab. Lett.*, **12**, 217–221.

Wiens, D. P. (1992) Minimax designs for approximately linear regression. *J. Statist. Planng Inf.*, **31**, 353–371.

Wiens, D. P. (2009) Asymptotic properties of a Neyman-Pearson test for model discrimination, with an application to experimental design. *J. Statist. Theory Pract.*, to be published.