The Annals of Statistics
2017, Vol. 45, No. 4, 1638–1663
DOI: 10.1214/16-AOS1503

ROBUST DISCRIMINATION DESIGNS OVER HELLINGER NEIGHBOURHOODS[1]

BY RUI HU AND DOUGLAS P. WIENS

MacEwan University and University of Alberta

To aid in the discrimination between two, possibly nonlinear, regression models, we study the construction of experimental designs. Considering that each of these two models might be only approximately specified, robust "maximin" designs are proposed. The rough idea is as follows. We impose neighbourhood structures on each regression response, to describe the uncertainty in the specifications of the true underlying models. We determine the least favourable—in terms of Kullback–Leibler divergence—members of these neighbourhoods. Optimal designs are those maximizing this minimum divergence. Sequential, adaptive approaches to this maximization are studied. Asymptotic optimality is established.

1. Introduction. Much of the experimental work in scientific disciplines—physics, chemistry, engineering, etc.—is concerned with the elucidation of a functional relationship between a response variable y and various covariates $\mathbf{x}$. However, in practice it is often the case that the investigator will not know the correct functional form, but instead will have several plausible models in mind. A first aim of the investigator is therefore to design an experiment distinguishing among these rival models. In this article, we assume that two rival models are available. Under the first model, the data arise from a population with density $f_0(y|\mathbf{x}, \boldsymbol{\varphi}_0)$ while under the other model the density is $f_1(y|\mathbf{x}, \boldsymbol{\varphi}_1)$; the conditional means are

$$\mu_j(\mathbf{x}) = \int y f_j(y|\mathbf{x}, \boldsymbol{\varphi}_j)\, dy, \qquad j = 0, 1. \tag{1.1}$$

Here, $\boldsymbol{\varphi}_0$ and $\boldsymbol{\varphi}_1$ represent nuisance parameters and will not be explicitly mentioned if there is no possibility of confusion. Given a design space $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^N$, suppose that $y_{il}, l = 1, \ldots, n_i \geq 0$, are observations made at the covariate $\mathbf{x}_i$. Usually, the model discrimination problem is cast as a problem of hypothesis testing [Atkinson and Fedorov (1975a, 1975b), Fedorov (1975)]:

$$H_0: \ f_0\big(y|\mathbf{x}, \mu_0(\mathbf{x})\big) \quad \text{versus} \quad H_1: \ f_1\big(y|\mathbf{x}, \mu_1(\mathbf{x})\big), \qquad \mathbf{x} \in \mathcal{S}.$$

Received April 2016; revised June 2016.

[1]Supported by the Natural Sciences and Engineering Research Council of Canada.

MSC2010 subject classifications. Primary 62K99, 62H30; secondary 62F35.

Key words and phrases. Adaptive design, Hellinger distance, Kullback–Leibler divergence, maximin, Michaelis–Menten model, Neyman–Pearson test, nonlinear regression, optimal design, robustness, sequential design.

The Neyman–Pearson test then can be used to compare these two hypotheses. Define $\mathcal{R} = \sum_{i,l} R(y_{il})$ with

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

The test rejects H_0 for large values of $\mathcal{R}$. We shall assume that the experimenter models $\mu_j(\mathbf{x})$ parametrically as $\eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$, with the form of $\eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$ specified but the parameters $\boldsymbol{\theta}_j$ unknown.

When only two rival models are available, the design of experiments for discrimination has been investigated by numerous authors, among them Fedorov (1975), Hill (1978), Dette and Titoff (2009). Extensions to discrimination between several models are also explored in Atkinson and Fedorov (1975b), Dette (1994), Dette and Haller (1998).

Sequential and static designs are the two most well-studied strategies. Hunter and Reiner (1965) proposed a sequential design assuming that both densities were Gaussian, namely, $f_j(y|\mathbf{x}, \mu_j, \sigma) = \sigma^{-1}\phi((y - \mu_j(\mathbf{x}))/\sigma)$, $j = 0, 1$. Fedorov and Pazman (1968) extended the method to heteroscedastic models. Static, that is, nonsequential, design strategies were constructed under the normality assumption by Atkinson and Fedorov (1975a, 1975b). López-Fidalgo, Tommasi and Trandafir (2007) extended static design to nonnormal models.

The criteria to be optimized in the articles cited above are, or are equivalent to, the integrated Kullback–Leibler (KL) divergence $\mathcal{D}$:

$$\mathcal{D}(f_0, f_1, \boldsymbol{\xi}|\mu_0, \mu_1) = \int_{\mathcal{S}} \mathcal{I}\{f_0, f_1|\mathbf{x}, \mu_0(\mathbf{x}), \mu_1(\mathbf{x})\}\boldsymbol{\xi}(d\mathbf{x}), \tag{1.2}$$

with ξ being the design measure placing mass $\xi_i = n_i/n$ at $\mathbf{x}_i$, and

$$\mathcal{I}\{f_0, f_1|\mathbf{x}, \mu_0(\mathbf{x}), \mu_1(\mathbf{x})\} = \int_{-\infty}^{\infty} f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))\log\left\{\frac{f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))}{f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))}\right\} dy \tag{1.3}$$

being the Kullback–Leibler divergence measuring the information lost when $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ is used to approximate $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$.

In the references above, one of $f_j(y|\mathbf{x}, \mu_j)$, $j = 0, 1$, is assumed to correctly represent the true physical mechanism. However, it is dangerous to apply a method that is highly dependent on a specific form [Box and Draper (1959), Huber (1981), Ford, Titterington and Kitsos (1989)]. From a viewpoint of robustness, it is more sensible to suppose only that the correct model lies in a neighbourhood of a specified density. Wiens (2009a) allowed for the means (but not the f_j) to be specified erroneously, and imposed the neighbourhood structure

$$\mu_j(\mathbf{x}) = \eta_j(\mathbf{x}|\boldsymbol{\theta}_j) + \psi_j(\mathbf{x}) \tag{1.4}$$

for specified $\eta_j(\mathbf{x}|\cdot)$. The vectors $\boldsymbol{\psi}_j = (\psi_j(\mathbf{x}_1), \ldots, \psi_j(\mathbf{x}_N))'$ were allowed to range over classes $\boldsymbol{\Psi}_j$, resulting in the neighbourhoods:

$$\mathcal{F}_j = \{f_j(\cdot|\mathbf{x}, \mu_j)|\mu_j(\mathbf{x}_i) = \eta_j(\mathbf{x}|\boldsymbol{\theta}_j) + \psi_j(\mathbf{x}), \boldsymbol{\psi}_j \in \boldsymbol{\Psi}_j\}, \qquad j = 0, 1.$$

Under this setting, robust Kullback–Leibler optimal designs were obtained in Wiens (2009a) by maximizing the minimum asymptotic power of the Neyman–Pearson test statistic $\mathcal{R}$ over $\mathcal{F}_0$ and $\mathcal{F}_1$. The asymptotic properties of $\mathcal{R}$ were derived in Wiens (2009b) for two rival models with common densities $f_j(y|\mathbf{x}, \mu_j(\mathbf{x})) = f(y|\mathbf{x}, \mu_j(\mathbf{x}))$.

Our work is a natural sequel to Wiens (2009a). Model misspecification is still the problem we would like to address, but under a more general scenario which we now describe. The two rival models are $f_j(y|\mathbf{x}, \mu_j(\mathbf{x}))$, $j = 0, 1$, with $\mu_j(\mathbf{x})$ determined by (1.1) and assumed to be of the form $\eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$ for some $\boldsymbol{\theta}_j$, that is, $\boldsymbol{\psi}_j \equiv 0$. Define $\mathcal{F}_j$ to be neighbourhoods of $f_j(y|\mathbf{x}, \mu_j(\mathbf{x}))$ used to describe inaccuracies in the specifications of the true underlying densities. The true model lies in one of $\mathcal{F}_j$, $j = 0, 1$. It is our purpose in this paper to propose methods of discrimination design which are robust against the possible model misspecification mentioned above.

We entertain the following two scenarios, but for the most part will concentrate on the first.

Case I: Under the null hypothesis, the density function $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ of the response variable is fixed and its mean μ_0 is as defined in (1.1); under the alternative hypothesis the density function varies over a Hellinger neighbourhood of a nominal density $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$. Recall that the Hellinger distance $d_h(f, g)$ between densities f, g is defined by

$$d_h^2(f, g) = \frac{1}{2}\int \bigl(f^{1/2}(y) - g^{1/2}(y)\bigr)^2 dy = 1 - \int \sqrt{f(y)g(y)}\, dy.$$

Here, the two classes are $\mathcal{F}_0 = \{f_0(y|\mathbf{x}, \mu_0(\mathbf{x})\}$ and $\mathcal{F}_1$ is a Hellinger neighbourhood defined as

$$\mathcal{F}_1(\varepsilon_1) = \Bigl\{f(y|\mathbf{x}) \bigm| \max_{\mathbf{x}\in\mathcal{S}} d_h\bigl(f(y|\mathbf{x}), f_1\bigl(y|\mathbf{x}, \mu_1(\mathbf{x})\bigr)\bigr) \le \varepsilon_1\Bigr\} \tag{1.5}$$

for some $\varepsilon_1 > 0$. The members of $\mathcal{F}_1$ may differ from f_1 because of differences in the functional form of the density, or in their mean structures, or both.

Case II: Under the null hypothesis, the response variable has density $f(y|\mathbf{x})$ varying over a Hellinger neighbourhood of a nominal density $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$. The members of $\mathcal{F}_0$ may differ from f_0 because of differences in the functional form of the density, or in their mean structures, or both. Under the alternative hypothesis, the density function is $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$ with the mean μ_1 defined in (1.1). In this case, the two classes $\mathcal{F}_0(\varepsilon_0)$ and $\mathcal{F}_1$ are defined as

$$\mathcal{F}_0(\varepsilon_0) = \Bigl\{f(y|\mathbf{x}) \bigm| \max_{\mathbf{x}\in\mathcal{S}} d_h\bigl(f(y|\mathbf{x}), f_0\bigl(y|\mathbf{x}, \mu_0(\mathbf{x})\bigr)\bigr) \le \varepsilon_0\Bigr\},$$

for some $\varepsilon_0 > 0$, and $\mathcal{F}_1 = \{f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))\}$, respectively.

Thus, Case I fixes the null model and allows the alternate to vary over a Hellinger class; in Case II these are reversed. In this paper, we will primarily focus

on Case I. In the next section, we will show that the Neyman–Pearson test for discriminating between any pair in $\mathcal{F}_0 \times \mathcal{F}_1$ is related to the KL-divergence defined in (1.2) between the pair of densities. These results are also applicable to Case II.

All derivations, and longer mathematical arguments, are in the Appendix or in the supplementary document [Hu and Wiens (2017)]. Some omitted details may be found in Hu (2016).

2. Asymptotic properties of the test statistic $\mathcal{R}$. In the asymptotics literature, one finds numerous results about the asymptotic distribution of $\mathcal{R}$, the test statistic for the discrimination between a pair of models $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$, under various conditions. Wiens (2009b) proved the asymptotic normality of $\mathcal{R}$ under standard regularity conditions for likelihood estimation. In Oosterhoff and van Zwet (2012), similar results are proved under certain contiguity assumptions. In the Appendix, we derive the asymptotic distribution of $\mathcal{R}$ under conditions tailored to our problem. The proof follows that in Oosterhoff and van Zwet (2012). It is rather long, and depends on a number of preliminary lemmas, and so is given in the supplementary document Hu and Wiens (2017).

THEOREM 2.1. *Given a design space $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^N$, assume that the experiment has n_i replicates at each covariate $\mathbf{x}_i$, with $\sum_{i=1}^N n_i = n$. Define $\mathcal{D}$ as in (1.2) and for any two densities f_0, f_1 define*

$$r(y|\mathbf{x}_i; f_0, f_1) = \frac{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))}.$$

Assume that the densities f_0, f_1 satisfy:

(a) *for the KL-divergence*,

$$n\mathcal{D} = O(1), \tag{2.1}$$

(b) *for all $\delta > 0$*

$$\lim_{n\to\infty} \sum_{i=1}^N n_i \int_{\{|\log r(y|\mathbf{x}_i; f_0, f_1)| \geq \delta\}} f_0\big(y|\mathbf{x}_i, \mu_0(\mathbf{x})\big) \times \big(\sqrt{r(y|\mathbf{x}_i; f_0, f_1)} - 1\big)^2 dy = 0, \tag{2.2}$$

(c) *there is a $\tau > 0$ such that*

$$\lim_{n\to\infty} \sum_{i=1}^N n_i \int_{\{\log r(y|\mathbf{x}_i; f_0, f_1) \geq \tau\}} f_1\big(y|\mathbf{x}_i, \mu_1(\mathbf{x})\big) \log r(y|\mathbf{x}_i; f_0, f_1)\, dy = 0. \tag{2.3}$$

Then $\mathcal{R} = 2\sum_{i=1}^N \sum_{l=1}^{n_i} \log r(y|\mathbf{x}_i; f_0, f_1)$ and:

(i) *under the null hypothesis*,

$$\frac{\mathcal{R}+2n\mathcal{D}}{\sqrt{8n\mathcal{D}}} \xrightarrow{L} N(0,1),$$

(ii) *under the alternative hypothesis*,

$$\frac{\mathcal{R}-2n\mathcal{D}}{\sqrt{8n\mathcal{D}}} \xrightarrow{L} N(0,1).$$

REMARK 1. As is also shown in the supplementary document [Hu and Wiens (2017)], conditions (2.1)–(2.3) in Theorem 2.1 hold in particular when the two densities $f_j(y|\mathbf{x},\mu_j(\mathbf{x}),\sigma)=\phi((y-\mu_j(\mathbf{x}))/\sigma)/\sigma$, $j=0,1$, have means which satisfy $\mu_1(\mathbf{x})=\mu_0(\mathbf{x})+n^{-1/2}\Delta(\mathbf{x})$ for a bounded function Δ. In particular, condition (2.3) holds for every $\tau>0$. The same conclusion holds for log-normal densities $f_j(y|\mathbf{x},\mu_j(\mathbf{x}),v_j^2)$ with $\mu_1(\mathbf{x})=\mu_0(\mathbf{x})+n^{-1/2}\Delta(\mathbf{x})$ and homogeneous variances $v_j^2=v^2$.

REMARK 2. Denote by F a distribution function whose density is f. To guarantee that the KL divergence between two densities f_0 and f_1 is finite, F_1 should be absolutely continuous with respect to F_0. Moreover, it is natural to assume that the two rival models are close to each other in some sense. Therefore, in the following we let $f_0(y|\mathbf{x},\mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x},\mu_1(\mathbf{x}))$ have the same support set $\Omega_{\mathbf{x}}$, and its complement set is

$$\Omega_{\mathbf{x}}^c=\{y: f_1(y|\mathbf{x},\mu_1(\mathbf{x}))=f_0(y|\mathbf{x},\mu_0(\mathbf{x}))=0\}.$$

Then to make sure that condition (2.1) is reasonable, F_1 should be absolutely continuous with respect to F_0. Therefore, we only consider $f(y|\mathbf{x})\in\mathcal{F}_1(\varepsilon_1)$ such that $f(y|\mathbf{x})=0$ on $\Omega_{\mathbf{x}}^c$. For simplicity, we assume that the densities we consider in this paper are continuous in the interiors of their support sets.

If the radii ε_j of the Hellinger neighbourhoods $\mathcal{F}_0(\varepsilon_0)$ of $f_0(\cdot|\mathbf{x},\mu_0(\mathbf{x}))$ and $\mathcal{F}_1(\varepsilon_1)$ of $f_1(\cdot|\mathbf{x},\mu_1(\mathbf{x}))$ shrink at a rate $o(n^{-1/2})$, then the results in Theorem 2.1 also hold for any pair of densities in $\mathcal{F}_0(\varepsilon_0)\times\mathcal{F}_1(\varepsilon_1)$. This is guaranteed by the result in the following corollary.

COROLLARY 1. *Assume that the central densities* $f_0(y|\mathbf{x},\mu_0(\mathbf{x}))$ *and* $f_1(y|\mathbf{x},\mu_1(\mathbf{x}))$ *satisfy conditions* (2.1)–(2.3) *in Theorem* 2.1 *and that* $\varepsilon_0=o(n^{-1/2})$, $\varepsilon_1=o(n^{-1/2})$. *Then for any pair* $(f^{(0)}(y|\mathbf{x}),f^{(1)}(y|\mathbf{x}))\in\mathcal{F}_0(\varepsilon_0)\times\mathcal{F}_1(\varepsilon_1)$ *satisfying* (2.3) *we have that* $f^{(0)}(y|\mathbf{x})$ *and* $f^{(1)}(y|\mathbf{x})$ *also satisfy conditions* (2.1) *and* (2.2).

The main results in Theorem 2.1 and Corollary 1 show the asymptotic normality of the statistic

$$\mathcal{R}(f^{(0)},f^{(1)})=2\sum_{i=1}^{N}\sum_{l=1}^{n_i}\log r(y|\mathbf{x}_i;f^{(0)},f^{(1)}),$$

under $f^{(0)}(y|\mathbf{x}) \in \mathcal{F}_0(\varepsilon_0)$ or $f^{(1)}(y|\mathbf{x}) \in \mathcal{F}_1(\varepsilon_1)$, that is, the density of the observation variable Y is $f^{(0)}(y|\mathbf{x})$ or $f^{(1)}(y|\mathbf{x})$. In practice, we are more interested in the asymptotic normality of the test statistic $\mathcal{R} := \mathcal{R}(f_0, f_1)$ for the discrimination of the two nominal models $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$, when, however, the true model is in $\mathcal{F}_0(\varepsilon_0)$ or $\mathcal{F}_1(\varepsilon_1)$. In Theorem 2.2, we show that the asymptotic normality of $\mathcal{R}$ still holds under any density $f \in \mathcal{F}_0(\varepsilon_0) \cup \mathcal{F}_1(\varepsilon_1)$.

THEOREM 2.2. *Assume that the two models* $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$ *and* $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ *satisfy conditions* (2.1)–(2.3) *in Theorem* 2.1. *Then*:

(i) *under* $f \in \mathcal{F}_1(\varepsilon_1)$,

$$\frac{\mathcal{R} - 2n\mathcal{D}(f_0, f)}{\sqrt{8n\mathcal{D}(f_0, f)}} \xrightarrow{L} N(0, 1);$$

(ii) *under* $f \in \mathcal{F}_0(\varepsilon_0)$,

$$\frac{\mathcal{R} + 2n\mathcal{D}(f, f_1)}{\sqrt{8n\mathcal{D}(f, f_1)}} \xrightarrow{L} N(0, 1).$$

The following theorem is immediate from Theorem 2.2, but summarizes the results in the context of the asymptotic power of the test against a density $f \in \mathcal{F}_0$ or $f \in \mathcal{F}_1$.

THEOREM 2.3. *The asymptotic power against a density* $f \in \mathcal{F}_0$ *or* $\mathcal{F}_1$ *is*

$$\pi(f) := P_f\big(\mathcal{R}(f_0, f_1) > c\big) = \Phi\left(\frac{-c + \gamma(f, f_0, f_1)}{2\sqrt{|\gamma(f, f_0, f_1)|}}\right) + o(1),$$

where $P_f(\cdot)$ *means that the calculations are to be made assuming that* f *is the density of* Y, *and where*

$$\gamma(f, f_0, f_1) = \begin{cases} -2n\mathcal{D}(f, f_1), & \text{if } f \in \mathcal{F}_0, \\ 2n\mathcal{D}(f_0, f), & \text{if } f \in \mathcal{F}_1. \end{cases}$$

The critical value is

$$c = -2n\mathcal{D}(f_0, f_1) + u_\alpha\sqrt{8n\mathcal{D}(f_0, f_1)},$$

determined by

$$\alpha = P_{f_0}\big(\mathcal{R}(f_0, f_1) > c\big),$$

with u_α *being the* $(1 - \alpha)$*-quantile of the standard normal distribution.*

Design criteria for Case I. In Case I, where $\mathcal{F}_1(\varepsilon_1)$ is a neighbourhood of f_1, the design problem is to find a design to maximize the "worst" power with controlled Type I error. In particular, a *robust maximin* design $\boldsymbol{\xi}^*$ is constructed which maximizes the minimum power, with significance level α, over $\mathcal{F}_1(\varepsilon_1)$, that is,

$$\boldsymbol{\xi}^* = \arg\max_{\boldsymbol{\xi}} \min_{f_1 \in \mathcal{F}_1(\varepsilon_1)} P_{f_1}(\mathcal{R} > c) \qquad \text{subject to } \alpha = P_{f_0}(\mathcal{R} > c),$$

where c is the critical value defining the rejection region $\{\mathcal{R} > c\}$. According to Theorem 2.3, asymptotically, the robust design is the solution to the optimality problem

$$\boldsymbol{\xi}^* = \arg\max \min_{f \in \mathcal{F}_1(\varepsilon_1)} \pi(f),$$

and the minimum asymptotic power is

$$\min_{f \in \mathcal{F}_1(\varepsilon_1)} \Phi\left(\frac{-c + 2n\mathcal{D}(f_0, f)}{2\sqrt{2n\mathcal{D}(f_0, f)}}\right), \tag{2.4}$$

where c is defined in Theorem 2.3. Under certain condition, we can solve the minimization problem by minimizing the integrated KL-divergence $\mathcal{D}(f_0, f)$, as shown in the following proposition.

PROPOSITION 1. *Define*

$$f_{1*} = \arg \min_{f \in \mathcal{F}_1(\varepsilon_1)} \mathcal{D}(f_0, f). \tag{2.5}$$

If

$$\mathcal{D}(f_0, f_{1*}) \geq -c, \tag{2.6}$$

then also f_{1*} *minimizes* $\pi(f)$ *in* $\mathcal{F}_1(\varepsilon_1)$, *and so is the desired minimizer in* (2.4).

REMARK 3. If $c \geq 0$, then (2.6) is automatic. Otherwise, we check it numerically.

The problem now is to find f_{1*} as at (2.5), and then

$$\boldsymbol{\xi}^* = \arg\max \min_{f \in \mathcal{F}_1(\varepsilon_1)} \pi(f) = \arg\max \mathcal{D}(f_0, f_{1*}). \tag{2.7}$$

For Case II, we view the null hypothesis as composite, in the sense that f_0 is the representative of the whole neighbourhood and the nominal size of the test is evaluated at f_0, and is to be α. We should accept the null if it appears that the f generating the data is anything in $\mathcal{F}_0(\varepsilon_0)$. Then, if $f \in \mathcal{F}_0(\varepsilon_0)$ is generating the data we make an error if we reject the null hypothesis, and we would like to minimize the (maximum) probability of this. A *robust maximin* design $\boldsymbol{\xi}^*$ is

then constructed which minimizes the maximum probability of such an error, with significance level α, over $\mathcal{F}_0(\varepsilon_0)$. This case will be investigated in future work.

As an illustration, we consider two models $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$ with $\mu_0(\mathbf{x}) = \eta_0(\mathbf{x}|\boldsymbol{\theta}_0)$ and $\mu_1(\mathbf{x}) = \eta_1(\mathbf{x}|\boldsymbol{\theta}_1)$. The design obtained from the criterion (2.7) is robust for testing the hypotheses

$$H_0 : f_0\big(y|\mathbf{x}, \mu_0(\mathbf{x})\big) \quad \text{vs.} \quad H_1 : f_1\big(y|\mathbf{x}, \mu_1(\mathbf{x})\big),$$

when the true model is in a small neighbourhood of one of the hypothesized models.

Of course, the values of the regression parameters are unknown. To address this problem, Atkinson and Fedorov (1975a) [see also López-Fidalgo, Tommasi and Trandafir (2007)] assumes a range of plausible values for the parameters. The worst possible values of the parameters, that is, those that minimize $\mathcal{D}$, are obtained within their respective ranges and the maximin optimal design that maximizes this minimum value is constructed. This method leads to static design strategies. In this paper, we proceed *sequentially* and *adaptively*, with the parameters $\boldsymbol{\theta}_j$ replaced by updated least squares (LS) estimates $\widehat{\boldsymbol{\theta}}_j$ before proceeding to the next stage. The next observation then will be made at the point $\mathbf{x}_{\text{new}}$ optimizing the discrepancy function [e.g., the KL divergence (1.3)] evaluated at the $\widehat{\boldsymbol{\theta}}_j$. This is repeated until sufficiently many design points and observations are obtained. See Hunter and Reiner (1965) and Fedorov and Pazman (1968) for background material.

In Section 3, the minimization problem is solved analytically at each fixed $\mathbf{x} \in \mathcal{S}$. The maximization leading to optimal designs is done numerically. In Section 4 a sequential strategy is proposed, with the unknown parameters in $\eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$ updated as described above. To see how the test performs with our robust designs, we simulate the sizes and powers of the model discrimination test to discriminate between two models $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$ when the true model may merely be close to the nominal model $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$.

3. Minimization of the discrepancy function.

Assume that

$$\varepsilon_1 < \min_{\mathbf{x}\in\mathcal{S}} d_h\big(f_0\big(y|\mathbf{x}, \mu_0(\mathbf{x})\big), f_1\big(y|\mathbf{x}, \mu_1(\mathbf{x})\big)\big); \tag{3.1}$$

this ensures that $\mathcal{F}_0$ and $\mathcal{F}_1(\varepsilon_1)$ are disjoint—otherwise, the minimum power of the test is zero. That ε_1 is sufficiently small will be checked numerically in each example.

For the first step, we minimize $\mathcal{D}$ over the neighbourhood $\mathcal{F}_1(\varepsilon_1)$. Equivalently, we consider the optimization problem

$$\min_f \mathcal{D}(f_0, f, \boldsymbol{\xi}|\mu_0) = \min_f \sum_{i=1}^{N} \xi_i \int f(y|\mathbf{x}_i) \log\left(\frac{f(y|\mathbf{x}_i)}{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))}\right) dy, \tag{3.2}$$

under the constraints (1.5) and $\int f(y|\mathbf{x})\, dy = 1$. (The requirement that f be nonnegative turns out to be satisfied automatically and need not be prescribed.)

It is sufficient to find, for each $\mathbf{x}$, the minimizer $f_{1*}(y|\mathbf{x})$ of

$$\mathcal{I}\{f_0, f|\mathbf{x}, \mu_0(\mathbf{x})\} = \int f(y|\mathbf{x}) \log\left(\frac{f(y|\mathbf{x})}{f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))}\right) dy, \tag{3.3}$$

subject to (i′) $\int \sqrt{f(y|\mathbf{x}) f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))}\, dy \geq 1 - \varepsilon_1^2$ and (ii′) $\int f(y|\mathbf{x})\, dy = 1$. That is, we consider the optimality problem at each $\mathbf{x}$.

To solve this minimization problem, we adopt the Lagrange multiplier method and obtain the following result. For each $\mathbf{x}$, this gives a value $f_{1*}(y|\mathbf{x}, \mu_{1*}(\mathbf{x}))$ of the least favourable density, with mean $\mu_{1*}(\mathbf{x})$ given by (1.1).

PROPOSITION 2. *For $\mathbf{x} \in \mathcal{S}$, consider the system*

$$\log \frac{f(y|\mathbf{x})}{f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))} + 1 + \frac{1}{2}\lambda_1 f_1^{1/2}\big(y|\mathbf{x}, \mu_1(\mathbf{x})\big) f^{-1/2}(y|\mathbf{x}) + \lambda_2 = 0, \tag{3.4}$$

$$\int_{\Omega_{\mathbf{x}}} f(y|\mathbf{x})\, dy = 1, \tag{3.5}$$

$$\int_{\Omega_{\mathbf{x}}} \sqrt{f(y|\mathbf{x}) f_1\big(y|\mathbf{x}, \mu_1(\mathbf{x})\big)}\, dy = 1 - \varepsilon_1^2. \tag{3.6}$$

Define $f_{1*}(y|\mathbf{x}, \mu_{1*}(\mathbf{x}))$ as follows: For $y \in \Omega_{\mathbf{x}}^c$, $f_{1*}(y|\mathbf{x}, \mu_{1*}(\mathbf{x})) \equiv 0$ and for $y \in \Omega_{\mathbf{x}}$, $f_{1*}(y|\mathbf{x}, \mu_{1*}(\mathbf{x}))$ is a solution to (3.4)–(3.6) with $\lambda_1(\mathbf{x}) < 0$ and $\lambda_2(\mathbf{x}) \in \mathbb{R}$. Then $f_{1*}(y|\mathbf{x}, \mu_{1*}(\mathbf{x}))$ is the minimizer of (3.3).

As examples, in Figure 1 we plot, for fixed values of $\mathbf{x}$, the densities $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$ and the least favourable density $f_{1*}(y|\mathbf{x}, \mu_{1*}(\mathbf{x}))$. In Figure 1(a), both f_0 and f_1 are normal; in Figure 1(b), both are log-normal. In both cases, the shape of the least favourable density obtained from Proposition 2 is, as one would expect, close to the nominal density $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$.

Based on results in Proposition 2, the answer to the optimality problem (3.2) is given in following theorem.

THEOREM 3.1. *For a design $\boldsymbol{\xi}$ placing a fraction ξ_i of the observations at $\mathbf{x}_i$, the minimum divergence $\mathcal{D}$ over $\mathcal{F}_1$ is*

$$\sum_{i=1}^{N} \xi_i \mathcal{I}\{f_0, f_{1*}|\mathbf{x}_i, \mu_0(\mathbf{x}_i), \mu_1(\mathbf{x}_i)\},$$

where f_{1*} is as given in Proposition 2.

4. Sequential, adaptive discrimination designs. We construct sequential adaptive designs, in which at each stage the choice of the next design point is informed by data gathered previously. The designs we propose are robust against the assumption that the hypotheses are correctly specified. To demonstrate this in finite

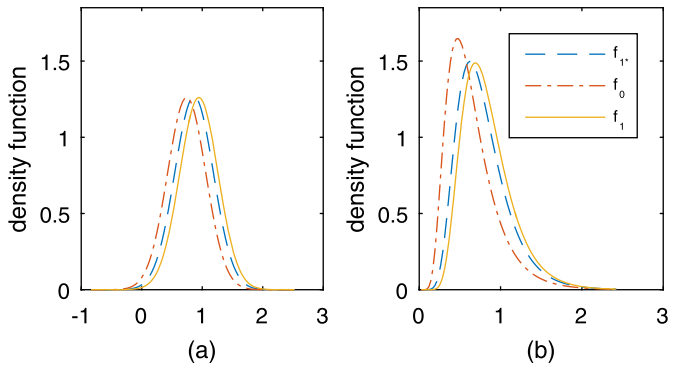

FIG. 1. *Dotted line represents* $f_0(y|x, \mu_0)$ *and solid line is* $f_1(y|x, \mu_1)$ *at* $x = 2.89$. (a) f_0 *and* f_1 *are* normal *densities* (*see Example* 1 *in Section* 4.1). (b) f_0 *and* f_1 *are* lognormal *densities* (*see Example* 2 *in Section* 4.2). *Here* f_0 *and* f_1 *have mean functions* $\mu_0(x)$ *and* $\mu_1(x)$ *being given by* (4.1) *and* (4.2), *respectively, with* $V = K = 1$. *The two nominal densities have the same variance* 0.1. *Dashed line represents least favourable density in* $\mathcal{F}_1(0.045)$.

samples, we shall simulate the sizes and minimum powers of the model discrimination tests based on our robust designs and those based on "classically optimal" designs—those which entertain no neighbourhood structure on the models, that is, $\varepsilon_0 = \varepsilon_1 = 0$. On this basis, we will compare the methods.

The method requires the experimenter to employ an initial design, of size n_{init}, and to first draw a sample of this size. In our simulations, we have chosen a parameter vector $\boldsymbol{\theta}_{0\text{true}}$ and then simulated the initial, and subsequent, data from either $f_0(y|\eta_0(\mathbf{x}|\boldsymbol{\theta}_{0\text{true}}))$ or from the least favourable member of $\mathcal{F}_1(\varepsilon_1)$. We simulate from f_0 when investigating the sizes of the tests associated with the robust and classically optimal designs. To investigate the minimum powers, we simulate from $f_{1*}(y|\mathbf{x}, \mu_{1*}(\mathbf{x}))$.

The experiment proceeds as follows:

Step 1: Choose an initial design $\boldsymbol{\xi}_0 = \{\xi_{0i}\}_{i=1}^N$ of size $n_{\text{init}} = \sum_{i=1}^N n_{i,\text{init}}$, where $n_{\text{init}} < n$ and $\xi_{0i} = n_{i,\text{init}}/n_{\text{init}}$.

Step 2: Draw $n_{i,\text{init}}$ observations at each covariate $\mathbf{x}_i$.

Carry out steps 3–5, starting with $m = 0$, until an n-point design is obtained.

Step 3: Estimate both parameter vectors $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_1$. In each case, estimation of $\boldsymbol{\theta}_j$ is done assuming that $\mu_j(\mathbf{x}) = \eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$. We denote these estimates by $\widehat{\boldsymbol{\theta}}_m = (\widehat{\boldsymbol{\theta}}_{0m}, \widehat{\boldsymbol{\theta}}_{1m})$.

Step 4: The next design point in the classical design is

$$\mathbf{x}_{\text{new}}^{(c)} = \arg\max_{\mathbf{x}\in\mathcal{S}} \mathcal{I}\{f_0, f_1|\mathbf{x}, \mu_0(\mathbf{x}) = \eta_0(\mathbf{x}|\widehat{\boldsymbol{\theta}}_{0m}), \mu_1(\mathbf{x}) = \eta_1(\mathbf{x}|\widehat{\boldsymbol{\theta}}_{1m})\}.$$

The next design point in the robust design is

$$\mathbf{x}_{\text{new}}^{(r)} = \arg\max_{\mathbf{x}\in\mathcal{S}} \mathcal{I}\{f_0, f_{1*}|\mathbf{x}, \mu_0(\mathbf{x}) = \eta_0(\mathbf{x}|\widehat{\boldsymbol{\theta}}_{0m}), \mu_1(\mathbf{x}) = \eta_1(\mathbf{x}|\widehat{\boldsymbol{\theta}}_{1m})\}.$$

We note, but for ease of presentation do not emphasize, that the estimates $\widehat{\boldsymbol{\theta}}_m$ will depend on the designs used up to this point.

Step 5: Draw y_{new} at the design point $\mathbf{x}_{\text{new}}$.

These steps are illustrated in detail in the examples that follow. Before doing this, we state the related result that, under appropriate conditions, the designs so obtained are asymptotically optimal. We entertain two sequences of models $f_{0n}(y|\mathbf{x}, \mu_0(\mathbf{x}))$, $f_{1n}(y|\mathbf{x}, \mu_1(\mathbf{x}))$ indexed by the sample size n. We assume that

$$\big|\eta_0(\mathbf{x}|\boldsymbol{\theta}_{0n}) - \eta_1(\mathbf{x}|\boldsymbol{\theta}_{1n})\big| = O\big(n^{-1/2}\big).$$

This guarantees, as in Remark 1, that if $f_{0n}(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_{1n}(y|\mathbf{x}, \mu_1(\mathbf{x}))$ are both normal or both lognormal densities with $\mu_0(\mathbf{x}) = \eta_0(\mathbf{x}|\boldsymbol{\theta}_{0n})$, $\mu_1(\mathbf{x}) = \eta_1(\mathbf{x}|\boldsymbol{\theta}_{1n})$ and the same nuisance parameters, then conditions (2.1)–(2.3) in Theorem 2.1 hold. It is also needed for (i) of Theorem 4.1, which is based on Theorem 3.1 in Sinha and Wiens (2003). By this, the LSEs $\widehat{\boldsymbol{\theta}}_{jn}$ updated in each iteration are consistent for sequences $\boldsymbol{\theta}_{jn}$ of parameters defined as

$$\boldsymbol{\theta}_{jn} = \arg\min_{\boldsymbol{\theta}}\Big[\sum_{i=1}^{N}\big\{E[Y_n|\mathbf{x}_i] - \eta_j(\mathbf{x}_i|\boldsymbol{\theta})\big\}^2\Big].$$

In fact, $\widehat{\boldsymbol{\theta}}_{jn} - \boldsymbol{\theta}_{jn} \xrightarrow{\text{a.s.}} 0$ as shown in Sinha and Wiens (2003). With this consistency, we then can obtain that the designs $\{\boldsymbol{\xi}_n\}$ constructed as in Steps 1–5 above are asymptotically optimal. We require assumptions (B1)–(B5) and A3′ as stated in Sinha and Wiens (2003). Moreover, we have two additional assumptions:

(B6) For each fixed $\mathbf{x}$, the KL-divergence in Proposition 2, given by $\mathcal{I}\{f_0, f_{1*}|\mathbf{x}, \mu_0(\mathbf{x}) = \eta_0(\mathbf{x}|\boldsymbol{\theta}_0), \mu_1(\mathbf{x}) = \eta_1(\mathbf{x}|\boldsymbol{\theta}_1)\}$, is Lipschitz continuous with respect to $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$.

(B7) The size of the initial sample n_{init} satisfies $\lim_{n_{\text{init}}\to\infty} n_{\text{init}}/n = 0$.

Sequential, adaptive optimality has been treated elsewhere in the literature. In particular, Wynn (1970) proposed a sequential, but not adaptive, method converging to a D-optimal design in linear models. Wiens and Li (2014) gave a sequential, adaptive estimation method yielding both consistent variance estimates and an asymptotically V-optimal design. Chaudhuri and Mykland (1993) study adaptive designs for nonlinear models and likelihood estimation. Our proof of the following theorem closely parallels those in Wynn (1970) and Wiens and Li (2014).

THEOREM 4.1. *Under assumptions* (B1)–(B7) *and* (A3′), *as* $n_{\text{init}} \to \infty$, *there are sequences* $\{\boldsymbol{\theta}_{jn}\}$ *for which*:

(i) *the LS estimates* $\widehat{\boldsymbol{\theta}}_{jn} - \boldsymbol{\theta}_{jn} \xrightarrow{\text{a.s.}} 0$, $j = 0, 1$, *and*

(ii) $\mathcal{D}(\boldsymbol{\xi}_n, \widehat{\boldsymbol{\theta}}_n) - \max_{\xi\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}_n) \xrightarrow{\text{pr}} 0$ *with* $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\theta}}_{0n}, \widehat{\boldsymbol{\theta}}_{1n})$ *and* $\boldsymbol{\theta}_n = (\boldsymbol{\theta}_{0n}, \boldsymbol{\theta}_{1n})$.

Here, $\mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta})$ *is the KL-divergence between* $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ *and the least favourable density* $f_{1*}(y|\mathbf{x}, \mu_{1*}(\mathbf{x}))$:

$$\mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \xi_i \mathcal{I}\{f_0, f_{1*}|\mathbf{x}_i, \mu_0(\mathbf{x}) = \eta_0(\mathbf{x}|\boldsymbol{\theta}_0), \mu_1(\mathbf{x}) = \eta_1(\mathbf{x}|\boldsymbol{\theta}_1)\},$$

and $\mathcal{P}$ *is the set of all possible n-point designs.*

In the following, we consider several examples in a 51-point design space $\mathcal{S} = \{1, 1.1, \ldots, 4.9, 5\}$, dividing $[1, 5]$ into 50 equal subintervals. Increasing the number of points in the design space may affect the speed of the algorithm (which is slow, due to the onerous computations called for by Proposition 2). But the theory supporting its asymptotic optimality is independent of the number of points in the design space.

4.1. *Example* 1. Assume that both f_0 and f_1 are normal densities with mean $\eta_j(x|\boldsymbol{\theta}_j)$, $j = 0, 1$, and common variance $\sigma^2 = 0.1$. We consider the Michaelis–Menten and exponential response models

$$\eta_0(x|\boldsymbol{\theta}_0) = \frac{V_0 x}{K_0 + x}, \tag{4.1}$$

$$\eta_1(x|\boldsymbol{\theta}_1) = V_1\big(1 - \exp\{-K_1 x\}\big), \tag{4.2}$$

where $\boldsymbol{\theta}_0 = (V_0, K_0)', \boldsymbol{\theta}_1 = (V_1, K_1)'$.

Following steps 1–5 as described above, we obtain robust (or classical) sequential designs with sample size 20. Figure 2 shows a robust design and a classical design obtained in this example.

After a robust (or classical) design, that is, a design measure $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$, is obtained, observations at the design points will be simulated from the "null"

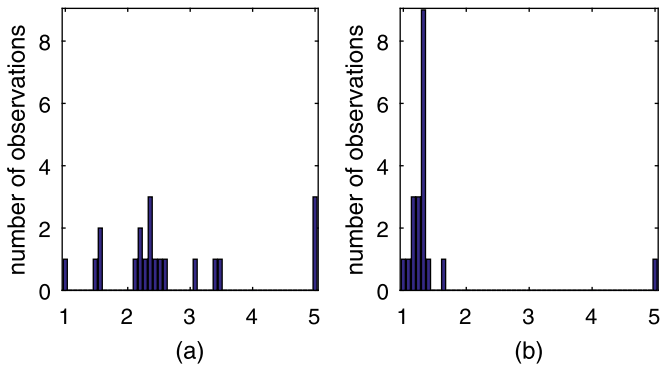

FIG. 2. *In Example* 1, (a) *classical design with sample size* 20 (b) *robust design with sample size* 20 *for* $\varepsilon_1 = 0.17$.

TABLE 1
Simulated sizes and minimum powers (standard errors in parentheses) for Example 1

	ε			
	0.058	**0.063**	**0.1**	**0.17**
Classical				
Size	0.049 (0.0068)	0.044 (0.0065)	0.054 (0.0071)	0.054 (0.0071)
Min-power	0.684 (0.0147)	0.703 (0.0144)	0.681 (0.0147)	0.670 (0.0149)
Robust				
Size	0.055 (0.0072)	0.043 (0.0064)	0.052 (0.0070)	0.052 (0.0070)
Min-power	0.730 (0.0140)	0.711 (0.0143)	0.707 (0.0144)	0.699 (0.0145)

model or the "alternative" model. To investigate the sizes of the tests, we simulate from the "null" model $f_0(y|\eta_0(x|\boldsymbol{\theta}_{0\text{true}}))$. We have used $\boldsymbol{\theta}_{0\text{true}} = (1,1)^T$ in this and the following example. To investigate the minimum powers we simulate from the "alternate" model, the least favourable density $f_{1*}(y|x, \mu_{1*}(x))$. Then the observations are substituted into the test statistic $\mathcal{R}$ and a model discrimination test is performed for the hypotheses

$$H_0: f_0\big(y|\eta_0(x|\boldsymbol{\theta}_0)\big) \quad \text{vs.} \quad H_1: f_1\big(y|\eta_1(x|\boldsymbol{\theta}_1)\big).$$

We simulate 1000 robust and classical designs and do the hypothesis tests of size $\alpha = 0.05$. The number of rejections is counted and the ratio of number of rejections to number of tests is the estimate of the size (if the data are simulated from the null model) or the minimum power (if the data are simulated from the alternate model). The radii of the neighbourhoods $\mathcal{F}_1(\varepsilon_1)$ were $\varepsilon_1 = 0.058, 0.063, 0.1, 0.17$. The simulated results are recorded in Table 1.

According to Table 1, the sizes of model discrimination tests for both classical and robust designs are close to the test size $\alpha = 0.05$. The minimum powers for robust designs are higher than those of classical designs. As the neighbourhood $\mathcal{F}_1(\varepsilon_1)$ is enlarged with respect to ε_1, the minimum powers decrease because the least favourable densities are found in a bigger neighbourhood.

4.2. *Example* 2. Suppose that under each model the observations are log-normal, that is, $\log Y$ is normally distributed. Assume that the logarithm of the observation $\log Y$ has mean $\alpha_j(x)$ and variance $\sigma_j^2(x)$, so that the density of Y is

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

with

$$E_{\text{model } j}[Y|x] = \mu_j(x) = \exp\big(\sigma_j^2(x)/2 + \alpha_j(x)\big),$$

$$\text{var}_{\text{model } j}[Y|x] = v_j^2(x) = \mu_j^2(x)\big\{\exp\big(\sigma_j^2(x)\big) - 1\big\}.$$

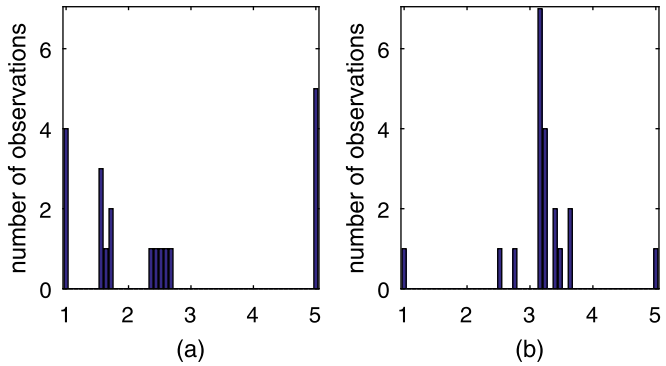

FIG. 3. *In Example* 2, (a) *classical design with sample size* 20 (b) *robust design with sample size* 20 *for* $\varepsilon_1 = 0.17$.

In the following, we assume homoscedastic models and specify the variance function $v_j^2(x) \equiv v^2 = 0.1$.

Let f_0 and f_1 be log-normal densities with means $\eta_0(x|\boldsymbol{\theta}_0)$ and $\eta_1(x|\boldsymbol{\theta}_1)$. Here, $\eta_j(x|\boldsymbol{\theta}_j)$, $j = 0, 1$, are the Michaelis–Menten and exponential response models defined in (4.1) and (4.2), respectively. The robust designs and classical optimal designs can be obtained by following steps 1–5. As an example a robust design and a classical design are illustrated in Figure 3.

To investigate the sizes of the tests, we simulate from the null model $f_0(y|\eta_0(x|\boldsymbol{\theta}_{0\text{true}}))$. To assess the minimum powers, we simulate from $f_{1*}(y|\mathbf{x}, \mu_{1*}(\mathbf{x}))$, the least favourable density in $\mathcal{F}_1(\varepsilon_1)$ (we use the same values of ε_1 as in Example 1). As described in Example 1, we simulate 1000 robust and classical designs and perform model discrimination tests with size $\alpha = 0.05$. The estimates of type-I error and minimum powers are given in Table 2.

The numerical results lead to the same conclusions as in Example 1.

TABLE 2
Simulated sizes and minimum powers (*standard errors in parentheses*) *for Example* 2

	$\boldsymbol{\varepsilon}$			
	0.01	**0.032**	**0.17**	**0.2**
Classical				
Size	0.052 (0.0070)	0.053 (0.0071)	0.0610 (0.0076)	0.043 (0.0064)
Min-power	0.602 (0.0155)	0.641 (0.0152)	0.601 (0.0155)	0.593 (0.0155)
Robust				
Size	0.049 (0.0068)	0.054 (0.0071)	0.050 (0.0069)	0.053 (0.0071)
Min-power	0.649 (0.0151)	0.675 (0.0148)	0.630 (0.0153)	0.623 (0.0153)

5. Summarizing remarks. We have considered the construction of robust model discrimination designs, to aid in the choice of regression models. In the existing literature on discrimination designs, the construction has generally been based on the assumption that the true model is one of the nominal models. Our method instead assumes that the two models range over Hellinger neighbourhoods of the nominal models. The model discrimination problem can be cast as a problem of hypothesis testing. In particular, we have considered the case that under the null hypothesis the "neighbourhood" is a singleton—a fixed density function—and under the alternative hypothesis the density lies in a Hellinger neighbourhood of the hypothesized density. We aimed at constructing experimental designs by maximizing the minimum power, over the Hellinger neighbourhood, of the Neyman–Pearson test. We derived the asymptotic properties of the Neyman–Pearson test statistic and proved that the power of the Neyman–Pearson test is a monotonic function of the Kullback–Leibler divergence between the two rival models under certain conditions. Therefore, we have proposed designs that maximize the minimum KL divergence in the neighbourhood.

The minimization part of this procedure has been carried out analytically; the optimal designs are obtained by maximizing the minimized discrepancy function sequentially and adaptively, with the parameters reestimated after each design point is chosen and implemented. Examples and small samples simulations have given empirical support for the validity of our asymptotic theory.

In the examples, we have assessed the sizes and powers of the post-design hypothesis tests by simulating observations from a "true" model which might differ from both nominal models. To illustrate that our designs are robust against slight deviation from the nominal model, we compared the minimum powers of model discrimination tests based on robust designs and classical designs. This was done for a range of values of ε_1, determining the size of the alternate neighbourhood. Subject to condition (3.1), necessary in order that the hypotheses be separated, we have used several values of ε_1 small enough that the hypothesis are widely separated, and several values large enough that the alternate neighbourhood is quite large. As seen from the results in Tables 1 and 2, the size of the test is quite stable under changes in ε_1, and the (minimum) power is quite robust, relative to that of the test based on nonrobust, "classical" design principles.

APPENDIX: DERIVATIONS

A.1. Proof of Corollary 1. We first show that (2.2) holds, under the conditions in the statement of the Corollary. For arbitrary $(f^{(0)}, f^{(1)})$ and any $\delta > 0$, denote

$$\Xi = \{y : |\log r(y|\mathbf{x}_i; f^{(0)}, f^{(1)})| \geq \delta\}.$$

Then

$$
\begin{aligned}
&\int_{\Xi} f^{(0)}(y|\mathbf{x}_i)\bigl(\sqrt{r\bigl(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\bigr)}-1\bigr)^2\,dy \\
&\quad=\int_{\Xi}\left(\begin{array}{c}[\sqrt{f^{(0)}(y|\mathbf{x}_i)}-\sqrt{f_0\bigl(y|\mathbf{x}_i,\mu_0(\mathbf{x}_i)\bigr)}\,] \\ +[\sqrt{f_0\bigl(y|\mathbf{x}_i,\mu_0(\mathbf{x}_i)\bigr)}-\sqrt{f_1\bigl(y|\mathbf{x}_i,\mu_1(\mathbf{x}_i)\bigr)}\,] \\ +[\sqrt{f_1\bigl(y|\mathbf{x}_i,\mu_0(\mathbf{x}_i)\bigr)}-\sqrt{f^{(1)}(y|\mathbf{x}_i)}\,]\end{array}\right)^2 dy \\
&\quad\le 3\sum_{j=0,1}\int_{\Xi}\bigl(\sqrt{f^{(j)}(y|\mathbf{x}_i)}-\sqrt{f_j\bigl(y|\mathbf{x}_i,\mu_j(\mathbf{x}_i)\bigr)}\bigr)^2 dy \\
&\qquad+3\int_{\Xi}\bigl(\sqrt{f_0\bigl(y|\mathbf{x}_i,\mu_0(\mathbf{x}_i)\bigr)}-\sqrt{f_1\bigl(y|\mathbf{x}_i,\mu_1(\mathbf{x}_i)\bigr)}\bigr)^2 dy.
\end{aligned}
\tag{A.1}
$$

We show that each term in (A.1) is $o(n^{-1})$. Since $\varepsilon_0 = o(n^{-1/2})$, $\varepsilon_1 = o(n^{-1/2})$, the first term is $o(n^{-1})$ for each j, and

$$
\lim_{n\longrightarrow\infty}\sum_{i=1}^{N} n_i\int_{\Xi}\bigl(\sqrt{f^{(j)}(y|\mathbf{x}_i)}-\sqrt{f_j\bigl(y|\mathbf{x}_i,\mu_j(\mathbf{x}_i)\bigr)}\bigr)^2 dy=0, \qquad j=0,1. \tag{A.2}
$$

With

$$
\Xi_1=\left\{|\log r\bigl(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\bigr)|\ge\delta \text{ and } |\log r(y|\mathbf{x}_i; f_0, f_1)|\ge\frac{\delta}{n}\right\},
$$

$$
\Xi_2=\left\{|\log r\bigl(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\bigr)|\ge\delta \text{ and } |\log r(y|\mathbf{x}_i; f_0, f_1)|<\frac{\delta}{n}\right\},
$$

the second term in (A.1) can be divided into two terms:

$$
\begin{aligned}
&\int_{\Xi}\bigl(\sqrt{f_0\bigl(y|\mathbf{x}_i,\mu_0(\mathbf{x}_i)\bigr)}-\sqrt{f_1\bigl(y|\mathbf{x}_i,\mu_1(\mathbf{x}_i)\bigr)}\bigr)^2 dy \\
&\quad=\int_{\Xi_1}\bigl(\sqrt{f_0\bigl(y|\mathbf{x}_i,\mu_0(\mathbf{x}_i)\bigr)}-\sqrt{f_1\bigl(y|\mathbf{x}_i,\mu_1(\mathbf{x}_i)\bigr)}\bigr)^2 dy \\
&\qquad+\int_{\Xi_2}\bigl(\sqrt{f_0\bigl(y|\mathbf{x}_i,\mu_0(\mathbf{x}_i)\bigr)}-\sqrt{f_1\bigl(y|\mathbf{x}_i,\mu_1(\mathbf{x}_i)\bigr)}\bigr)^2 dy.
\end{aligned}
\tag{A.3}
$$

Notice that $\Xi_1\subseteq\Xi_3=\{|\log r(y|\mathbf{x}_i; f_0, f_1)|\ge\frac{\delta}{n}\}$. Then the first term in (A.3) satisfies

$$
\begin{aligned}
&\int_{\Xi_1}\bigl(\sqrt{f_0\bigl(y|\mathbf{x}_i,\mu_0(\mathbf{x}_i)\bigr)}-\sqrt{f_1\bigl(y|\mathbf{x}_i,\mu_1(\mathbf{x}_i)\bigr)}\bigr)^2 dy \\
&\quad\le\int_{\Xi_3}\bigl(\sqrt{f_0\bigl(y|\mathbf{x}_i,\mu_0(\mathbf{x}_i)\bigr)}-\sqrt{f_1\bigl(y|\mathbf{x}_i,\mu_1(\mathbf{x}_i)\bigr)}\bigr)^2 dy.
\end{aligned}
$$

Moreover, since $f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))$ and $f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))$ satisfy (2.2), we have

$$\begin{aligned}&\lim_{n\longrightarrow\infty}\sum_{i=1}^{N} n_i \int_{\Xi_1} \bigl(\sqrt{f_0\bigl(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i)\bigr)} - \sqrt{f_1\bigl(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i)\bigr)}\bigr)^2 dy \\ &\quad\le \lim_{n\longrightarrow\infty}\sum_{i=1}^{N} n_i \int_{\Xi_3} \bigl(\sqrt{f_0\bigl(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i)\bigr)} - \sqrt{f_1\bigl(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i)\bigr)}\bigr)^2 dy = 0.\end{aligned} \tag{A.4}$$

For the second term in (A.3), noticing that

$$\bigl|\log r(y|\mathbf{x}_i; f_0, f_1)\bigr| < \frac{\delta}{n} \quad\Leftrightarrow\quad e^{-\delta/n} \le r(y|\mathbf{x}_i; f_0, f_1) \le e^{\delta/n},$$

we have

$$\begin{aligned}&\int_{\Xi_2} \bigl(\sqrt{f_0\bigl(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i)\bigr)} - \sqrt{f_1\bigl(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i)\bigr)}\bigr)^2 dy \\ &\qquad\le \max\bigl\{\bigl(1 - e^{-\delta/2n}\bigr)^2, \bigl(1 - e^{\delta/2n}\bigr)^2\bigr\} = o\bigl(n^{-1}\bigr),\end{aligned}$$

and then

$$\lim_{n\longrightarrow\infty}\sum_{i=1}^{N} n_i \int_{\Xi_2} \bigl(\sqrt{f_0\bigl(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i)\bigr)} - \sqrt{f_1\bigl(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i)\bigr)}\bigr)^2 dy = 0. \tag{A.5}$$

Therefore, combining (A.2), (A.4) and (A.5) we have that

$$\lim_{n\longrightarrow\infty}\sum_{i=1}^{N} n_i \int_{\Xi} f^{(0)}(y|\mathbf{x}_i)\bigl(r\bigl(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\bigr) - 1\bigr)^2 dy = 0,$$

that is, $f^{(0)}(y|\mathbf{x})$ and $f^{(1)}(y|\mathbf{x})$ satisfy (2.2).

To prove that $f^{(0)}(y|\mathbf{x})$ and $f^{(1)}(y|\mathbf{x})$ satisfy (2.1), first write

$$\begin{aligned}n\mathcal{D} &= \sum_{i=1}^{N} n_i \int_{\Xi} f^{(1)}(y|\mathbf{x}_i) \log r\bigl(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\bigr) dy \\ &\quad + \sum_{i=1}^{N} n_i \int_{\Xi^c} f^{(1)}(y|\mathbf{x}_i) \log r\bigl(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\bigr) dy.\end{aligned} \tag{A.6}$$

We will prove that the limit of the first term in (A.6) is 0 as $n \to \infty$ and the limit of the second term is finite. By the triangle inequality for the Hellinger distance, we have

$$\begin{aligned}d_h^2\bigl(f^{(0)}(y|\mathbf{x}_i), f^{(1)}(y|\mathbf{x}_i)\bigr) &\le \varepsilon_0^2 + \varepsilon_1^2 + d_h^2\bigl(f_0\bigl(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i)\bigr), f_1\bigl(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i)\bigr)\bigr) \\ &\le O\bigl(n^{-1}\bigr).\end{aligned} \tag{A.7}$$

According to Oosterhoff and van Zwet (2012), (A.7) and condition (2.2) imply that $\{\Pi_{i=1}^{N}(F_i^{(0)})^{n_i}\}$ and $\{\Pi_{i=1}^{N}(F_i^{(1)})^{n_i}\}$ are contiguous with respect to each other, where $F_i^{(0)}$ and $F_i^{(1)}$ are the distributions corresponding to $f^{(0)}(y|\mathbf{x}_i)$ and $f^{(1)}(y|\mathbf{x}_i)$, respectively. Therefore, we conclude that for all $\delta > 0$,

$$\begin{aligned}
&\lim_{n\longrightarrow\infty}\sum_{i=1}^{N} n_i F_i^{(1)}\big(\big|\log r\big(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\big)\big| \geq \delta\big) = 0,\\
&\lim_{n\longrightarrow\infty}\sum_{i=1}^{N} n_i F_i^{(0)}\big(\big|\log r^{-1}\big(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\big)\big| \geq \delta\big) = 0.
\end{aligned} \tag{A.8}$$

Moreover, according to the contiguity of $\{\Pi_{i=1}^{N}(F_i^{(0)})^{n_i}\}$ and $\{\Pi_{i=1}^{N}(F_i^{(1)})^{n_i}\}$, we have

$$\lim_{n\longrightarrow\infty}\sum_{i=1}^{N} n_i F_i^{(0)}\big(\big|\log r\big(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\big)\big| \geq \delta\big) = 0. \tag{A.9}$$

Then by similar analysis as in the proof of Theorem 2.1 and using condition (2.3), the limit of the first term in (A.6) is 0. To prove that the second term in (A.6) has a finite limit, we expand $\log r(y|\mathbf{x}_i)$ [as in the proof of Lemma 2.5 in Hu and Wiens (2017)] and obtain

$$\begin{aligned}
&\Big|\sum_{i=1}^{N} n_i \int_{\Xi^c} f^{(1)}(y|\mathbf{x}_i)\log r\big(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\big)\,dy\Big|\\
&\quad\leq \left\{\begin{array}{l}
2\sum_{i=1}^{N} n_i \int_{\Xi^c}\big(\sqrt{f^{(1)}(y|\mathbf{x}_i)} - \sqrt{f^{(0)}(y|\mathbf{x}_i)}\big)^2\,dy\\
\quad + \Big|\sum_{i=1}^{N} n_i \int_{\Xi^c}\big(f^{(0)}(y|\mathbf{x}_i) - f^{(1)}(y|\mathbf{x}_i)\big)\,dy\Big|\\
+ \Big|\sum_{i=1}^{N} n_i \int_{\Xi^c}\rho_{1i\delta}\big(\sqrt{f^{(1)}(y|\mathbf{x}_i)} - \sqrt{f^{(0)}(y|\mathbf{x}_i)}\big)^2\,dy\Big|
\end{array}\right\}\\
&\quad\leq \left\{\begin{array}{l}
4\sum_{i=1}^{N} n_i d_h^2\big(f^{(0)}(y|\mathbf{x}_i), f^{(1)}(y|\mathbf{x}_i)\big)\\
+ \Big|\sum_{i=1}^{N} n_i \int_{\Xi^c}\big(f^{(0)}(y|\mathbf{x}_i) - f^{(1)}(y|\mathbf{x}_i)\big)\,dy\Big|\\
\quad + 3\delta\sum_{i=1}^{N} n_i d_h^2\big(f^{(0)}(y|\mathbf{x}_i), f^{(1)}(y|\mathbf{x}_i)\big)
\end{array}\right\}\\
&\quad = O(1).
\end{aligned} \tag{A.10}$$

Here, the second term in (A.10) satisfies

$$\begin{aligned}&\left|\sum_{i=1}^{N} n_i \int_{\Xi^c} \left(f^{(0)}(y|\mathbf{x}_i) - f^{(1)}(y|\mathbf{x}_i)\right) dy\right| \\ &\quad = \left|\sum_{i=1}^{N} n_i \int_{\Xi} \left(f^{(1)}(y|\mathbf{x}_i) - f^{(0)}(y|\mathbf{x}_i)\right) dy\right| \to 0\end{aligned} \tag{A.11}$$

due to (A.8) and (A.9). Combining (A.10) and (A.11), we can conclude that $f^{(0)}(y|\mathbf{x})$ and $f^{(1)}(y|\mathbf{x})$ satisfy (2.1).

A.2. Proof of Theorem 2.2. We prove only (i); (ii) can be proved in a similar manner. According to Theorem 2.1, if $f(y|\mathbf{x}) \in \mathcal{F}_1(\varepsilon_1)$ is the true model, the test statistic

$$\mathcal{R}(f_0, f) = 2\sum_{i=1}^{N}\sum_{l=1}^{n_i} \log\left\{\frac{f(y_{il}|\mathbf{x}_i)}{f_0(y_{il}|\mathbf{x}_i, \mu_0(\mathbf{x}_i))}\right\}$$

is normally distributed with mean $-2n\mathcal{D}(f_0, f)$ and standard deviation $\sqrt{8n\mathcal{D}(f_0, f)}$. Recall that

$$\mathcal{R} := \mathcal{R}(f_0, f_1) = 2\sum_{i=1}^{N}\sum_{l=1}^{n_i} \log\left\{\frac{f_1(y_{il}|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}{f_0(y_{il}|\mathbf{x}_i, \mu_0(\mathbf{x}_i))}\right\}.$$

Notice that

$$\mathcal{R} = \mathcal{R}(f_0, f) - 2z_n$$

with

$$z_n = \sum_{i,l} \log\left\{\frac{f(y_{il}|\mathbf{x}_i)}{f_1(y_{il}|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right\},$$

and $n\mathcal{D}(f_0, f) = O(1)$ according to Corollary 1. Therefore, if we can prove that when $f(y|\mathbf{x}) \in \mathcal{F}_1(\varepsilon_1)$ is the true model

$$z_n = o_p(1), \tag{A.12}$$

then the asymptotic normality of $\mathcal{R}$ is proved. With the notation in Theorem 2.1, to prove (A.12) under $f(y|\mathbf{x})$, we need to show that for any $\epsilon > 0$,

$$\lim_{n\to\infty} F^{(n)}\left(\left|\sum_{i,l} \log\left\{\frac{f(Y_{il}|\mathbf{x}_i)}{f_1(Y_{il}|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right\}\right| > \epsilon\right) = 0. \tag{A.13}$$

Notice that

$$F^{(n)}\left(\left|\sum_{i,l}\log\left\{\frac{f(Y_{il}|\mathbf{x}_i)}{f_1(Y_{il}|\mathbf{x}_i,\mu_1(\mathbf{x}_i))}\right\}\right|>\epsilon\right)$$

$$\le F^{(n)}\left(\sum_{i,l}\left|\log\left\{\frac{f(Y_{il}|\mathbf{x}_i)}{f_1(Y_{il}|\mathbf{x}_i,\mu_1(\mathbf{x}_i))}\right\}\right|>\epsilon\right) \tag{A.14}$$

$$\le F^{(n)}\left(\max_{i,l}\left|\log\left\{\frac{f(Y_{il}|\mathbf{x}_i)}{f_1(Y_{il}|\mathbf{x}_i,\mu_1(\mathbf{x}_i))}\right\}\right|>\frac{\epsilon}{n}\right).$$

Then if we can prove that

$$\sum_{i,l}F^{(n)}\left(\left|\log\left\{\frac{f(Y_{il}|\mathbf{x}_i)}{f_1(Y_{il}|\mathbf{x}_i,\mu_1(\mathbf{x}_i))}\right\}\right|>\frac{\epsilon}{n}\right)\to 0, \tag{A.15}$$

according to (A.14), (A.13) holds.

Notice that the Hellinger distance between $f_1(y|\mathbf{x}_i,\mu_1(\mathbf{x}_i))$ and $f(y|\mathbf{x}_i)$ is at most $\varepsilon_1 = o(n^{-1/2})$. Then

$$\sum_{i=1}^{N} n_i\int\left(\sqrt{f(y|\mathbf{x}_i)}-\sqrt{f_1\big(y|\mathbf{x}_i,\mu_1(\mathbf{x}_i)\big)}\right)^2 dy = o(1).$$

Based on (1.5) in Oosterhoff and van Zwet (2012), we have that $F_1^{(n)}$ and $F^{(n)}$ are mutually contiguous.

Because of the contiguity, according to the proof of Theorem 2 in Oosterhoff and van Zwet (2012), to prove (A.15), it is equivalent to prove that for any $\epsilon>0$

$$\lim_{n\longrightarrow\infty}\sum_{i=1}^{N} n_i\int_{\{|\log r(y|\mathbf{x}_i;f,f_1)|\ge\frac{\epsilon}{n}\}}\left(\sqrt{f(y|\mathbf{x}_i)}-\sqrt{f_1\big(y|\mathbf{x}_i,\mu_1(\mathbf{x}_i)\big)}\right)^2 dy = 0.$$

This follows, since

$$\lim_{n\longrightarrow\infty}\sum_{i=1}^{N} n_i\int_{\{|\log r(y|\mathbf{x}_i;f,f_1)|\ge\frac{\epsilon}{n}\}}\left(\sqrt{f(y|\mathbf{x}_i)}-\sqrt{f_1\big(y|\mathbf{x}_i,\mu_1(\mathbf{x}_i)\big)}\right)^2 dy$$

$$\le \lim_{n\longrightarrow\infty} 2n\varepsilon_1^2 = 0.$$

This completes the proof of (i).

A.3. Proof of Proposition 1. For an arbitrary $f\in\mathcal{F}_1(\varepsilon_1)$ set $t=\mathcal{D}(f_0,f)$, and define $t_0=\mathcal{D}(f_0,f_{1*})$, so that, by definition and assumption,

$$t\ge t_0\ge -c. \tag{A.16}$$

In this notation, we are to show that $\Phi(\frac{-c+t_0}{2\sqrt{t_0}})\le\Phi(\frac{-c+t}{2\sqrt{t}})$, that is, that

$$\frac{-c+t_0}{\sqrt{t_0}}\le\frac{-c+t}{\sqrt{t}}\qquad\text{for } t\ge t_0.$$

After a rearrangement, this condition becomes

$$-c \le \sqrt{t_0}\sqrt{t}.$$

This is obvious if $c \ge 0$, otherwise it follows from (A.16).

A.4. Proof of Proposition 2. In the following, we write $f(y)$, $f_0(y|\mu_0)$, $f_1(y|\mu_1)$ for $f(y|\mathbf{x})$, $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$, $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$. Define

$$\begin{aligned}&\mathcal{L}\big(f(y), \lambda_1, \lambda_2\big)\\ &\quad = f(y)\log\frac{f(y)}{f_0(y|\mu_0)} + \lambda_1\sqrt{f(y)f_1(y|\mu_1)} + \lambda_2 f(y) \qquad \text{for } y \in \Omega_{\mathbf{x}}.\end{aligned}$$

For each fixed $y \in \Omega_{\mathbf{x}}$, the function $\mathcal{L}(f(y), \lambda_1, \lambda_2)$ is convex with respect to $f(y) > 0$. It follows that the critical point which is a solution to (3.4) is a minimizer of $\mathcal{L}(f(y), \lambda_1, \lambda_2)$. Then the solution to (3.4)–(3.6) is also the solution to the optimality problem (3.3), as we now show. Assume that $(f_{1*}(y|\mu_{1*}), \lambda_1, \lambda_2)$ is a solution to the equation system. For any $f(y)$ such that $f(y)$ vanishes on $\Omega_{\mathbf{x}}^c$ and satisfies the constraints of the optimization problem (3.3), it is clear that

$$\mathcal{L}\big(f(y), \lambda_1, \lambda_2\big) \ge \mathcal{L}\big(f_{1*}(y), \lambda_1, \lambda_2\big),$$

that is,

$$\begin{aligned}&f(y)\log\bigg(\frac{f(y)}{f_0(y|\mu_0)}\bigg) + \lambda_1\sqrt{f(y)f_1(y|\mu_1)} + \lambda_2 f(y)\\ &\quad \ge f_{1*}(y|\mu_{1*})\log\bigg(\frac{f_{1*}(y|\mu_{1*})}{f_0(y|\mu_0)}\bigg) + \lambda_1\sqrt{f_{1*}(y|\mu_{1*})f_1(y|\mu_1)}\\ &\qquad + \lambda_2 f_{1*}(y|\mu_{1*}),\end{aligned}$$

and then

$$\begin{aligned}\mathcal{I}\{f_0, f|\mu_0\} &\ge \int_{\Omega_{\mathbf{x}}} f_{1*}(y|\mu_{1*})\log\bigg(\frac{f_{1*}(y|\mu_{1*})}{f_0(y|\mu_0)}\bigg)dy\\ &\quad + \lambda_1\int_{\Omega_{\mathbf{x}}}\big(\sqrt{f_{1*}(y|\mu_{1*})f_1(y|\mu_1)} - \sqrt{f(y)f_1(y|\mu_1)}\big)dy\\ &= \mathcal{I}\{f_0, f_{1*}|\mu_0, \mu_{1*}\}\\ &\quad + \lambda_1\int_{\Omega_{\mathbf{x}}}\big(\sqrt{f_{1*}(y|\mu_{1*})f_1(y|\mu_1)} - \sqrt{f(y)f_1(y|\mu_1)}\big)dy\\ &\ge \mathcal{I}\{f_0, f_{1*}|\mu_0, \mu_{1*}\},\end{aligned}$$

since

$$\lambda_1\int_{\Omega_{\mathbf{x}}}\big(\sqrt{f_{1*}(y|\mu_{1*})f_1(y|\mu_1)} - \sqrt{f(y)f_1(y|\mu_1)}\big)dy \ge 0.$$

Therefore, the solution to equations (3.4), (3.5) and (3.6) is the minimizer of the optimality problem (3.3).

The multiplier λ_1 is strictly negative. For, if $\lambda_1 = 0$ then according to (3.4) we have $f_{1*}(y|\mu_{1*}) = f_0(y|\mu_0)\exp\{-1-\lambda_2\}$. However, constraint (3.5) then implies $\lambda_2 = -1$ and $f_{1*}(y|\mu_{1*}) = f_0(y|\mu_0)$. But then constraint (3.6) cannot be satisfied, since $\min_{\mathbf{x}\in\mathcal{S}} d_h(f_0, f_1|\mathbf{x}) > \varepsilon_1$. Therefore, $\lambda_1 < 0$.

Finally, by using the constraints (3.5) and (3.6), the minimum of the optimality problem (3.3) can be simplified:

$$\mathcal{I}\{f_0, f_{1*}|\mu_0, \mu_{1*}\} = -1 - \frac{1}{2}\lambda_1\bigl(1-\varepsilon_1^2\bigr) - \lambda_2.$$

A.5. Proof of Theorem 4.1. (i) The consistency of the LS estimates is a direct result of Theorem 3.1 in Sinha and Wiens (2003).

(ii) We prove that $\mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_n) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}_n) \xrightarrow{\text{pr}} 0$ by verifying that

(E1) $\max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_n) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}_n) \xrightarrow{\text{pr}} 0$ as $n_{\text{init}} \to \infty$,

(E2) $\mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_n) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_n) \xrightarrow{\text{pr}} 0$ as $n_{\text{init}} \to \infty$.

We first prove (E1). Let $\boldsymbol{\xi}_n^*$ be a design such that $\mathcal{D}(\boldsymbol{\xi}_n^*, \widehat{\boldsymbol{\theta}}_n) = \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_n)$ and let $\boldsymbol{\xi}_{n0}$ be the design such that $\mathcal{D}(\boldsymbol{\xi}_{n0}, \boldsymbol{\theta}_n) = \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}_n)$. Then

$$\begin{aligned} L_n &:= \mathcal{D}(\boldsymbol{\xi}_{0n}, \widehat{\boldsymbol{\theta}}_n) - \mathcal{D}(\boldsymbol{\xi}_{0n}, \boldsymbol{\theta}_n) \le \mathcal{D}\bigl(\boldsymbol{\xi}_n^*, \widehat{\boldsymbol{\theta}}_n\bigr) - \mathcal{D}(\boldsymbol{\xi}_{0n}, \boldsymbol{\theta}_n) \\ &\le \mathcal{D}\bigl(\boldsymbol{\xi}_n^*, \widehat{\boldsymbol{\theta}}_n\bigr) - \mathcal{D}\bigl(\boldsymbol{\xi}_n^*, \boldsymbol{\theta}_n\bigr) =: U_n. \end{aligned}$$

Recall that

$$\mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \xi_i \mathcal{I}\bigl\{f_0, f_{1*}|\mathbf{x}_i, \eta_0(\mathbf{x}_i|\boldsymbol{\theta}_0), \eta_1(\mathbf{x}_i|\boldsymbol{\theta}_1)\bigr\}.$$

According to condition (B6), the integrand $\mathcal{I}\{f_0, f_{1*}|\mathbf{x}, \eta_0(\mathbf{x}|\boldsymbol{\theta}_0), \eta_1(\mathbf{x}|\boldsymbol{\theta}_1)\}$ is Lipschitz continuous with respect to $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$. Via the consistency of $\widehat{\boldsymbol{\theta}}_n$, and the linearity of $\mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta})$ with respect to ξ we have that for any design ξ,

$$\begin{aligned} &\bigl|\mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_n) - \mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}_n)\bigr| \\ &\quad\le \max_{i=1,\ldots,N} \left| \begin{array}{l} \mathcal{I}\bigl\{f_0, f_{1*}|\mathbf{x}_i, \eta_0(\mathbf{x}_i|\widehat{\boldsymbol{\theta}}_{0n}), \eta_1(\mathbf{x}_i|\widehat{\boldsymbol{\theta}}_{1n})\bigr\} \\ -\mathcal{I}\bigl\{f_0, f_1|\mathbf{x}_i, \eta_0(\mathbf{x}_i|\boldsymbol{\theta}_{0n}), \eta_1(\mathbf{x}_i|\boldsymbol{\theta}_{1n})\bigr\} \end{array} \right| \xrightarrow{\text{a.s.}} 0. \end{aligned} \tag{A.17}$$

Therefore, $L_n, U_n \xrightarrow{\text{a.s.}} 0$ and (E1) follows.

To prove (E2), we first write

$$\begin{aligned} &\mathcal{D}\bigl(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_n\bigr) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_n) \\ &= \bigl(\mathcal{D}\bigl(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_n\bigr) - \mathcal{D}\bigl(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr)\bigr) + \Bigl(\mathcal{D}\bigl(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}})\Bigr) \\ &\quad + \Bigl(\max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_n)\Bigr). \end{aligned} \tag{A.18}$$

The first and last terms in (A.18) converge to 0 in probability as $n_{\text{init}} \to \infty$, due to (A.17) and (E1). Then it suffices to prove that, for any $\varepsilon > 0$,

$$\Pr\Big(\mathcal{D}\big(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}\big(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) \geq -\varepsilon\Big) \to 1. \tag{A.19}$$

Recall that given the $(n-1)$th design $\boldsymbol{\xi}^{(n-1)}$ and the estimates $\widehat{\boldsymbol{\theta}}_n$, the next design point is

$$\mathbf{x}_{\text{new}} = \arg \max_{i=1,\ldots,N} \mathcal{I}\big\{f_0, f_{1*}|\mathbf{x}_i, \eta_0(\mathbf{x}_i|\widehat{\boldsymbol{\theta}}_{0n}), \eta_1(\mathbf{x}_i|\widehat{\boldsymbol{\theta}}_{1n})\big\}.$$

Then the nth design is

$$\boldsymbol{\xi}^{(n)} = \frac{n-1}{n}\boldsymbol{\xi}^{(n-1)} + \frac{1}{n}\delta_n(\mathbf{x}),$$

where $\delta_n(\mathbf{x}) = I(\mathbf{x} = \mathbf{x}_{\text{new}})$. Therefore, the KL-divergence for the nth design is

$$\mathcal{D}\big(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) = \frac{n-1}{n}\mathcal{D}\big(\boldsymbol{\xi}^{(n-1)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) + \frac{1}{n}\mathcal{D}\big(\delta_n(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big),$$

and the difference between the KL-divergence with the nth design and that with $(n-1)$th design is

$$\mathcal{D}\big(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) - \mathcal{D}\big(\boldsymbol{\xi}^{(n-1)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) = \frac{\mathcal{D}(\delta_n(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}) - \mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}})}{n-1}. \tag{A.20}$$

Now to establish (A.19), denote $\boldsymbol{\xi}^*_{\text{init}} = \arg\max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}})$. For any $\varepsilon > 0$, divide the sequence $\{\boldsymbol{\xi}^{(n)}\}$ into two disjoint subsequences $S_1(\varepsilon)$ and $S_2(\varepsilon)$ such that

$$\begin{aligned} S_1(\varepsilon) &:= \big\{\boldsymbol{\xi}^{(n)} : \mathcal{D}\big(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) \geq \mathcal{D}\big(\boldsymbol{\xi}^*_{\text{init}}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) - \varepsilon/2\big\}, \\ S_2(\varepsilon) &:= \big\{\boldsymbol{\xi}^{(n)} : \mathcal{D}\big(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) < \mathcal{D}\big(\boldsymbol{\xi}^*_{\text{init}}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) - \varepsilon/2\big\}. \end{aligned}$$

We first show that $S_1(\varepsilon)$ is nonempty for each $\varepsilon > 0$. If not, there must exist an ε such that for any n we have

$$\mathcal{D}\big(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) < \mathcal{D}\big(\boldsymbol{\xi}^*_{\text{init}}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) - \varepsilon/2.$$

Then according to (A.20), and with $z_{1n} = (\mathcal{D}(\delta_n(\mathbf{x}); \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}) - \mathcal{D}(\delta_n(\mathbf{x}); \widehat{\boldsymbol{\theta}}_n)) + (\mathcal{D}(\boldsymbol{\xi}^*_{\text{init}}; \widehat{\boldsymbol{\theta}}_n) - \mathcal{D}(\boldsymbol{\xi}^*_{\text{init}}; \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}))$, we have that

$$\begin{aligned} &\mathcal{D}\big(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) - \mathcal{D}\big(\boldsymbol{\xi}^{(n-1)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) \\ &\quad > \frac{\mathcal{D}(\delta_n(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}) - \mathcal{D}(\boldsymbol{\xi}^*_{\text{init}}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}) + \varepsilon/2}{n-1} \\ &\quad = \frac{\varepsilon}{2(n-1)} + \frac{z_{1n}}{n-1} + \frac{\mathcal{D}(\boldsymbol{\xi}^*_{\text{init}}, \widehat{\boldsymbol{\theta}}_n) - \mathcal{D}(\boldsymbol{\xi}^*_{\text{init}}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}})}{n-1} \\ &\quad \geq \frac{\varepsilon}{2(n-1)} + \frac{z_{1n}}{n-1}, \end{aligned} \tag{A.21}$$

since $\mathcal{D}(\delta_n(\mathbf{x}), \widehat{\boldsymbol{\theta}}_n) \geq \mathcal{D}(\xi^*_{\text{Init}}, \widehat{\boldsymbol{\theta}}_n)$ by the definition of $\delta_n(\mathbf{x})$.

We can prove $z_{1n} \overset{\text{a.s.}}{\to} 0$ by applying (A.17). According to the proof of Theorem 3.1(i) in Sinha and Wiens (2003), there exists small enough q (>0) such that $\lim_{n\to\infty} n^q(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) < \infty$ almost surely. Then because of condition (B6) [or (B6′)], we also have $\lim_{n\to\infty} n^q z_{1n} < \infty$ almost surely. Moreover, since z_{1n} is bounded by the maximum KL-divergence, we have

$$\sum_{m=1}^{\infty} \frac{z_{1m}}{m} < \infty \qquad \text{a.s.} \tag{A.22}$$

Since $n_{\text{init}}/n \to 0$ as $n_{\text{init}} \to \infty$, we have

$$\begin{aligned}\mathcal{D}\bigl(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr) &= \mathcal{D}\bigl(\boldsymbol{\xi}^{(n_{\text{init}})}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr) + \sum_{m=n_{\text{Init}}+1}^{n} \bigl(\mathcal{D}\bigl(\boldsymbol{\xi}^{(m)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr) - \mathcal{D}\bigl(\boldsymbol{\xi}^{(m-1)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr)\bigr) \\ &> \mathcal{D}\bigl(\boldsymbol{\xi}^{(n_{\text{init}})}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr) + \sum_{m=n_{\text{Init}}+1}^{n} \left(\frac{\varepsilon}{2(m-1)} + \frac{z_{1m}}{m}\right) \overset{\text{a.s.}}{\to} \infty\end{aligned}$$

as $n_{\text{init}} \to \infty$, a contradiction to the assumption that the maximum KL-divergence is finite. Therefore, for any $\varepsilon > 0$, $S_1(\varepsilon)$ is nonempty and we can find a sequence $\{\boldsymbol{\xi}^{(n_l)}\}_{l=1}^{\infty} \subset S_1(\varepsilon)$, that is, $\mathcal{D}(\boldsymbol{\xi}^{(n_l)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}})$ arbitrarily close to $\mathcal{D}(\boldsymbol{\xi}^*_{\text{init}}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}})$. By (A.20), we have

$$\begin{aligned}&\mathcal{D}\bigl(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr) \\ &\quad= \mathcal{D}\bigl(\boldsymbol{\xi}^{(n_l)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr) + \frac{\mathcal{D}(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}) - \mathcal{D}\bigl(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr)}{n_l} \\ &\quad= \mathcal{D}\bigl(\boldsymbol{\xi}^{(n_l)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr) + \frac{\mathcal{D}(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_l+1}) - \mathcal{D}\bigl(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_l+1}\bigr)}{n_l} \\ &\qquad+ \frac{\mathcal{D}(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}) - \mathcal{D}(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_l+1})}{n_l} \\ &\qquad+ \frac{\mathcal{D}\bigl(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_l+1}\bigr) - \mathcal{D}\bigl(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr)}{n_l} \\ &\quad= \mathcal{D}\bigl(\boldsymbol{\xi}^{(n_l)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr) + \frac{\mathcal{D}(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_l+1}) - \mathcal{D}\bigl(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_l+1}\bigr)}{n_l} + \frac{z_{2n_l}}{n_l} \\ &\quad\geq \mathcal{D}\bigl(\boldsymbol{\xi}^*_{\text{init}}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr) - \frac{\varepsilon}{2} + \frac{z_{2n_l}}{n_l},\end{aligned}$$

for $\boldsymbol{\xi}^{(n_l+1)} \in S_1(\varepsilon)$ or $S_2(\varepsilon)$, where

$$\begin{aligned}z_{2n} = &\bigl(\mathcal{D}\bigl(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr) - \mathcal{D}\bigl(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_l+1}\bigr)\bigr) \\ &+ \bigl(\mathcal{D}\bigl(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_l+1}\bigr) - \mathcal{D}\bigl(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\bigr)\bigr)\end{aligned}$$

and, similar to z_{1n}, we also have $z_{2n} \to 0$ a.s.

In summary, as in (A.21), for all $\boldsymbol{\xi}^{(n_k)} \in S_2(\varepsilon)$ we have

$$\mathcal{D}\big(\boldsymbol{\xi}^{(n_k)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) > \mathcal{D}\big(\boldsymbol{\xi}^{(n_k-1)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) + \frac{\varepsilon}{2(n_k-1)} + \frac{z_{1n_k}}{n_k} > \mathcal{D}\big(\boldsymbol{\xi}^{(n_k-1)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) + \frac{z_{1n_k}}{n_k};$$

iterating this gives

$$\begin{aligned}
\mathcal{D}\big(\boldsymbol{\xi}^{(n_k)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) &> \mathcal{D}\big(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) + \sum_{m=n_l+2}^{n_k} \frac{z_{1m}}{m} \\
&> \mathcal{D}\big(\boldsymbol{\xi}^{*}_{\text{init}}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) - \frac{\varepsilon}{2} + \frac{z_{2n_l}}{n_l} + \sum_{m=n_l+2}^{n_k} \frac{z_{1m}}{m} \\
&= \mathcal{D}\big(\boldsymbol{\xi}^{*}_{\text{init}}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) - \frac{\varepsilon}{2} + Y_{n_k},
\end{aligned}$$

with

$$Y_n = \begin{cases} \dfrac{z_{2n_l}}{n_l}, & \text{if } n = n_l + 1, \\ \dfrac{z_{2n_l}}{n_l} + \displaystyle\sum_{m=n_l+2}^{n_k} \frac{z_{1m}}{m}, & \text{if } n \neq n_l + 1. \end{cases}$$

Notice that $Y_n \xrightarrow{\text{pr}} 0$ by (A.22). Therefore, for any $\varepsilon > 0$,

$$\Pr\big(\mathcal{D}\big(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) - \mathcal{D}\big(\boldsymbol{\xi}^{*}_{\text{init}}, \widehat{\boldsymbol{\theta}}_{n_{\text{init}}}\big) \geq -\varepsilon\big) \to 1$$

as $n_{\text{init}} \to \infty$. Then we have proved (A.19) holds which implies (E2).

Acknowledgements. We are grateful for the insightful comments of two anonymous referees.

SUPPLEMENTARY MATERIAL

Supplement to "Robust discrimination designs over Hellinger neighbourhoods" (DOI: 10.1214/16-AOS1503SUPP; .pdf). There we give the rather lengthy proof of Theorem 2.1, which depends on a number of preliminary lemmas. We also show that the conditions of this theorem apply to normal and log-normal densities.

REFERENCES

ATKINSON, A. C. and FEDOROV, V. V. (1975a). The design of experiments for discriminating between two rival models. *Biometrika* **62** 57–70. MR0370955

ATKINSON, A. C. and FEDOROV, V. V. (1975b). Optimal design: Experiments for discriminating between several models. *Biometrika* **62** 289–303. MR0381163

BOX, G. E. P. and DRAPER, N. R. (1959). A basis for the selection of a response surface design. *J. Amer. Statist. Assoc.* **54** 622–654. MR0108872

CHAUDHURI, P. and MYKLAND, P. A. (1993). Nonlinear experiments: Optimal design and inference based on likelihood. *J. Amer. Statist. Assoc.* **88** 538–546. MR1224379

DETTE, H. (1994). Discrimination designs for polynomial regression on compact intervals. *Ann. Statist.* **22** 890–903. MR1292546

DETTE, H. and HALLER, G. (1998). Optimal designs for the identification of the order of a Fourier regression. *Ann. Statist.* **26** 1496–1521. MR1647689

DETTE, H. and TITOFF, S. (2009). Optimal discrimination designs. *Ann. Statist.* **37** 2056–2082. MR2533479

FEDOROV, V. V. (1975). Optimal experimental designs for discriminating two rival regression models. In *A Survey of Statistical Design and Linear Models* (*Proc. Internat. Sympos., Colorado State Univ., Ft. Collins, Colo.*, 1973) 155–164. North-Holland, Amsterdam. MR0375666

FEDOROV, V. V. and PAZMAN, A. (1968). Design of physical experiments. *Fortschr. Phys.* **16** 325–355.

FORD, I., TITTERINGTON, D. M. and KITSOS, C. P. (1989). Recent advances in nonlinear experimental design. *Technometrics* **31** 49–60. MR0997670

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

HU, R. (2016). Robust designs for model discrimination and prediction of a threshold probability. Ph.D. dissertation, Univ. Alberta.

HU, R. and WIENS, D. P. (2017). Supplement to "Robust discrimination designs over Hellinger neighbourhoods." DOI:10.1214/16-AOS1503SUPP.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York. MR0606374

HUNTER, W. G. and REINER, A. M. (1965). Designs for discriminating between two rival models. *Technometrics* **7** 307–323. MR0192615

LÓPEZ-FIDALGO, J., TOMMASI, C. and TRANDAFIR, P. C. (2007). An optimal experimental design criterion for discriminating between non-normal models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 231–242. MR2325274

OOSTERHOFF, J. and VAN ZWET, W. R. (2012). *A Note on Contiguity and Hellinger Distance*. Springer, New York.

SINHA, S. and WIENS, D. P. (2003). Asymptotics for robust sequential designs in misspecified regression models. In *Mathematical Statistics and Applications*: *Festschrift for Constance van Eeden*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **42** 233–247. IMS, Beachwood, OH. MR2138295

WIENS, D. P. (2009a). Robust discrimination designs. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 805–829. MR2750096

WIENS, D. P. (2009b). Asymptotic properties of an Neyman–Pearson test for model discrimination, with an application to experimental design. *J. Stat. Theory Pract.* **3** 419–427. MR2751609

WIENS, D. P. and LI, P. (2014). V-optimal designs for heteroscedastic regression. *J. Statist. Plann. Inference* **145** 125–138. MR3125354

WYNN, H. P. (1970). The sequential generation of D-optimum experimental designs. *Ann. Math. Statist.* **41** 1655–1664. MR0267704

DEPARTMENT OF MATHEMATICS
AND STATISTICS
MACEWAN UNIVERSITY
EDMONTON, AB T5P 2P7
CANADA
E-MAIL: rhu@ualberta.ca

DEPARTMENT OF MATHEMATICAL
AND STATISTICAL SCIENCES
UNIVERSITY OF ALBERTA
EDMONTON, AB T6G 2G1
CANADA
E-MAIL: doug.wiens@ualberta.ca