Journal of Statistical Planning and Inference 145 (2014) 125–138

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

V-optimal designs for heteroscedastic regression

Douglas P. Wiens [a,*], Pengfei Li [b]

[a] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1
[b] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

ARTICLE INFO

Article history:
Received 12 February 2013
Received in revised form
14 August 2013
Accepted 13 September 2013
Available online 23 September 2013

Keywords:
Adaptive
Forcing
Robustness
Sequential
Weighted least squares

ABSTRACT

We obtain V-optimal designs, which minimize the average variance of predicted regression responses, over a finite set of possible regressors. We assume a general and possibly heterogeneous variance structure depending on the design points. The variances are either known (or at least reliably estimated) or unknown. For the former case we exhibit optimal static designs; our methods are then modified to handle the latter case, for which we give a sequential estimation method which is fully adaptive, yielding both consistent variance estimates and an asymptotically V-optimal design.

1. Introduction and summary

Suppose that one can observe random variables Y_i $(i=1,\ldots,I)$, with means $\mathbf{x}_i^T\boldsymbol{\beta}$ and possibly heterogeneous variances $\sigma_i^2>0$. Here $\mathbf{x}_i$ is a d-dimensional vector of regressors, and $\boldsymbol{\beta}$ is a d-dimensional vector of unknown parameters. Quite typically the ultimate goal of the experiment will be prediction – at pre-specified levels of the regressors – in which case the designer will aim for minimization of the prediction variance at these levels. Such 'V-optimal' designs are defined as minimizers of the average variance of the predictions $\hat{Y}_i$.

To set up our approach, and the manner in which we intend to handle the heteroscedasticity, suppose that one makes $n_i = n\lambda_i$ $(n=\sum_{i=1}^I n_i)$ observations $\{Y_{ij}\}_{j=1}^{n_i}$, with the intention of estimating $\boldsymbol{\beta}$ and then estimating $E[Y_i]$ by $\hat{Y}_i = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$. We assume that all n observations are independent, so that for the weighted least squares (wls) estimate with weights $\{w_i\}$, viz.

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin} \sum_{i,j} w_i (Y_{ij} - \mathbf{x}_i^T\boldsymbol{\beta})^2,$$

the covariance matrix of $\hat{\mathbf{Y}} = (\hat{Y}_1,\ldots,\hat{Y}_I)^T$ is $n^{-1}\mathbf{C}$, where

(i) $\mathbf{C}=\mathbf{BW\Sigma W\Lambda B}$ for $\mathbf{B}=\mathbf{X}(\mathbf{X}^T\mathbf{W\Lambda X})^{-1}\mathbf{X}^T$;
(ii) $\mathbf{W}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Sigma}$ are $I\times I$ diagonal matrices with diagonal elements $\{w_i\}$, $\{\lambda_i\}$ and $\{\sigma_i^2\}$ $(\sigma_i^2 = \mathrm{VAR}[Y_{ij}])$ respectively; and
(iii) $\mathbf{X}$ is the $I\times d$ model matrix with rows $\{\mathbf{x}_i^T\}$, assumed to have full rank $d\le I$.

A more transparent representation of $\mathbf{C}$ is as $\mathbf{C}=\mathbf{B}[diag(\ldots,\lambda_i w_i^2\sigma_i^2,\ldots)]\mathbf{B}$. If the error variances are known then the optimal weights are $w_i=\sigma_i^{-2}$, leading to

$$\mathbf{C}=\mathbf{B}=\mathbf{X}(\mathbf{V}^T\boldsymbol{\Lambda}\mathbf{V})^{-1}\mathbf{X}^T, \tag{1}$$

[*] Corresponding author.
E-mail addresses: doug.wiens@ualberta.ca (D.P. Wiens), pengfei.li@uwaterloo.ca (P. Li).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.09.007

where $\mathbf{V}=\mathbf{\Sigma}^{-1/2}\mathbf{X}$. In the more common case that the variances are unknown they are replaced by estimates, the accuracy of which partially determines the efficiency of resulting procedures.

There is a variety of optimality criteria that one might adopt, each of which has been modified to accommodate heteroscedasticity – but typically only of a *known* form. Wong (1995) established the equivalence of the D- and G-optimality criteria for heteroscedastic linear models. Heiligers (1996) studied E-optimal designs – both design points and design weights – for heteroscedastic polynomial models with known variances. Li and Chan (2001) studied L_p - optimal designs for more general models, still assuming a known variance structure. Dette et al. (2005) obtained Bayesian and minimax D-optimal designs, assuming certain one-parameter variance functions.

Wiens (2011) obtained D- and I-optimal regression designs, assuming wls estimates, which explicitly account for the error in estimating the variances.

Here we consider the computation of V-optimal design weights for fixed model matrices $\mathbf{X}$. These are minimizers of the average, or total, variance of $\sqrt{n}\hat{\mathbf{Y}}$ over the I locations $\{\mathbf{x}_i\}$ forming the columns of $\mathbf{X}^T$. We initially assume that the variances are known, so that optimal weighting can be done. Then, using (1), $\text{VAR}[\sqrt{n}\hat{Y}_i]=\mathbf{x}_i^T(\mathbf{V}^T\mathbf{\Lambda}\mathbf{V})^{-1}\mathbf{x}_i$, and the total variance is, with $\boldsymbol{\lambda}\overset{def}{=}(\lambda_1,\ldots,\lambda_I)^T$, given by

$$L(\lambda)=\sum_{i=1}^{I}\mathbf{x}_i^T(\mathbf{V}^T\mathbf{\Lambda}\mathbf{V})^{-1}\mathbf{x}_i=tr\,\mathbf{C}. \tag{2}$$

The restriction that $n\lambda_i$ be an integer will be dropped – see Pukelsheim and Rieder (1992) for methods of approximating the resulting designs with implementable ones. We seek to minimize (2) subject to $\lambda_i\ge 0, \sum_{i=1}^{I}\lambda_i=1$. We note that λ can be identified with a probability distribution on $\{1,\ldots,I\}$, and that the set $\mathcal{P}$ of all such distributions for which $\mathbf{V}^T\mathbf{\Lambda}\mathbf{V}$ is non-singular is convex. Note as well that we allow the possibility that $\lambda_i=0$ – there is no requirement that observations be made at all of the $\mathbf{x}_i$, although we will impose such a requirement in Section 4, in order to be able to consistently estimate all of the σ_i.

The following two simple yet motivating examples, to which we will return, illustrate some of the limitations of the current literature on this problem.

Example 1. Suppose that the regressors are merely the group indicators $\mathbf{e}_i$ – the ith column of $\mathbf{X}=\mathbf{I}_I$. More generally, suppose that $I=d$, so that $\mathbf{X}$ is square and non-singular. Then the predictions $\hat{Y}_i=\overline{y}_i$ do not depend on the regressors and $\mathbf{C}=\mathbf{\Sigma}\mathbf{\Lambda}^{-1}$, with $L(\lambda)=\sum_{i=1}^{I}(\sigma_i^2/\lambda_i)$. A simple argument gives the optimal allocations

$$\lambda_{0i}=\frac{\sigma_i}{\sigma_1+\cdots+\sigma_I},\quad i=1,\ldots,I, \tag{3}$$

with $\min_\lambda L(\lambda)=L(\lambda_0)=(\sum_{i=1}^{I}\sigma_i)^2$. This is also a special case of the results of Pukelsheim and Torsney (1991), who studied the construction of optimal design weights for a variety of criteria, with $I=d$ and homoscedastic errors. The present work extends Pukelsheim and Torsney by relaxing both of these restrictions.

Example 2. Suppose that $d=1$. Then $\mathbf{X}\overset{def}{=}\mathbf{x}=(x_1,\ldots,x_I)^T$ and $L(\lambda)=\|\mathbf{x}\|^2/\sum_{i=1}^{I}(\lambda_i x_i^2/\sigma_i^2)$. This is clearly minimized by the degenerate choice

$$\lambda_{0i}=\begin{cases}1 & \text{if } \dfrac{x_i^2}{\sigma_i^2}=\max\limits_k\dfrac{x_k^2}{\sigma_k^2},\\ 0 & \text{otherwise.}\end{cases} \tag{4}$$

In both of these examples there are only d points at which $\lambda_i>0$ – a common situation in optimal design theory but one which will not necessarily continue to hold in our framework.

In Section 2 we treat static, i.e. non-sequential, V-optimal designs. We give, in Theorem 1, a necessary and sufficient condition for the V-optimality of an allocation vector λ. We give as well an easily checked necessary condition; this affords a considerable reduction in the numerical complexity of the ensuing computational problem. In Theorem 2 we specialize to the case of d-point designs, which, when they exist, are numerically simplest to construct. We give a computing algorithm to carry out this case, and another which gives the V-optimal designs on more than d points. Several examples illustrate the designs constructed using the theory of this section. In Section 2.1 we give an explicit expression for the change in the loss if the variances are misspecified; a robust strategy for handling this is suggested. Then in Section 3 we present (Theorem 3) a method of sequentially allocating design points. The method is very quick, and is in particular useful if the error variances are to be estimated, and the estimates updated as the data accrue. As well, the designs obtained sequentially with known variances suggest which rows of $\mathbf{X}$ will form the support of a static V-optimal design. One can then use the algorithm developed in Section 2 to compute the optimal allocations to these rows, and check the conjecture of V-optimality. We illustrate this method (Example 7 of Section 3) by using it to construct a V-optimal design for cubic regression on $I=41$ points.

In Section 4 we drop the assumption of known variances. Our sequential techniques are modified so as to force a certain minimum proportion of the observations to be allocated to each design point; this yields (Theorem 4) both consistent estimates of the variances and an asymptotically V-optimal design. The – very efficient – computing algorithms and large sample results are illustrated in these sections, and as well in a case study in Section 5. All derivations are in the Appendix.

2. Construction of static V-optimal designs

Note that $L(\lambda)$ is invariant under transformations $\mathbf{X}\to\mathbf{XT}$ with $\mathbf{T}$ nonsingular. We assume that this has already been done, in such a way that

$$\mathbf{X}^T\mathbf{X}=\mathbf{I}_d, \tag{5}$$

so that (2) becomes

$$L(\lambda)=\sum_{i=1}^{I}\mathbf{e}_i^T(\mathbf{V}^T\mathbf{\Lambda}\mathbf{V})^{-1}\mathbf{e}_i=tr[(\mathbf{V}^T\mathbf{\Lambda}\mathbf{V})^{-1}]. \tag{6}$$

A design formed from the rows of $\mathbf{X}$ can be characterized by the 'allocations vector' λ, viewed as a member of $\mathcal{P}$. We denote by $\mathcal{I}=\mathcal{I}_\lambda$ the support of λ, i.e. the set of indices i for which $\lambda_i>0$, and by $\mathbf{C}_\lambda$ the covariance matrix (1). Define

$$\alpha_i(\lambda)=[\mathbf{V}(\mathbf{V}^T\mathbf{\Lambda}\mathbf{V})^{-2}\mathbf{V}^T]_{ii},\quad i=1,\dots,I; \tag{7a}$$

$$\alpha_{\max}(\lambda)=\max_{1\le i\le I}\alpha_i(\lambda). \tag{7b}$$

We sometimes prefer the simpler expression $\alpha_i(\lambda)=[\mathbf{C}_\lambda^2]_{ii}/\sigma_i^2$. Note also that

$$E_\lambda[\alpha(\lambda)]\overset{def}{=}\sum_{i=1}^{I}\lambda_i\alpha_i(\lambda)=tr[(\mathbf{V}^T\mathbf{\Lambda}\mathbf{V})^{-1}]=L(\lambda). \tag{8}$$

Theorem 1. (*a*) *There is a V-optimal design whose allocations vector we denote by* λ_0. *In order that a vector* λ_0 *define a V-optimal design, it is* necessary and sufficient *that*

$$\lambda_{0i}=0\text{ if }\alpha_i(\lambda_0)<\alpha_{\max}(\lambda_0), \tag{9}$$

i.e. the support $\mathcal{I}_{\lambda_0}$ *must be such that* $\alpha_i=\alpha_{\max}$ *for* $i\in\mathcal{I}_{\lambda_0}$. *Equivalently,*

$$E_{\lambda_0}[\alpha(\lambda_0)]=\alpha_{\max}(\lambda_0).$$

(*b*) *If* (9) *holds then the minimum loss is*

$$\min_\lambda L(\lambda)=L(\lambda_0)=\alpha_{\max}(\lambda_0), \tag{10}$$

and so a necessary *condition for the optimality of* λ_0 *is that*

$$\alpha_{\max}(\lambda_0)=\min_{\lambda\in\mathcal{P}}\alpha_{\max}(\lambda). \tag{11}$$

Example 1.1 (*Example 1 continued*). With λ_0 given by (3) we have $\mathbf{C}_{\lambda_0}=\mathbf{\Sigma}^{1/2}tr(\mathbf{\Sigma}^{1/2})$, with $\mathcal{I}_{\lambda_0}=\{1,\dots,I\}$ and $\alpha_i=(\sigma_1+\cdots+\sigma_I)^2\equiv\alpha_{\max}=L(\lambda_0)$.

Example 2.1 (*Example 2 continued*). If λ_0 is given by (4) we have $\mathbf{C}_{\lambda_0}=\mathbf{xx}^T/\max_k(x_k^2/\sigma_k^2)$, with

$$\mathcal{I}_{\lambda_0}=\left\{i\left|\frac{x_i^2}{\sigma_i^2}=\max_k\frac{x_k^2}{\sigma_k^2}\right.\right\},\alpha_i(\lambda_0)=\|\mathbf{x}\|^2\frac{x_i^2}{\sigma_i^2}\Big/\left(\max_k\frac{x_k^2}{\sigma_k^2}\right)^2.,\alpha_{\max}(\lambda_0)=\|\mathbf{x}\|^2\Big/\left(\max_k\frac{x_k^2}{\sigma_k^2}\right).=L(\lambda_0).$$

Denote by $|\mathcal{I}_\lambda|$ the number of elements in the set $\mathcal{I}_\lambda$. Assume that $I>d$ – we have seen (Example 1) that the solution is straightforward if $I=d$. We first specialize Theorem 1 to the case $|\mathcal{I}_\lambda|=d$, in which case the conditions of the theorem translate quite simply. Then the, numerically more involved, case $|\mathcal{I}_\lambda|=d'>d$ is dealt with.

Let $\mathbf{X}_0=\mathbf{X}_0(\mathcal{I})$ be a submatrix formed from the d rows of $\mathbf{X}$, indexed by $\mathcal{I}$ and assumed to be linearly independent. Denote by $(\mathbf{X}_0\mathbf{X}_0^T)^{ii}$ the ith diagonal element of $(\mathbf{X}_0\mathbf{X}_0^T)^{-1}$, and by $\mathbf{x}_0^{(i)}$ the ith column of $\mathbf{X}_0^{-1}$ (so that $(\mathbf{X}_0\mathbf{X}_0^T)^{ii}=\|\mathbf{x}_0^{(i)}\|^2$).

Theorem 2. (*a*) *In order that* λ_0 *define a V-optimal design on the d rows of* $\mathbf{X}_0(\mathcal{I}_{\lambda_0})$, *it is* necessary and sufficient *that* (*i*)

$$\lambda_{0i}=\left\{\sigma_i\|\mathbf{x}_0^{(i)}\|\Big/\sqrt{\alpha_{\max}lpar\lambda_0)}.,i\in\mathcal{I}_{\lambda_0},0,i\notin\mathcal{I}_{\lambda_0},\right. \tag{12}$$

for

$$\alpha_{\max}(\lambda_0)=\sum_{i\in\mathcal{I}_{\lambda_0}}\sigma_i\|\mathbf{x}_0^{(i)}\|, \tag{13}$$

and that (*ii*) *all columns of* $\mathbf{H}_0\mathbf{X}^T\mathbf{\Sigma}^{-1/2}$, *where* $\mathbf{H}_0=\mathbf{X}_0^{-1}\mathbf{\Sigma}_0^{1/2}\mathbf{D}_0^{-1}\mathbf{X}_0^{-T}$ *for* $\mathbf{D}_0=diag(\|\mathbf{x}_0^{(i)}\|,i\in\mathcal{I}_{\lambda_0})$ *and* $\mathbf{\Sigma}_0^{1/2}=diag(\sigma_i,i\in\mathcal{I}_{\lambda_0})$, *have norms* ≤ 1. (*b*) *If the conditions of* (*a*) *hold then the minimum loss is* $L(\lambda_0)=\alpha_{\max}(\lambda_0)$. *Thus a* necessary condition, *if* (*a*) *is to* *hold, is that*

$$\sum_{i \in \mathcal{I}_{\lambda_0}} \sigma_i \| \mathbf{x}_0^{(i)} \| = \min, \tag{14}$$

among all $d \times d$ *submatrices of* $\mathbf{X}$.

The computations necessary to implement Theorem 2, and all others in this article, have been programmed in MATLAB; the code is available from the authors. The algorithm is as follows:

Step A1 Cycle through the $\binom{I}{d}$ choices of $\mathbf{X}_0(\mathcal{I})$ with $|\mathcal{I}| = d$. Arrange these by decreasing values of the determinants $|\mathbf{X}_0^T\mathbf{X}_0|$ – this greatly increases the speed of the algorithm, when there is in fact a d-point optimal design, since the optimal choice typically also results in a large value of this determinant. Define $\alpha_{\max} = \infty$ if $\mathbf{X}_0$ is singular.

Step A2 If $\alpha_{\max} = \sum_{i \in \mathcal{I}} \sigma_i \| \mathbf{x}_0^{(i)} \|$ corresponding to $\mathbf{X}_0(\mathcal{I})$ is not a minimum, among those choices of $\mathcal{I}$ considered to this point, then it is rejected. This is because a *necessary* condition for an optimum is that $\mathbf{X}_0$ minimize $\alpha_{\max}$. If a local minimum *is* found, then (12) is computed and the sufficient condition in Theorem 2 (a) (ii) is checked.

Step A3 If the sufficient condition is satisfied then an optimal d-point design has been found and no more choices are checked. If not, test another choice $\mathbf{X}_0(\mathcal{I})$. If no optimal d-point design exists, go to Step B1 below.

If $d=1$ then (Example 2) there is always a d-point V-optimal design – several, if $\max(x_k^2/\sigma_k^2)$ is not attained uniquely. In fact the necessary condition (14) with $d=1$ is that x_i^2/σ_i^2 be maximal for $i = \mathcal{I}_{\lambda_0}$ and this condition is also sufficient. For $d > 1$ the necessary condition is not always sufficient. If it is not, then there is no d-point optimal design and one must search for an optimal design on $d' > d$ points of support. For $|\mathcal{I}| = d' = d+1, d+2, \ldots$ we then carry out the following:

Step B1 Let $\mathbf{X}_0 = \mathbf{X}_0(\mathcal{I})$ be a submatrix formed from the d' rows of $\mathbf{X}$ indexed by $\mathcal{I}$, and relabel the rows of $\mathbf{X}$ so that those of $\mathbf{X}_0$ come first:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_0 \\ \mathbf{X}_1 \end{pmatrix} \begin{array}{l} \leftarrow d' \\ \leftarrow I - d' \end{array}$$

Arrange these by decreasing values of the determinants $|\mathbf{X}_0^T\mathbf{X}_0|$. Decompose $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$ compatibly, viz., $\mathbf{\Sigma} = \mathbf{\Sigma}_0 \oplus \mathbf{\Sigma}_1$ and $\mathbf{\Lambda} = \mathbf{\Lambda}_0 \oplus \mathbf{\Lambda}_1$. Then with

$$\mathbf{A}_\lambda \overset{def}{=} \mathbf{X}_0^T \mathbf{\Sigma}_0^{-1/2} \mathbf{\Lambda}_0 \mathbf{\Sigma}_0^{-1/2} \mathbf{X}_0,$$

we have $\mathbf{C} = \mathbf{X}\mathbf{A}_\lambda^{-1}\mathbf{X}^T$, $\mathbf{C}^2 = \mathbf{X}\mathbf{A}_\lambda^{-2}\mathbf{X}^T$, and (with × denoting submatrices of no interest to us)

$$\mathbf{\Sigma}^{-1/2}\mathbf{C}^2\mathbf{\Sigma}^{-1/2} = \begin{pmatrix} \mathbf{\Sigma}_0^{-1/2}\mathbf{X}_0\mathbf{A}_\lambda^{-2}\mathbf{X}_0^T\mathbf{\Sigma}_0^{-1/2} & \times \\ \times & \mathbf{\Sigma}_1^{-1/2}\mathbf{X}_1\mathbf{A}_\lambda^{-2}\mathbf{X}_1^T\mathbf{\Sigma}_1^{-1/2} \end{pmatrix}.$$

Cycle through the $\binom{I}{d'}$ choices of $\mathbf{X}_0(\mathcal{I})$. Solve the equations '$\alpha_i(\lambda) = \alpha_{\max}$ $(i = 1, \ldots, d')$; $\sum \lambda_{0i} = 1$' for the $d'+1$ unknowns $\lambda_{01}, \ldots, \lambda_{0d'}, \alpha_{\max}$. The constraint that the λ_{0i} be non-negative is addressed by writing $\lambda_i = \mu_i^2$ and requiring that $\sum \mu_i^2 = 1$. These equations are then $\mathbf{F}(\boldsymbol{\theta}_0) = \mathbf{0}$, where $\boldsymbol{\theta} = (\boldsymbol{\mu}^T, \alpha)^T$ and

$$\mathbf{F}(\boldsymbol{\theta}) = \begin{pmatrix} diag(\mathbf{\Sigma}_0^{-1/2}\mathbf{X}_0\mathbf{A}_\lambda^{-2}\mathbf{X}_0^T\mathbf{\Sigma}_0^{-1/2}) - \alpha_{\max}\mathbf{1}_{d'} \\ \sum_{i=1}^{d'} \mu_i^2 - 1. \end{pmatrix}.$$

Step B2 If $\alpha_{\max}$ corresponding to $\mathbf{X}_0(\mathcal{I}_\lambda)$ is not a minimum, among those choices of $\mathcal{I}_\lambda$ considered to this point, then it is rejected. If a local minimum *is* found, then we check the sufficient condition:

$$\frac{1}{\sigma_i^2}\mathbf{x}_i^T\mathbf{A}_{\lambda_0}^{-2}\mathbf{x}_i \le \alpha_{\max} \quad \text{for } i = d'+1, \ldots, I.$$

Equivalently, $[\mathbf{\Sigma}^{-1/2}\mathbf{C}_{\lambda_0}^2\mathbf{\Sigma}^{-1/2}]_{ii}/\alpha_{\max} = [\mathbf{\Sigma}^{-1/2}\mathbf{X}\mathbf{A}_{\lambda_0}^{-2}\mathbf{X}^T\mathbf{\Sigma}^{-1/2}]_{ii}/\alpha_{\max} \le 1$, for $i = 1, \ldots, I$.

Step B3 If the sufficient condition is satisfied then the optimal $d'-$point design has been found and no more choices are checked. If not, test another choice $\mathbf{X}_0(\mathcal{I})$. If no optimal $d'-$point design exists, increment d' and repeat.

Example 3. That steps B1–B3 above are typically necessary is brought out forcefully by this example. We set I=8, d=4 and generated an $I \times d$ matrix $\mathbf{X}$, along with standard deviations $\{\sigma_i\}_{i=1}^I$. These standard deviations, together with the first three columns of $\mathbf{X}$, were

$\{\sigma_i\}_{i=1}^8$	Columns 1–3 of **X**		
1.0	1.0	−.2	−.9
.7	−1.4	.1	−.7

.3	−.1	−.5	−.5
1.1	1.3	.7	−.3
.4	−.7	−.1	0
.6	.3	.3	−3.0
.2	.2	.0	−.5
1.8	−.1	−1.3	1.2

We then chose a fourth column of $\mathbf{X}$, transformed $\mathbf{X}$ to satisfy (5), and carried out the algorithm described in the preceding section. Depending on this final column we obtained V-optimal designs on $d=4$ and $d'=5,6,7,8$ points. These five columns, followed by the five designs, were

Column 4 of $\mathbf{X}$					V-optimal design λ_0				
−1.1	−.6	−.2	−.8	−.6	0	0	.2012	.0779	.1126
.9	.1	.9	.4	1.2	.2565	.1647	.1673	.1729	.1441
.4	−.9	−.8	−.2	−.7	.1980	.1674	.0956	.1057	.0953
0	−.2	−1.4	1.1	.2	.3286	.3036	.1581	.1901	.1004
.2	−1.7	−1.4	−1.1	−.4	0	.1808	.0971	.1086	.0751
−1.6	.6	.5	0	−1.5	.2169	.1836	0	0	.0858
−.1	−.1	−.2	.6	.5	0	0	0	.0639	.0683
1.6	.7	−.2	1.1	−1.2	0	0	.2807	.2810	.3184
Minimum loss tr $(\mathbf{C}_{\lambda_0})$:					12.30	14.42	18.79	17.16	18.81

Example 4. Here we chose $\mathbf{X}$ to be the model matrix for quadratic regression, with rows $(1,x,x^2)$ for $x=-1(.2)1$, so that $I=11, d=3$. We first took constant variances. The algorithm returned an optimal 3-point design with weights $\{.2715,.4569,.2715\}$ on $x=\{-1,0,1\}$ respectively. Increasing I resulted in 3-point designs more and more closely approximating the A-optimal design for this situation, which has weights $\{.25,.50,.25\}$ on $x=\{-1,0,1\}$ respectively. With standard deviations $\boldsymbol{\sigma}_0=(0.7,1.3,.1,.4,.4,.3,.3,.4,.2,1.5,1.2)^T$ we instead obtained a 4-point design with weights $\{.1612,.1260,.4068,.3060\}$ on $x=\{-1,-.6,0,.6\}$ respectively.

Example 5. We chose $\mathbf{X}$ to be the model matrix for cubic regression, with x as in Example 4. With constant variances we obtained the symmetric 6-point design with weights $\{.1886,.0107,.3007\}$ on $x=\{\pm 1,\pm .6,\pm .4\}$ respectively. This design can be viewed as a discretization, to these chosen x-values, of the arcsine design for cubic regression and A-loss presented by Pukelsheim and Torsney (1991). The arcsine design places weights $\{.158,.342\}$ on $x=\{\pm 1,\pm .5\}$ respectively.

Example 6. With $\mathbf{X}$ as in Example 5 but with standard deviations $\boldsymbol{\sigma}_0$ as in Example 4 we instead obtained a 6-point design with weights $\{.2682,.0672,.0890,.0740,.1226,.3790\}$ on $x=\{-1,-.6,0,.2,.6,1\}$ respectively.

We revisit these examples in Section 3.

2.1. Effect of misspecified variances

Suppose that a design λ_0, optimal for variances $\{\sigma_{0,i}^2\}$ and employing weights $w_i=\sigma_{0,i}^{-2}$, is employed when the variances are in fact $\{\sigma_i^2\}$. A calculation, employing (5), (7a) and (7b), yields that the total prediction variance is then

$$L(\lambda_0;\boldsymbol{\sigma})=\sum_{i=1}^{I}\left(\frac{\sigma_i^2}{\sigma_{0,i}^2}\right)\lambda_{0,i}\alpha_i(\lambda_0)=\alpha_{\max}(\lambda_0)\sum_{\lambda_{0,i}>0}\left(\frac{\sigma_i^2}{\sigma_{0,i}^2}\right)\lambda_{0,i}.$$

This is to be compared with the total prediction variance $L(\lambda_0;\boldsymbol{\sigma}_0)$ for correctly specified variances, given by (10), from which we obtain the ratio

$$\frac{L(\lambda_0;\boldsymbol{\sigma})}{L(\lambda_0;\boldsymbol{\sigma}_0)}=\sum_{\lambda_{0,i}>0}\left(\frac{\sigma_i^2}{\sigma_{0,i}^2}\right)\lambda_{0,i}, \tag{15}$$

representing the decrease in efficiency. This ratio is the average, with respect to λ_0, ratio of the true to assumed variances. It can of course be arbitrarily large – or small – if the variances are arbitrarily badly specified, but under realistic scenarios the change is expected to be only slight. In fact one could achieve a measure of robustness by intentionally overspecifying the $\sigma_{0,i}^2$. Note also that (15) is unaffected by variances which are misspecified at points not in the support of λ_0.

3. A sequential algorithm

A common alternative to a static design is a sequential allocation of resources, allowing the experimenter to terminate sampling once his goals have been realized. In such a scheme, if n observations have currently been made then one might estimate the σ_i^2 and determine the $(n+1)$th location in order to maximize the decrease in the resulting estimated loss. This can be based on the following preliminary result, in which we assume that the variances are *known*.

Theorem 3. *Start with a design $\lambda^{(n_0)} \in \mathcal{P}$ allocating n_0 observations. Given $\{\sigma_i\}_{i=1}^{I}$ and a current allocation of n observations, construct a design in the following manner. Define*

$$\Lambda^{(n)} = diag\left(\frac{n_1}{n}, \ldots, \frac{n_I}{n}\right),$$
$$\mathbf{C}_n = \mathbf{X}(\mathbf{V}^T\mathbf{\Lambda}^{(n)}\mathbf{V})^{-1}\mathbf{X}^T,$$
$$\alpha_{i,n} = \alpha_i(\lambda^{(n)}) = [\mathbf{C}_n^2]_{ii}/\sigma_i^2,$$

and allocate the $(n+1)$th *observation to the location* $i^{(n)} = \arg\max_{1 \le i \le I} \alpha_{i,n}$. *Then with*

$$\Delta_n \overset{def}{=} \frac{[\mathbf{C}_n^2]_{i^{(n)}i^{(n)}}}{n\sigma_{i^{(n)}}^2 + \mathbf{C}_{n,i^{(n)}i^{(n)}}} = \frac{\alpha_{\max}(\lambda^{(n)})}{n + (\mathbf{C}_{n,i^{(n)}i^{(n)}}/\sigma_{i^{(n)}}^2)},$$

and with $\mathbf{c}_{n,i}$ denoting the ith column of C_n, we have

$$\mathbf{C}_{n+1} = \frac{n+1}{n}\left\{\mathbf{C}_n - \frac{\mathbf{c}_{n,i^{(n)}}\mathbf{c}_{n,i^{(n)}}^T}{n\sigma_{i^{(n)}}^2 + \mathbf{C}_{n,i^{(n)}i^{(n)}}}\right\}, \tag{16}$$

$$L\left(\lambda^{(n+1)}\right) = \frac{n+1}{n}\left(L\left(\lambda^{(n)}\right) - \Delta_n\right). \tag{17}$$

The sequence constructed in this manner is asymptotically optimal:

$$L(\lambda^{(n)}) \to \min_{\lambda \in \mathcal{P}} L(\lambda) \quad \text{as} \quad n \to \infty. \tag{18}$$

The proof of (18) closely parallels that of Wynn (1970) for D-optimal designs; see also Section 4.2 of Fedorov (1972). For the rest of this section we denote by N the final size of a sequentially obtained design.

Example 1.2 (*Example 1 continued*). If $\mathbf{X}$ is square and non-singular, Theorem 3 leads to placing the next observation at location $i^{(n)} = \arg\max_{1 \le i \le I} \frac{\sigma_i}{n_i}$. The sequence of choices $\{i^{(n)}\}$ clearly yields designs converging to (3).

Example 2.2 (*Example 2 continued*). If $d=1$, Theorem 3 leads to $i^{(n)} = \arg\max(x_k^2/\sigma_k^2)$, thus attaining (4) exactly, regardless of the current $\lambda^{(n)}$.

Example 5.1 (*Example 5 continued*). See Fig. 1, which is typical. The initial design had one point at each of $I=11$ locations. (These have been removed from the final design presented here.) The final ($N=100$) design λ_0 has all mass where $\alpha_i(\lambda_0) = \alpha_{\max}(\lambda_0) = E_{\lambda_0}[\alpha(\lambda_0)] = 37.0551$. It places weights $\{.2022, .0112, .2921, .2921, .0112, .1910\}$ on the design points $x = \{-1, -.6, -.4, .4, .6, 1\}$ respectively – very close to the weights $\{.1886, .0107, .3007\}$ placed at $x = \{\pm 1, \pm .6, \pm .4\}$ by the static V-optimal design, with loss 37.0039, found in Section 2. This bodes well for our approach in Section 4, where the current approach is applied to unknown, but consistently estimated, variances.

In practice the variances will typically be estimated, and re-estimated prior to assigning the next location in the sequential construction of the design. The convergence of this method requires the variance estimates to be consistent; this in turn requires a certain minimum proportion of the allocations to be made at each location. This is the subject of Section 4 below. The following example illustrates a special case, in which some of the variance estimates are based throughout only on the two observations assigned to their locations in the initial design, since no further observations are allocated to these locations.

Example 6.1 (*Example 6 continued*). Using the same inputs as in Example 6, we generated designs sequentially. We started with an initial sample of 2 observations at each of the 11 values of x, and carried out an ordinary least squares regression in order to obtain initial estimates of the σ_i. On the basis of this output a further observation was made, at a location determined as in Theorem 3. At each subsequent stage a weighted least squares regression was carried out in order to re-estimate the variances, using the inverses of the variance estimates obtained at the previous stage as weights. This process was continued until a total of $N=1000$ observations had been made. Finally the initial design points were removed, leaving the final design plotted in Fig. 2(b). We also carried out a similar but simpler process, using the true standard deviations at each stage; this resulted in Fig. 2(c). The minimum loss, using the true variances, was only very slightly larger than that in Example 6 – 7.4833 vs. 7.3685.

Example 7. The construction of the V-optimal designs, as in Section 2, is perhaps not feasible for large values of I, if a large number of values of d' must be tested. If however the correct value of d' is known in advance, then the program need be run only for this value. One might also know which rows of $\mathbf{X}$ are to be used in the design, and so need only compute one allocation λ and check the appropriate sufficient condition from Section 2. In this example we illustrate this technique, to obtain a V-optimal design for homoscedastic cubic regression on the $I=41$ points $x = -1(.05)1$. We first obtained a design sequentially, using Theorem 3 and constant variances with a final sample of size $N=500$. The result is shown in

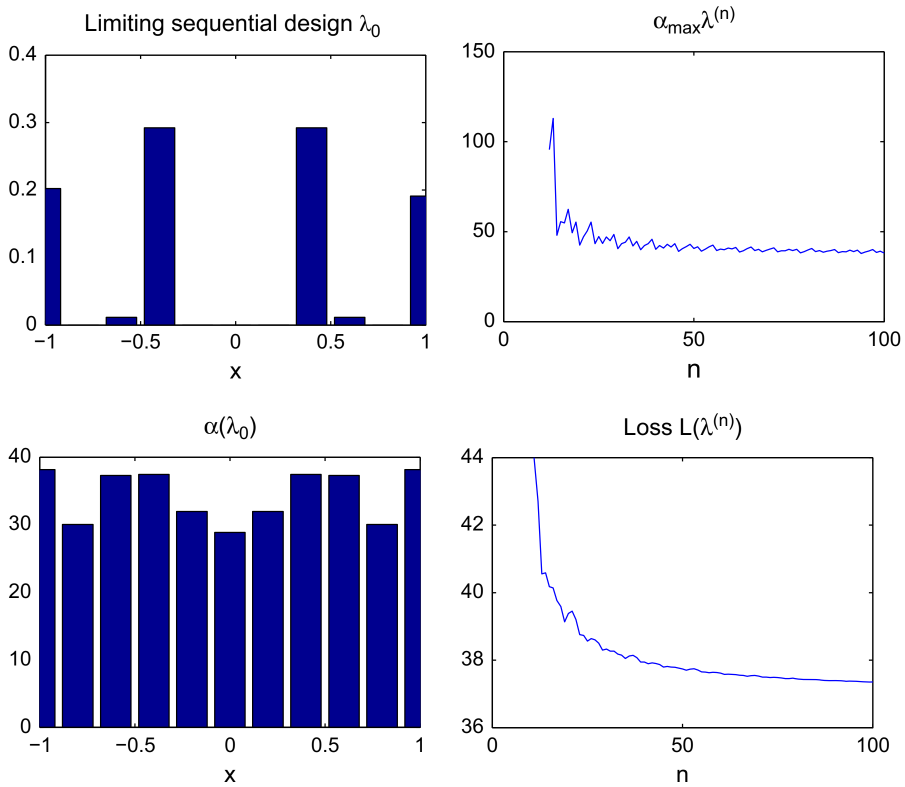

Fig. 1. Sequential design on $I = 11$ points with inputs as in Example 5. Top left is ‘limiting’ design λ_0 with final size $N = 100$; top right the sequence of maxima $\max_i \alpha_i(\lambda^{(n)})$ for $n = 11, \ldots, 100$; bottom left is $\{\alpha_i(\lambda_0)\}$ and bottom right is loss $L(\lambda^{(n)}) = E_{\lambda^{(n)}}[\alpha(\lambda^{(n)})]$.

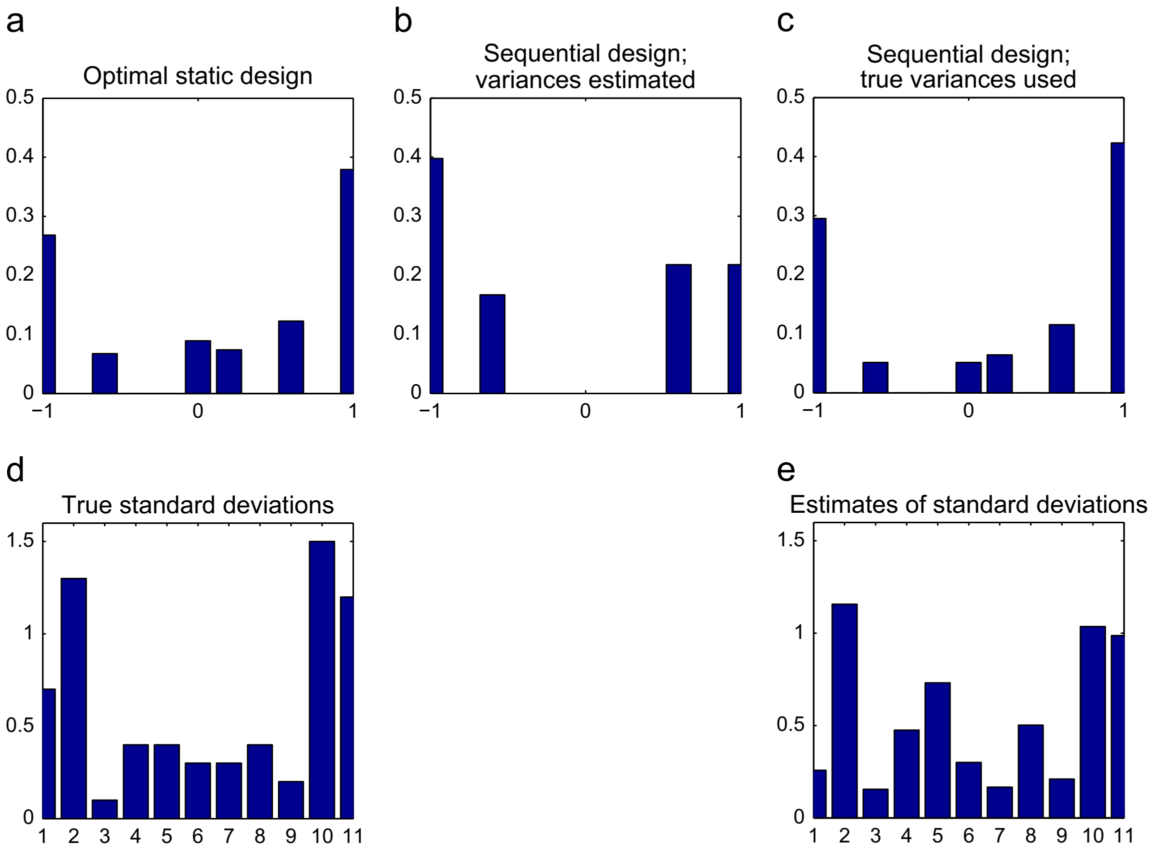

Fig. 2. Example 6.1; heteroscedastic cubic regression: (a) optimal static design; (b) sequentially obtained design using estimated variances; (c) sequentially obtained design using true variances; (d) true standard deviations versus group labels; and (e) estimates of standard deviations versus group labels.

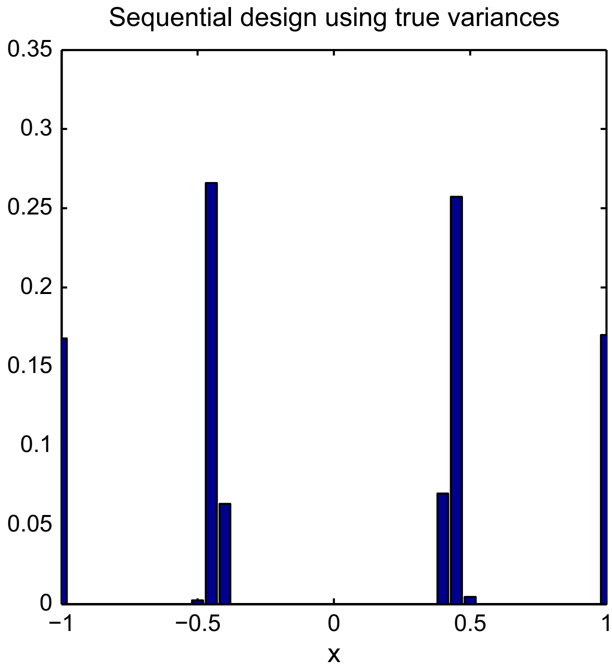

Fig. 3. Sequentially obtained design for Example 7 – homoscedastic cubic regression with $x=-1(.05)1$.

Fig. 3, and suggests that the V-optimal design will place all mass at the 6 points $\pm .40$, $\pm .45$, ± 1. We next ran the algorithm of Section 2, forcing it to start with $d'=6$ and to compute λ only for this one choice of J. The sufficient condition held, thus yielding a V-optimal design with λ assigning mass .0797 to $\pm .40$, .2566 to $\pm .45$ and .1638 to ± 1. This design is very similar to the A-optimal design for cubic regression (see Pukelsheim, 1993, p. 224) – an outcome which is somewhat expected since under homoscedasticity the efficiency of the predictions is directly related to that of the estimates of the regression coefficients.

4. Optimal adaptive allocations with forcing

In heteroscedastic regression one typically seeks consistent estimates of the variances as well as of the regression parameters. A sequential design which achieves both objectives can be obtained via the following procedure. We are partially motivated by Chaudhuri and Mykland (1993), who exhibit adaptive, sequential design procedures for nonlinear models.

A rough description of the method is as follows. The experimenter decides in advance on the total number n of observations which he will make. The first n_{init} of these are primarily used to estimate the variances. In the asymptotic theory we require n_{init} to grow with n in such a way that the resulting estimates be consistent; for this we use 'forcing', requiring always that a certain proportion of the observations be allocated to each of the I design points. This then extends a proposal made for $d=1$ (location estimation) by Antos et al. (2008). The variance estimates obtained after n_{init} observations are used to determine the remaining $n-n_{init}$ allocations, in a sequential manner which parallels that described in Section 3. Finally, a weighted least squares regression is carried out. As well, the variance estimates may be updated.

Specifically, the procedure is as follows:

Step C1 Make 2 observations at each of the I locations. Compute preliminary estimates $\{\hat{\sigma}_i\}_{i=1}^{I}$.

For $m=2I, 2I+1, \ldots, n_{init}-1$:

Step C2 Denote by $\{Y_{ij}\}_{j=1}^{N_{im}}$ the observations that have been allocated to $\mathbf{x}_i$ $(i=1,\ldots,I)$. Then $\sum_{i=1}^{I} N_{im}=m$. Estimate the standard deviations:

$$\hat{\sigma}_{im}=\sqrt{\frac{\sum_{j=1}^{N_{im}}(Y_{ij}-\overline{Y}_{im})^2}{N_{im}-1}}. \tag{19}$$

We use this estimate – which does not employ the regression structure but instead records only the variation around the group average $\overline{Y}_{im}=\sum_{j=1}^{N_{im}}Y_{ij}/N_{im}$ – for robustness against the possibility of misspecifying this structure, as well as for its more tractable asymptotic properties. There is of course a possible loss in efficiency if the mean structure is correctly specified; it is however typically true that this is a small premium to pay for the insurance, in the form of robustness, so obtained. This analogy, often employed in robustness studies, is due to Anscombe (1960).

We say that a case $\mathbf{x}_i$ is 'under-represented' if $N_{im}/m \le c/(I\sqrt{m})$. If any case is under-represented, then the $(m+1)$th observation is made at the most under-represented case $\mathbf{x}_{i^*}$, where $i^* = \arg\min N_{im}$. If i^* is not unique, it is randomly chosen from among the most under-represented cases. If all N_{im}/m exceed $c/(I\sqrt{m})$, then the $(m+1)$th location is determined by Theorem 3, with the σ_i replaced by their current estimates $\hat{\sigma}_{im}$. (An alternate, but much slower, method is to recompute a design, using the theory of Section 2, with the σ_i replaced by their current estimates $\hat{\sigma}_{im}$. This yields an allocation vector λ_m, and then i^* is randomly chosen from the multinomial distribution assigning probabilities $\lambda_{m1}, \ldots, \lambda_{mI}$ to $1, \ldots, I$, respectively.)

Step C3 Make the $(m+1)$th observation. Increment m and repeat Step C2.

For $m = n_{init}, \ldots, n-1$:

Step C4 The $(m+1)$th location is determined by Theorem 3, with the σ_i replaced by their estimates $\hat{\sigma}_{in_{init}}$. Make the $(m+1)$th observation. Increment m and repeat.

Step C5 Estimate σ_i by $\hat{\sigma}_{in}$ as at (19). Carry out a weighted least squares regression, using weights $w_i = \hat{\sigma}_{in}^{-2}$.

Let $\{\hat{\lambda}^{(n)}\}$ be the design constructed as in Steps C1–C4 above, define $\hat{\boldsymbol{\sigma}}_n = (\hat{\sigma}_{1n}, \ldots, \hat{\sigma}_{In})^T$, and $\hat{\boldsymbol{\Sigma}}_n = diag(\hat{\sigma}_{1n}, \ldots, \hat{\sigma}_{In})$. The estimated loss function is

$$L(\hat{\lambda}^{(n)}; \hat{\boldsymbol{\sigma}}_n) = tr\{(\mathbf{X}^T \hat{\boldsymbol{\Sigma}}_n^{-1/2} \hat{\lambda}^{(n)} \hat{\boldsymbol{\Sigma}}_n^{-1/2} \mathbf{X})^{-1}\}.$$

To ensure the consistency of $\hat{\sigma}_{in}$ and $L(\hat{\lambda}^{(n)}; \hat{\boldsymbol{\sigma}}_n)$, we require two additional assumptions. In the notation of Section 1 , put $\varepsilon_{ij} = (Y_{ij} - \mathbf{x}_i^T \beta)/\sigma_i$ and suppose that the ε_{ij} are i.i.d., with

$$E[|\varepsilon_{ij}|^{4+\delta}] \overset{def}{=} \kappa < \infty \quad \text{for some } \delta > 0. \tag{20}$$

Further, the size n_{init} of the initial sample should satisfy

$$\lim_{n\to\infty} \frac{(\log n)^2}{n_{init}} = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{n_{init}}{n} = 0. \tag{21}$$

Theorem 4. *Assume that the ε_{ij} are i.i.d., and that (20) and (21) holds. Then as $n \to \infty$,*

(i) $\hat{\sigma}_{in} \xrightarrow{pr} \sigma_i$ *for* $i = 1, \ldots, I$; *and*

(ii) $L(\hat{\lambda}^{(n)}; \hat{\boldsymbol{\sigma}}_n) \xrightarrow{pr} \min_{\lambda \in \mathcal{P}} L(\lambda)$.

5. Case study

Guess and Crump (1977) and Guess et al. (1977) discuss experiments in which animals are subjected to doses, at varying levels, of certain carcinogens. The general purpose is efficient estimation of the probability $P(x)$ of a particular response – the development of a tumour prior to death – at a dose x, and the construction of confidence intervals on the predictions. For this problem designs minimizing the asymptotic variance of the predicted value *at one level* were constructed by Hoel and Jennrich (1979). If instead one is interested in minimizing the sum of the variances of the predictions at all fitted levels, then V-optimality is the appropriate principle.

Following these authors we adopt a model $P(x) = 1 - e^{-Q(x)}$, where $Q(x)$ is a cubic polynomial in the dose level x. The coefficients of this polynomial may be estimated by fitting the cubic model, by weighted least squares, to $y_i = -\log(1 - p(x_i))$. Here $p(x_i)$ is the observed proportion of animals exhibiting the response at level x_i.

We use the framework of the 'benzopyrene' experiment of Guess and Crump (1977), in which n_i animals are each exposed to levels x_i of the known carcinogen benzopyrene. They took $x_i = 6 \cdot 2^{i-1}$, $i = 1, 2, 3, 4$ and $n_1 = \cdots = n_4 = 300$, so that $n = 1200$.

We consider both static and adaptive designs for this experiment. To construct static designs, we begin by taking a prior form of $Q(x)$ as in Guess and Crump (1977):

$$P(x) = 1 - \exp\{-.000097x^2 - .0000017x^3\}. \tag{22}$$

From this we obtain prior estimates of the asymptotic variances; these are $\sigma_i^2 = P(x_i)/(1 - P(x_i))$. Then the theory of Section 2 is applied, to obtain a V-optimal design. We constructed two designs in this manner. In the first we entertained only the four levels $\{6, 12, 24, 48\}$ of x as above, and obtained the optimal design weights $\lambda = \{0.0521, 0.1094, 0.2408, 0.5977\}$ and group sizes, if $n = 1200$, of $n_1 = 63, n_2 = 131, n_3 = 289, n_4 = 717$. In the second we began with a larger collection of possible levels: $\{3, 6, 9, 12, 18, 24, 36, 48\}$. In this case there is no V-optimal four-point design, but the five-point design with weights $\lambda = \{0.0252, 0.1293, 0.2594, 0.1145, 0.4717\}$ on $x = \{3, 9, 24, 36, 48\}$ is V-optimal.

To illustrate the theory of Section 4, we simulated data Y_{ij}, as required, from the $N(Q(x_i), P(x_i)/(1 - P(x_i)))$ distribution, with $P(x)$ as at (22) and $x_i \in \{6, 12, 24, 48\}$. In practice of course $P(x)$ might not be given, but would not be needed as long as data were being collected. Then steps C1–C4 of Section 4 were carried out until 1200 observations had been allocated. In the notation of Section 4 we used $n_{init}=40$, $c=5$ and obtained the final – and very nearly V-optimal – design $\lambda=\{0.0558, 0.1108, 0.2275, 0.6058\}$, with group sizes $n_1=67, n_2=133, n_3=273, n_4=727$.

Acknowledgments

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

Appendix A. Derivations

Proof of Theorem 1. (a) For any $\lambda_0, \lambda_1 \in \mathcal{P}$ define, for $t \in [0,1]$, $\lambda_t = (1-t)\lambda_0 + t\lambda_1$. From (6) and Lemma 2 of Wiens (1993, p. 63) it follows that $L(\lambda_t)$ is a convex function of t, hence that $L(\lambda)$ is a convex function of λ. Thus a minimum is attained, and the necessary and sufficient condition for a minimum at λ_0 is

$$\frac{\partial}{\partial t} L(\lambda_t)_{|_{t=0}} \geq 0 \quad \text{for all } \lambda_1 \in \mathcal{P}. \tag{A.1}$$

We find that $(\partial L(\lambda_t)/\partial t)_{|_{t=0}} = -(\alpha_1(\lambda_0), \ldots \alpha_I(\lambda_0))(\lambda_1 - \lambda_0)$, whence (A.1) is equivalent to

$$\sum_{i=1}^{I} \alpha_i(\lambda_0)(\lambda_{1i} - \lambda_{0i}) \leq 0 \quad \text{for all } \lambda_1 \in \mathcal{P}. \tag{A.2}$$

Split the sum into terms for which $\alpha_i = \alpha_{\max}$ and terms for which $\alpha_i < \alpha_{\max}$. After a rearrangement, (A.2) becomes

$$\sum_{\alpha_i < \alpha_{\max}} (\alpha_{\max}(\lambda_0) - \alpha_i(\lambda_0))(\lambda_{1i} - \lambda_{0i}) \geq 0 \quad \text{for all } \lambda_1 \in \mathcal{P}. \tag{A.3}$$

This requires λ_{0i} to vanish on the set $\{\alpha_i < \alpha_{\max}\}$, and then (A.3) is equivalent to (9).

(b) Suppose that (9) holds. Then (10) follows from (9) and (8). Now (11) follows since, if it fails, then there is a design λ for which

$$L(\lambda_0) = \alpha_{\max}(\lambda_0) > \alpha_{\max}(\lambda) \geq E_\lambda[\alpha(\lambda)] = L(\lambda),$$

contradicting the optimality of λ_0. □

Proof of Theorem 2. (a) Let λ have d points of support. Relabel the rows of $\mathbf{X}$ in such a way that $\mathbf{X}_0(\mathcal{I}_\lambda)$ forms the *first* d rows. Then the other rows are linear combinations of these, and so we have

$$\mathbf{X} = \begin{pmatrix} \mathbf{X_0} \\ \mathbf{PX_0} \end{pmatrix} = \begin{pmatrix} \mathbf{I_d} \\ \mathbf{P} \end{pmatrix} \mathbf{X}_0,$$

for some $(I-d) \times d$ matrix $\mathbf{P}$ with $\mathbf{I}_d + \mathbf{P}^T\mathbf{P} = (\mathbf{X}_0\mathbf{X}_0^T)^{-1}$. With $\mathbf{\Lambda}_0 = diag(\lambda_1, \ldots, \lambda_d)$ and $\mathbf{\Sigma}$ partitioned compatibly as $\mathbf{\Sigma}_0 \oplus \mathbf{\Sigma}_1$, we have

$$\mathbf{C}_\lambda = \begin{pmatrix} \mathbf{I_d} \\ \mathbf{P} \end{pmatrix} \mathbf{X}_0 (\mathbf{X}_0^T \mathbf{\Sigma}_0^{-1/2} \mathbf{\Lambda}_0 \mathbf{\Sigma}_0^{-1/2} \mathbf{X}_0)^{-1} \mathbf{X}_0^T (\mathbf{I}_d, \mathbf{P}^T) = \begin{pmatrix} \mathbf{I_d} \\ \mathbf{P} \end{pmatrix} \mathbf{\Sigma}_0^{-1/2} \mathbf{\Lambda}_0 \mathbf{\Sigma}_0^{-1/2} (\mathbf{I}_d, \mathbf{P}^T),$$

and

$$\begin{aligned} \mathbf{\Sigma}^{-1/2}\mathbf{C}_\lambda^2\mathbf{\Sigma}^{-1/2} &= \begin{pmatrix} \mathbf{\Sigma_0^{-1/2}} \\ \mathbf{\Sigma_1^{-1/2}P} \end{pmatrix} \left\{ \begin{array}{c} \mathbf{\Sigma}_0^{1/2}\mathbf{\Lambda}_0^{-1}\mathbf{\Sigma}_0^{1/2} \cdot (\mathbf{I}_d + \mathbf{P}^T\mathbf{P}) \\ \cdot \mathbf{\Sigma}_0^{1/2}\mathbf{\Lambda}_0^{-1}\mathbf{\Sigma}_0^{1/2}(\mathbf{\Sigma}_0^{-1/2}, \mathbf{P}^T\mathbf{\Sigma}_1^{-1/2}) \end{array} \right\} \\ &= \begin{pmatrix} \mathbf{\Lambda}_0^{-1}\mathbf{\Sigma}_0^{1/2}(\mathbf{X}_0\mathbf{X}_0^T)^{-1}\mathbf{\Sigma}_0^{1/2}\mathbf{\Lambda}_0^{-1} & \times \\ \times & \begin{array}{c} \mathbf{\Sigma}_1^{-1/2}\mathbf{P}\mathbf{\Sigma}_0^{1/2}\mathbf{\Lambda}_0^{-1}\mathbf{\Sigma}_0^{1/2} \cdot (\mathbf{I}_d + \mathbf{P}^T\mathbf{P}) \\ \cdot \mathbf{\Sigma}_0^{1/2}\mathbf{\Lambda}_0^{-1}\mathbf{\Sigma}_0^{1/2}\mathbf{P}^T\mathbf{\Sigma}_1^{-1/2} \end{array} \end{pmatrix}. \end{aligned}$$

Thus $\alpha_i(\lambda) = (\sigma_i^2/\lambda_i^2) \| \mathbf{x}_0^{(i)} \|^2$ for $i=1,\ldots,d$. It follows that if (9) is to hold with $\sum_{i=1}^d \lambda_{0i} = 1$ then (12) and (13) are necessary. We require as well that $\alpha_i(\lambda_0) \leq \alpha_{\max}$ if $i \notin \mathcal{I}_{\lambda_0}$, i.e. that for $i = d+1, \ldots, I$,

$$\alpha_{\max} \geq \begin{bmatrix} \mathbf{\Sigma}_1^{-1/2}\mathbf{P}\mathbf{\Sigma}_0^{1/2}\mathbf{\Lambda}_0^{-1}\mathbf{\Sigma}_0^{1/2}(\mathbf{I}_d + \mathbf{P}^T\mathbf{P}) \\ \cdot \mathbf{\Sigma}_0^{1/2}\mathbf{\Lambda}_0^{-1}\mathbf{\Sigma}_0^{1/2}\mathbf{P}^T\mathbf{\Sigma}_1^{-1/2} \end{bmatrix}_{i-d,i-d}. \tag{A.4}$$

A calculation shows that (A.4) is equivalent to

$$1 \geq \left\| \mathbf{H}_0 \frac{\mathbf{x}_i}{\sigma_i} \right\| \quad \text{for } i > d. \tag{A.5}$$

Since $\|\mathbf{H}_0\mathbf{x}_i/\sigma_i\| = 1$ for $i = 1, \ldots, d$, (A.5) is equivalent to the statement in (a) (ii). That (a) (i), (ii) imply Theorem 1 (a) is now immediate, as is the necessity of (14). □

Proof of Theorem 3. If the next observation is made at location '*i*', then the new vector and matrix of allocations are

$$\lambda^{(n+1)} = \frac{n}{n+1}\lambda^{(n)} + \frac{1}{n+1}\mathbf{e}_i \quad \text{and} \quad \boldsymbol{\Lambda}^{(n+1)} = \frac{n}{n+1}\boldsymbol{\Lambda}^{(n)} + \frac{1}{n+1}\mathbf{e}_i\mathbf{e}_i^T.$$

With $\mathbf{u}_i \overset{def}{=} (\mathbf{V}^T\boldsymbol{\Lambda}^{(n)}\mathbf{V})^{-1/2}\mathbf{V}^T\mathbf{e}_i/\sqrt{n}$ we calculate that

$$\begin{aligned}(\mathbf{V}^T\boldsymbol{\Lambda}^{(n+1)}\mathbf{V})^{-1} &= \frac{n+1}{n}((\mathbf{V}^T\boldsymbol{\Lambda}^{(n)}\mathbf{V})^{1/2}[\mathbf{I}+\mathbf{u}_i\mathbf{u}_i^T](\mathbf{V}^T\boldsymbol{\Lambda}^{(n)}\mathbf{V})^{1/2})^{-1} \\ &= \frac{n+1}{n}\left\{(\mathbf{V}^T\boldsymbol{\Lambda}^{(n)}\mathbf{V})^{-1} - \frac{(\mathbf{V}^T\boldsymbol{\Lambda}^{(n)}\mathbf{V})^{-1}\mathbf{V}^T\mathbf{e}_i\mathbf{e}_i^T\mathbf{V}(\mathbf{V}^T\boldsymbol{\Lambda}^{(n)}\mathbf{V})^{-1}}{n+\frac{\mathbf{C}_{n,ii}}{\sigma_i^2}}\right\}.\end{aligned}$$

Then

$$\mathbf{C}_{n+1} = \boldsymbol{\Sigma}^{1/2}\mathbf{V}(\mathbf{V}^T\boldsymbol{\Lambda}^{(n+1)}\mathbf{V})^{-1}\mathbf{V}^T\boldsymbol{\Sigma}^{1/2} = \frac{n+1}{n}\left\{\mathbf{C}_n - \frac{\mathbf{c}_{n,i}\mathbf{c}_{n,i}^T}{n\sigma_i^2+\mathbf{C}_{n,ii}}\right\}.$$

This is (16); (17) follows from the observation that $tr(\mathbf{c}_{n,i}\mathbf{c}_{n,i}^T) = [\mathbf{C}_n^2]_{ii}$.

To establish (18) we use the notation of the proof of Theorem 1. Thus, let λ_0 denote a minimizer of $L(\lambda)$ over $\mathcal{P}$. We make frequent use of the fact, which follows from (8) and (11), that

$$L(\lambda) \le \alpha_{\max}(\lambda), \quad \text{with equality if } \lambda = \lambda_0. \tag{A.6}$$

Let λ_1 denote any member of $\mathcal{P}$. Then since $L(\lambda_t)$ is a convex function of t,

$$L(\lambda_1) - L(\lambda_0) = \int_0^1 \left(\frac{d}{dt}L(\lambda_t)\right)dt \le \int_0^1 \left(\frac{d}{dt}L(\lambda_t)_{|_{t=1}}\right)dt = \frac{d}{dt}L(\lambda_t)_{|_{t=1}}. \tag{A.7}$$

A calculation, followed by (A.6), gives

$$\frac{d}{dt}L(\lambda_t)_{|_{t=1}} = \sum_{i=1}^{I}(\lambda_{0i}-\lambda_{1i})\alpha_i(\lambda_1) = \sum_{i=1}^{I}\lambda_{0i}\alpha_i(\lambda_1) - L(\lambda_1) \le \alpha_{\max}(\lambda_1) - L(\lambda_1);$$

this together with (A.7) results in

$$L(\lambda_1) \le \frac{\alpha_{\max}(\lambda_1)+\alpha_{\max}(\lambda_0)}{2}. \tag{A.8}$$

Now replace λ_1 by $\lambda^{(n)}$ and substitute (A.8) into (17) to obtain

$$L\left(\lambda^{(n+1)}\right) - L(\lambda^{(n)}) \le \frac{1}{2n}\left[L(\lambda_0) - b_n \cdot \alpha_{\max}\left(\lambda^{(n)}\right)\right], \tag{A.9}$$

for

$$b_n = \left[1+\frac{1}{n+1}\left(1-\frac{\mathbf{C}_{n,i^{(n)}i^{(n)}}}{\sigma_{i^{(n)}}^2}\right)\right] \Big/ 1-\frac{1}{n+1}\left(1-\frac{\mathbf{C}_{n,i^{(n)}i^{(n)}}}{\sigma_{i^{(n)}}^2}\right)..$$

We shall require that

$$b_n \to 1 \quad \text{as } n \to \infty. \tag{A.10}$$

Since b_n depends only on those locations at which observations will be made, we can assume that $\lambda_{i^{(n)}}^{(n)} > 0$. Then

$$\frac{\mathbf{C}_{n,i^{(n)}i^{(n)}}}{\sigma_{i^{(n)}}^2} = [\mathbf{V}(\mathbf{V}^T\boldsymbol{\Lambda}\mathbf{V})^{-1}\mathbf{V}^T]_{i^{(n)}i^{(n)}} = \frac{[\boldsymbol{\Lambda}^{1/2}\mathbf{V}(\mathbf{V}^T\boldsymbol{\Lambda}\mathbf{V})^{-1}\mathbf{V}^T\boldsymbol{\Lambda}^{1/2}]_{i^{(n)}i^{(n)}}}{\lambda_{i^{(n)}}^{(n)}}.$$

The matrix $\boldsymbol{\Lambda}^{1/2}\mathbf{V}(\mathbf{V}^T\boldsymbol{\Lambda}\mathbf{V})^{-1}\mathbf{V}^T\boldsymbol{\Lambda}^{1/2}$ is idempotent, hence its diagonal elements lie in (0, 1). Thus $0 < \mathbf{C}_{n,i^{(n)}i^{(n)}}/\sigma_{i^{(n)}}^2 < 1/\lambda_{i^{(n)}}^{(n)}$ and (A.10) will follow from $n\lambda_{i^{(n)}}^{(n)} \to \infty$. This in turn is clear – it says merely that the frequency of those, finitely many, locations to which assignments continue to be made, must necessarily become arbitrarily large.

We must now show that given $\varepsilon > 0$ there is n^* such that

$$L(\lambda^{(n)}) \le L(\lambda_0) + \varepsilon \quad \text{for } n \ge n^*. \tag{A.11}$$

Divide the sequence $\{\lambda^{(n)}\}_{n \ge n_0}$ into disjoint subsequences $\mathcal{S}_1(\varepsilon)$ and $\mathcal{S}_2(\varepsilon)$ such that

$$\begin{aligned}&\lambda^{(n)} \in \mathcal{S}_1(\varepsilon) \Leftrightarrow L(\lambda^{(n)}) \le L(\lambda_0)+\varepsilon/2, \\ &\lambda^{(n)} \in \mathcal{S}_2(\varepsilon) \Leftrightarrow L(\lambda^{(n)}) > L(\lambda_0)+\varepsilon/2.\end{aligned}$$

We first show that $\mathcal{S}_1(\varepsilon)$ is non-empty for each ε; this will imply that we can find $L(\lambda^{(n)})$ arbitrarily close to $L(\lambda_0)$.

Suppose that $\mathcal{S}_1(\varepsilon)$ is empty. Then for all n we have

$$\alpha_{\max}(\lambda^{(n)}) \ge L(\lambda^{(n)}) > L(\lambda_0)+\frac{\varepsilon}{2}.$$

By (A.10), for $\delta \in (0,1]$ there is n_δ such that $n>n_\delta$ implies that $b_n > 1-\delta$. Then from (A.9), and with $n>n_\delta$ for $\delta = \varepsilon/\left(4\left(L(\lambda_0)+\frac{\varepsilon}{2}\right)\right)$ we have

$$L\left(\lambda^{(n+1)}\right) - L(\lambda^{(n)}) \le \frac{1}{2n}\left[L(\lambda_0) - b_n \cdot \left(L(\lambda_0)+\frac{\varepsilon}{2}\right)\right] < \frac{1}{2n}\left[\delta\left(L(\lambda_0)+\frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2}\right] = -\frac{\varepsilon}{8n}, \tag{A.12}$$

implying that

$$L(\lambda^{(n+1)}) = L(\lambda^{(n_\delta)}) + \sum_{m=n_\delta}^{n} (L(\lambda^{(m+1)}) - L(\lambda^{(m)})) \to -\infty,$$

as $n \to \infty$, a contradiction.

We have shown that there is a sequence $\{\lambda^{(n_l)}\}_{l=1}^{\infty}$ in $\mathcal{S}_1(\varepsilon)$ with $L(\lambda^{(n_l)}) \to L(\lambda_0)$. By (A.9) we have that, for $n_l > n_\delta$ with $\delta = 1$,

$$L\left(\lambda^{(n_l+1)}\right) - L\left(\lambda^{(n_l)}\right) \le \frac{L(\lambda_0)}{2n_l},$$

and so we can find n' such that for $n_l > n'$ and $\lambda^{(n_l)} \in \mathcal{S}_1(\varepsilon)$ we have

$$L(\lambda^{(n_l+1)}) \le L(\lambda_0) + \varepsilon, \tag{A.13}$$

even if $\lambda^{(n_l+1)} \in \mathcal{S}_2(\varepsilon)$. Furthermore, as at (A.12), we can find n'' such that for all $\lambda^{(n_k)} \in \mathcal{S}_2(\varepsilon)$ with $n_k > n''$ we have

$$L(\lambda^{(n_k+1)}) < L(\lambda^{(n_k)}). \tag{A.14}$$

Now select $n^* > \max(n', n'')$ for which $\lambda^{(n^*)} \in \mathcal{S}_1(\varepsilon)$, then (A.(13) and A.14) imply (A.11), whether $\lambda^{(n)} \in \mathcal{S}_1(\varepsilon)$ or $\lambda^{(n)} \in \mathcal{S}_2(\varepsilon)$. □

A version of the key inequality (A.8), used by Wynn (1970) in his construction of sequentially D-optimal designs, was established by Kiefer (1961).

Proof of Theorem 4. Define $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ and let $\boldsymbol{\mu}_1, \boldsymbol{m}_2$ and $\boldsymbol{\sigma}$ be $I \times 1$ vectors with elements $\mu_i, \mu_i^2 + \sigma_i^2$ and σ_i respectively. We denote by Y_n the nth response in the forced design, with associated random error ε_n. Let $\boldsymbol{i}_n$ be the $I \times 1$ incidence vector, with elements $\boldsymbol{i}_{ni} = I(Y_n$ is observed at $\mathbf{x}_i)$. Then Y_n can be written as $Y_n = \boldsymbol{i}_n^T \boldsymbol{\mu}_1 + \eta_n$ with $\eta_n = \boldsymbol{i}_n^T \boldsymbol{\sigma} \varepsilon_n$. This defines the sequence $\{\varepsilon_n\}$ of i.i.d. random variables. Furthermore, define a r.v. γ_n by $Y_n^2 = \boldsymbol{i}_n^T \boldsymbol{m}_2 + \gamma_n$; viz., $\gamma_n = 2\boldsymbol{i}_n^T \boldsymbol{\mu}_1 \boldsymbol{i}_n^T \boldsymbol{\sigma} \varepsilon_n + (\boldsymbol{i}_n^T \boldsymbol{\sigma})^2 (\varepsilon_n^2 - 1)$.

(i) To establish the consistency of $\hat{\sigma}_{in}^2$, $i = 1,\dots,I$, we first note that by the nature of the forcing in Step C2, N_{in} is at least of the order $n_{init}^{1/2}$. Thus

$$\hat{\sigma}_{in}^2 = \left(\frac{1}{N_{in}} \sum_{j=1}^{N_{in}} Y_{ij}^2 - \overline{Y}_{in}^2\right)\left(1 + O\left(n_{init}^{-1/2}\right)\right),$$

and so it suffices to show that

$$\overline{Y}_{in} \overset{pr}{\to} \mu_i \quad \text{and} \quad \frac{1}{N_{in}} \sum_{j=1}^{N_{in}} Y_{ij}^2 \overset{pr}{\to} \mu_i^2 + \sigma_i^2. \tag{A.15}$$

Let $\mathbf{U}_n = (\overline{Y}_{1n}, \dots, \overline{Y}_{In})^T$ and $\mathbf{V}_n = (\frac{1}{N_{in}} \sum_{j=1}^{N_{in}} Y_{1j}^2, \dots, \frac{1}{N_{in}} \sum_{j=1}^{N_{in}} Y_{Ij}^2)^T$. Then (A.15) is equivalent to

$$\mathbf{U}_n \overset{pr}{\to} \boldsymbol{\mu}_1 \quad \text{and} \quad \mathbf{V}_n \overset{pr}{\to} \boldsymbol{m}_2. \tag{A.16}$$

Since $\mathbf{U}_n$ and $\mathbf{V}_n$ can be respectively written as

$$\mathbf{U}_n = \left(\sum_{j=1}^{n} \boldsymbol{i}_j \boldsymbol{i}_j^T\right)^{-1} \sum_{j=1}^{n} \boldsymbol{i}_j Y_j = \boldsymbol{\mu}_1 + \left(\sum_{j=1}^{n} \boldsymbol{i}_j \boldsymbol{i}_j^T\right)^{-1} \sum_{j=1}^{n} \boldsymbol{i}_j \eta_j,$$

and

$$\mathbf{V}_n = \left(\sum_{j=1}^{n} \boldsymbol{i}_j \boldsymbol{i}_j^T\right)^{-1} \sum_{j=1}^{n} \boldsymbol{i}_j Y_j^2 = \boldsymbol{m}_2 + \left(\sum_{j=1}^{n} \boldsymbol{i}_j \boldsymbol{i}_j^T\right)^{-1} \sum_{j=1}^{n} \boldsymbol{i}_j \gamma_j,$$

these vectors $\mathbf{U}_n$ and $\mathbf{V}_n$ can be viewed as the least square estimates in stochastic regression (Lai and Wei, 1982). Theorem 1 of Lai and Wei (1982) then implies that (A.16) holds under the following assumptions:

(D1) $\boldsymbol{i}_n$ is $\mathcal{F}_{n-1} mbox-$measurable with $\{\mathcal{F}_n\}$ being an increasing sequence of σ-fields;

(D2) $\rho_{min}(n) \to \infty$ a.s. and $\log \rho_{max}(n) = o(\rho_{min}(n))$ a.s., where $\rho_{min}(n)$ and $\rho_{max}(n)$ are the minimum and maximum eigenvalues of $\sum_{j=1}^{n} \boldsymbol{i}_j \boldsymbol{i}_j^T$, respectively;

(D3) $\{\eta_n\}$ and $\{\gamma_n\}$ are martingale difference sequences with respect to $\{\mathcal{F}_n\}$;

(D4) for some $\alpha > 2$, we have that $\sup_n E(|\eta_n|^\alpha | \mathcal{F}_{n-1}) < \infty$ and $\sup_n E(|\gamma_n|^\alpha | \mathcal{F}_{n-1}) < \infty$.

To verify (D1), let $\{\mathcal{F}_n\}$ denote the σ-field generated by $\varepsilon_1, \ldots, \varepsilon_n$. An important feature of the forced design is that $\boldsymbol{i}_n$ depends only on $Y_1, \ldots, Y_{n-1}$ and therefore only on $\varepsilon_1, \ldots, \varepsilon_{n-1}$, hence is $\mathcal{F}_{n-1}$-measurable. For (D2), since $\sum_{j=1}^{n} \boldsymbol{i}_j \boldsymbol{i}_j^T$ is a diagonal matrix with diagonal elements $N_{1n}, \ldots, N_{In}$, and due to the forcing in Step C2 and the assumption in (21),

$$\rho_{min}(n) = \min_i N_{in} \overset{a.s.}{\to} \infty, \frac{\log \rho_{max}(n)}{\rho_{min}(n)} = \frac{\log \max_i N_{in}}{\min_i N_{in}} \le \frac{\log n}{c\sqrt{n_{init}}/I} \overset{a.s.}{\to} 0.$$

For (D3), note that (D1) implies that $\eta_n = \boldsymbol{i}_n^T \boldsymbol{\sigma} \varepsilon_n$ is $\mathcal{F}_n$-measurable; then

$$E[\eta_n | \mathcal{F}_{n-1}] = E[\boldsymbol{i}_n^T \boldsymbol{\sigma} \varepsilon_n | \mathcal{F}_{n-1}] = \boldsymbol{i}_n^T \boldsymbol{\sigma} E[\varepsilon_n | \mathcal{F}_{n-1}] = 0.$$

That is, $\{\eta_n\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_n\}$. Similarly, we can check that $\{\gamma_n\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_n\}$. Finally, (D4) follows from (20) and the observation that

$$\sup_n E(|\eta_n|^{4+\delta} | \mathcal{F}_{n-1}) \le \sup_n \{(\boldsymbol{i}_n^T \boldsymbol{\sigma})^{4+\delta} E(|\varepsilon_n|^{4+\delta})\} \le (\mathbf{1}^T \boldsymbol{\sigma})^{4+\delta} \kappa < \infty,$$

since $\{\varepsilon_n\}$ can be viewed as a sequence of i.i.d. random variables with the same distribution as the ε_{ij}. That $\sup_n E(|\gamma_n|^{2+\delta/2} | \mathcal{F}_{n-1}) < \infty$ is shown in a similar fashion.

(*ii*) Let $\hat{\boldsymbol{\sigma}}_{n_{init}}$ be the estimate of $\boldsymbol{\sigma}$ after Step C3. From (i), we have that

$$\hat{\boldsymbol{\sigma}}_{n_{init}} \overset{pr}{\to} \boldsymbol{\sigma} \tag{A.17}$$

as $n \to \infty$. The proof of (ii) consists of verifying that

(E1) $\min_{\lambda \in \mathcal{P}} L(\lambda; \hat{\boldsymbol{\sigma}}_{n_{init}}) \overset{pr}{\to} \min_{\lambda \in \mathcal{P}} L(\lambda; \boldsymbol{\sigma})$ as $n \to \infty$;
(E2) $L(\hat{\lambda}^{(n)}; \hat{\boldsymbol{\sigma}}_{n_{init}}) - \min_{\lambda \in \mathcal{P}} L(\lambda; \hat{\boldsymbol{\sigma}}_{n_{init}}) \overset{pr}{\to} 0$ as $n \to \infty$;
(E3) $L(\hat{\lambda}^{(n)}; \hat{\boldsymbol{\sigma}}_n) - L(\hat{\lambda}^{(n)}; \hat{\boldsymbol{\sigma}}_{n_{init}}) \overset{pr}{\to} 0$ as $n \to \infty$.

To verify (E1), let $\lambda^*_{n_{init}}$ be a design such that $L(\lambda^*_{n_{init}}; \hat{\boldsymbol{\sigma}}_{n_{init}}) = \min_{\lambda \in \mathcal{P}} L(\lambda; \hat{\boldsymbol{\sigma}}_{n_{init}})$ and let λ_0 be a design such that $L(\lambda_0; \boldsymbol{\sigma}) = \min_{\lambda \in \mathcal{P}} L(\lambda; \boldsymbol{\sigma})$. Then we have that $L(\lambda^*_{n_{init}}; \hat{\boldsymbol{\sigma}}_{n_{init}}) \le L(\lambda_0; \hat{\boldsymbol{\sigma}}_{n_{init}})$. Since $L(\lambda_0; \boldsymbol{\sigma})$ is continuous with respect to $\boldsymbol{\sigma}$, (A.17) implies that

$$L(\lambda^*_{n_{init}}; \hat{\boldsymbol{\sigma}}_{n_{init}}) \le L(\lambda_0; \hat{\boldsymbol{\sigma}}_{n_{init}}) \overset{pr}{\to} L(\lambda_0; \boldsymbol{\sigma}) = \min_{\lambda \in \mathcal{P}} L(\lambda; \boldsymbol{\sigma}). \tag{A.18}$$

Next we derive a lower bound for $L(\lambda^*_{n_{init}}; \hat{\boldsymbol{\sigma}}_{n_{init}})$. Define matrices

$$\mathbf{C}_{n_{init}} = \mathbf{X}^T \hat{\boldsymbol{\Sigma}}_{n_{init}}^{-1/2} \lambda^*_{n_{init}} \hat{\boldsymbol{\Sigma}}_{n_{init}}^{-1/2} \mathbf{X} = \sum_{i=1}^{I} \frac{\lambda^*_{in_{init}}}{\hat{\sigma}^2_{in_{init}}} \mathbf{x}_i \mathbf{x}_i^T \quad \text{and} \quad \mathbf{A}_{n_{init}}(\lambda) = \sum_{i=1}^{I} \frac{\lambda^*_{in_{init}}}{\sigma_i^2} \mathbf{x}_i \mathbf{x}_i^T.$$

Note that by (A.17),

$$a_{n_{init}} \overset{def}{=} \max_i \left\{ \left| \frac{\sigma_i^2}{\hat{\sigma}^2_{in_{init}}} - 1 \right| \Big/ \sigma_i^2 \right\} \overset{pr}{\to} 0 \mathrm{as} n \to \infty. \tag{A.19}$$

Some algebra (which uses (5)) leads to

$$\mathbf{C}_{n_{init}} \le \mathbf{A}_{n_{init}} + a_{n_{init}} \mathbf{I}_d, \tag{A.20}$$

here we use the Loewner ordering: $\mathbf{A} \le \mathbf{B}$ iff $\mathbf{B} - \mathbf{A}$ is positive semidefinite. Denote by $\rho_l(\mathbf{A})$ the l-th eigenvalue of a matrix $\mathbf{A}$. Then

$$\begin{aligned} L\left(\lambda^*_{n_{init}}; \hat{\boldsymbol{\sigma}}_{n_{init}}\right) &= tr\left[\mathbf{C}_{n_{init}}^{-1}\right] \ge tr\left[(\mathbf{A}_{n_{init}} + a_{n_{init}} \mathbf{I}_d)^{-1}\right] \\ &= \sum_{l=1}^{d} \frac{1}{\rho_l(\mathbf{A}_{n_{init}}) + a_{n_{init}}} \ge \sum_{l=1}^{d} \frac{1}{\rho_l(\mathbf{A}_{n_{init}})} - a_{n_{init}} \sum_{l=1}^{d} \frac{1}{\rho_l^2(\mathbf{A}_{n_{init}})}. \end{aligned} \tag{A.21}$$

Note that

$$\begin{aligned} \sum_{l=1}^{d} \frac{1}{\rho_l(\mathbf{A}_{n_{init}})} &= tr\left[\mathbf{A}_{n_{init}}^{-1}\right] = L\left(\lambda^*_{n_{init}}; \boldsymbol{\sigma}\right) \\ &\ge L(\lambda_0; \boldsymbol{\sigma}) = \min_{\lambda \in \mathcal{P}} L(\lambda; \boldsymbol{\sigma}). \end{aligned} \tag{A.22}$$

We claim that each $\rho_l(\mathbf{A}_{n_{init}})$ has asymptotically a positive lower bound; if so then (A.(21), A.22) and (A.19) will together ensure that

$$L\left(\lambda^*_{n_{init}}; \hat{\boldsymbol{\sigma}}_{n_{init}}\right) \ge \min_{\lambda \in \mathcal{P}} L(\lambda; \boldsymbol{\sigma}) - a_{n_{init}} \sum_{l=1}^{d} \frac{1}{\rho_l^2(\mathbf{A}_{n_{init}})} \overset{pr}{\to} \min_{\lambda \in \mathcal{P}} L(\lambda; \boldsymbol{\sigma}). \tag{A.23}$$

This claim follows from (A.20) and the fact that $L(\lambda_0;\hat{\boldsymbol{\sigma}}_{n_{init}}) = \sum_{l=1}^{d} \rho_l(\mathbf{C}_{n_{init}})$:

$$\rho_l\left(\mathbf{A}_{n_{init}}\right) \ge \rho_l\left(\mathbf{C}_{n_{init}}\right) - a_{n_{init}} \ge \frac{1}{L(\lambda_0;\hat{\boldsymbol{\sigma}}_{n_{init}})} - a_{n_{init}} \overset{pr}{\to} L(\lambda_0;\boldsymbol{\sigma}) > 0.$$

Combining (A.(18) and A.23), we now have $L(\lambda^*_{n_{init}};\hat{\boldsymbol{\sigma}}_{n_{init}}) \overset{pr}{\to} \min_{\lambda\in\mathcal{P}} L(\lambda;\boldsymbol{\sigma})$, completing the proof of (E1).

The verification of (E2) is very similar to the proof of Theorem 3 and is omitted. That of (E3) follows from the continuity of $L(\lambda^{(n)};\boldsymbol{\sigma})$ with respect to $\boldsymbol{\sigma}$ and the consistency of each of $\hat{\boldsymbol{\sigma}}_{n_{init}}, \hat{\boldsymbol{\sigma}}_n$. □

References

Anscombe, F.J., 1960. Rejection of Outliers. Technometrics 2, 123–147.

Antos, A., Grover, V., Szepesvári, C., 2008. Active learning in multi-armed bandits. In: Proceedings of the 19th International Conference, ALT 2008, Budapest, Hungary, October 13–16, 2008. In: Algorithmic Learning Theory: Lecture Notes in Computer Science, Springer, pp. 287–302.

Chaudhuri, P., Mykland, P.A., 1993. Nonlinear Experiments: Optimal Design and Inference Based on Likelihood. Journal of the American Statistical Association 88, 538–546.

Dette, H., Haines, L.M., Imhof, L.A., 2005. Bayesian and maximin optimal designs for heteroscedastic regression models. Canadian Journal of Statistics 33, 221–241.

Fedorov, V.V., 1972. Theory of Optimal Experiments. Academic Press.

Guess, H., Crump, K., 1977. Can we use animal data to estimate safe doses for chemical carcinogens? In: Whittmore, A. (Ed.), Environmental Health: Quantitative Methods, Society for Industrial and Applied Mathematics, pp. 13–28.

Guess, H., Crump, K., Peto, R., 1977. Uncertainly estimates for low dose rate extrapolation of animal carcinogenicity data. Cancer Research 37, 3475–3483.

Heiligers, B., 1996. Computing E-optimal polynomial regression designs. Journal of Statistical Planning and Inference 55, 219–233.

Hoel, P.G., Jennrich, R.I., 1979. Optimal designs for dose response experiments in cancer research. Biometrika 66, 307–316.

Kiefer, J., 1961. Optimal experimental designs V with applications to systematic and rotatable designs. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability I, University of California Press, Berkeley, pp. 381–405.

Lai, T.L., Wei, C.Z., 1982. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. Annals of Statistics 10, 154–166.

Li, K.-H., Chan, N.N., 2001. L_p-optimality for regression designs with heteroscedastic errors. Journal of Statistical Planning and Inference 92, 253–257.

Pukelsheim, F., Torsney, B., 1991. Optimal weights for experimental designs on linearly independent support points. Annals of Statistics 19, 1614–1625.

Pukelsheim, F., Rieder, S., 1992. Efficient rounding of approximate designs. Biometrika 79, 763–770.

Pukelsheim, F., 1993. Optimal Design of Experiments. Wiley.

Wiens, D.P., 1993. Designs for approximately linear regression: maximizing the minimum coverage probability of confidence ellipsoids. Canadian Journal of Statistics 21, 59–70.

Wiens, D.P., 2011. Designs for weighted least squares regression, with estimated weights. Statistics and Computing 23, 391–401.

Wong, W.-K., 1995. On the equivalence of *D* and *G*-optimal designs in heteroscedastic models. Statistics and Probability Letters 25, 317–321.

Wynn, H.P., 1970. The sequential generation of D-optimum experimental design. Annals of Mathematical Statistics 5, 1655–1664.