Journal of Statistical Planning and Inference 139 (2009) 3 – 15

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Robust designs for misspecified logistic models

Adeniyi J. Adewale[a], Douglas P. Wiens[b,*]

[a]Merck Research Laboratories, North Wales, Pennsylvania 19454, United States
[b]Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1

ARTICLE INFO

Available online 24 May 2008
MSC:
primary 62K05;62F35
secondary 62J05

Keywords:
Fisher information
Logistic regression
Linear predictor
Monte Carlo sample
Polynomial
Random walk
Simulated annealing

ABSTRACT

We develop criteria that generate robust designs and use such criteria for the construction of designs that insure against possible misspecifications in logistic regression models. The design criteria we propose are different from the classical in that we do not focus on sampling error alone. Instead we use design criteria that account as well for error due to bias engendered by the model misspecification. Our robust designs optimize the average of a function of the sampling error and bias error over a specified misspecification neighbourhood. Examples of robust designs for logistic models are presented, including a case study implementing the methodologies using beetle mortality data.

1. Introduction

Experimental designs have been treated extensively in the statistical literature, starting with designs for linear models and extending to nonlinear models. A large volume of literature is devoted to designs assuming the exact correctness of the relationship between the response variable and the design (explanatory) variables. Box and Draper (1959) added another dimension to the theory by investigating the impact of model misspecification. Following the work of Box and Draper the literature has since been replete with regression designs which are robust against violations of various model assumptions—linearity of the response, independence and homoscedasticity of the errors, etc. Authors who have considered designs with an eye on the approximate nature of the assumed linear models include Marcus and Sacks (1976), Li and Notz (1982), Wiens (1992) and Wiens and Zhou (1999), to mention but a few.

For nonlinear designs, Fedorov (1972), Ford and Silvey (1980), Chaloner and Larntz (1989) and Chaudhuri and Mykland (1993) have explored the construction of optimal designs while assuming that the nonlinear model of interest is correctly specified. Still others have investigated designs for generalized linear models, a class of possibly nonlinear models in which the response follows a distribution from the exponential family such as the normal, binomial, Poisson or gamma (McCullagh and Nelder, 1989). The expository article (Ford et al., 1989) hinted that in the context of nonlinear models, as in the case of linear models, the misspecification of the model itself is of serious concern. They asserted that "indeed, if the model is seriously in doubt, the forms of design that we have considered may be completely inappropriate." Sinha and Wiens (2002) have explored some designs for nonlinear models with due consideration for the approximate nature of the assumed model. In this work we consider designs for misspecified logistic regression models.

For the logistic model, the mean response $E(Y)=\mu$ depends on the parameters, $\boldsymbol{\beta}$, and the vector of explanatory variables, $\mathbf{x}$, through the nonlinear function $\mu=e^{\eta}/(1+e^{\eta})$, where $\eta=\mathbf{z}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}$. The function η is termed the linear predictor, with regressors $z_1(\mathbf{x}),\ldots,z_p(\mathbf{x})$ depending on the q-dimensional independent variable $\mathbf{x}$. The variance of the response, written $var(Y|\mathbf{x})$, is a

* Corresponding author. Tel.: +1 780 492 4406; fax: +1 780 492 6826.
E-mail addresses: adeniyiadewale@hotmail.com (A.J. Adewale), doug.wiens@ualberta.ca (D.P. Wiens).

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.05.022

nonlinear function of the linear predictor. Abdelbasit and Plackett (1983), Minkin (1987), Ford et al. (1992), Chaudhuri and Mykland (1993), Burridge and Sebastiani (1994), Atkinson and Haines (1996) and King and Wong (2000) have investigated designs for binary data, and in particular for logistic regression. As illustrated in these papers, the general approach to optimal design is to seek a design that optimizes certain functions of the information matrix of the model parameters. The information matrix for $\boldsymbol{\beta}$ from a design consisting of the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is given by

$$\sum_{i=1}^{n} w(\mathbf{x}_i, \boldsymbol{\beta}) \mathbf{z}(\mathbf{x}_i) \mathbf{z}^{\mathrm{T}}(\mathbf{x}_i) = \mathbf{Z}^{\mathrm{T}} \mathbf{W} \mathbf{Z}, \tag{1}$$

where $\mathbf{Z} = (\mathbf{z}(\mathbf{x}_1), \mathbf{z}(\mathbf{x}_2), \ldots, \mathbf{z}(\mathbf{x}_n))^{\mathrm{T}}$ and

$$\mathbf{W} = \mathrm{diag}(w(\mathbf{x}_1, \boldsymbol{\beta}), w(\mathbf{x}_2, \boldsymbol{\beta}), \ldots, w(\mathbf{x}_n, \boldsymbol{\beta}))$$

for weights $w(\mathbf{x}_i, \boldsymbol{\beta}) = (\mathrm{d}\mu/\mathrm{d}\eta_i)^2 / var(Y|\mathbf{x}_i)$. Thus, as with nonlinear experiments the information matrix depends on the unknown parameters $\boldsymbol{\beta}$. Designing an experiment for the estimation of these parameters would then seem to require that these parameters be known. The following are some of the approaches that have been explored in the literature for handling the dependency of the information matrix on $\boldsymbol{\beta}$.

(1) Locally optimal designs: A traditional approach in designing a nonlinear experiment is to aim for maximum efficiency at a best guess (initial estimate) of the parameter values (Chernoff, 1953). Designs that are optimal for given parameter values are dubbed locally optimal designs. These designs may be stable over a range of parameter values. However, if unstable, a design which is optimal for a best guess may not be efficient for parameter values in even a small neighbourhood of this guess.
(2) Bayesian optimal designs: A natural generalization of locally optimal designs is to use a prior distribution on the unknown parameters rather than a single guess. The approach which assumes such a prior and incorporates this distribution into the appropriate design criteria is termed Bayesian optimal design—see Chaloner and Larntz (1989) and Dette and Wong (1996).
(3) Minimax optimal designs: Rather than assume a prior distribution, this approach assumes a range of plausible values for the parameters. The minimax optimal design is the design with the least loss when the parameters take the worst possible value within their respective ranges. These least favourable parameter values are those that maximize the loss (King and Wong, 2000; Dette et al., 2003).
(4) Sequential designs: In sequential designs, the experiment is done in stages. Parameter estimates from a previous stage are used as initial estimates in the current stage. The process continues until optimal designs are obtained (Abdelbasit and Plackett, 1983; Sinha and Wiens, 2002).

Suppose an experimenter is faced with a set $\mathscr{S} = \{\mathbf{x}_i\}_{i=1}^{N}$ of possible design points from which he is interested in choosing n, not necessarily distinct, points at which to observe a binary response Y. The experimenter makes $n_i \geqslant 0$ observations at $\mathbf{x}_i$ such that $\sum_{i=1}^{N} n_i = n$. The design problem is to choose $n_1, \ldots, n_N$ in an optimal manner. The objective then is to choose a probability distribution $\{p_i\}_{i=1}^{N}$, with $p_i = n_i/n$, on the design space $\mathscr{S}$. The commonalities in the work of the authors who have considered logistic design is the salient assumption that the assumed model form is exactly correct. In this work, we propose the construction of robust designs for logistic models with due consideration for possible misspecification in the assumed form of the systematic component—the linear predictor. The linear predictor could be said to be misspecified when it does not reflect the influence of the covariates correctly, possibly due to omitted covariates or to omission of some transformation of existing covariates in the model. In this section we formalize our notion of model misspecification.

We suppose that the experimenter fits a logistic model with the mean response

$$\mu_i = \mu(\eta_i), \quad i = 1, \ldots, n, \tag{2}$$

for $\eta_i = \mathbf{z}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\beta}_0$ when in fact the true mean response is represented by

$$\mu_{T,i} = \mu(\eta_i + f(\mathbf{x}_i)). \tag{3}$$

The target parameter $\boldsymbol{\beta}_0$ is defined by

$$\boldsymbol{\beta}_0 = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \{E(Y|\mathbf{x}_i) - \mu(\mathbf{z}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\beta})\}^2.$$

Thus the target parameter is that which guarantees the least sum of squares of discrepancies, over all points in the design space, between the assumed mean response and the true mean response. The contamination function $f(\mathbf{x})$ may or may not be known. It would be known, for example, when an experimenter decides to fit the more parsimonious model (2) despite the knowledge of a more appropriate model (3) with a specified $f(\mathbf{x})$. For instance, the simplified model might be required if the number of support

points is not sufficient to handle a more complicated but more appropriate model. Knowing that the parsimonious model might result in an inferior analysis, the experimenter may seek a design that remedies the anticipated model inadequacy.

The contamination function would be unknown in a situation where the experimenter is aware of the possible uncertainties in the assumed model form and might have clues about the properties of the possible misspecification, without knowing its exact structure. When $f(\mathbf{x})$ is unknown, some knowledge about its properties or conditions it satisfies would be required to construct any appropriate design. This is so because no single design which takes a finite number of observations can protect against all possible forms of bias. Thus, we must impose some conditions on the contamination function when its precise form is unknown.

To bound the bias of an estimator $\hat{\beta}$, we assume that

$$\frac{1}{N}\sum_{i=1}^{N} f^2(\mathbf{x}_i) \leqslant \tau^2, \tag{4}$$

with $\tau^2 = \mathrm{O}(n^{-1})$. This latter requirement is analogous to the notion of contiguity in the asymptotic theory of hypothesis testing, and is justified in the same manner. The choice of τ is discussed following Theorem 3 in the next section. In order to ensure identifiability of the model parameters $\boldsymbol{\beta}$ and the contamination function $f(\mathbf{x})$ we require that the vector of regressors and the contamination be orthogonal. That is,

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}(\mathbf{x}_i) f(\mathbf{x}_i) = \mathbf{0}. \tag{5}$$

Let $\mathscr{F}$ denote the class of contamination functions $f(\mathbf{x})$ satisfying (4) and (5).

2. Loss functions: estimated and averaged mean squared errors of prediction

The basis for the construction of classical designs for logistic regression models has typically been the minimization of (a function of) the inverse of Fisher's information matrix (1)—see Atkinson and Haines (1996). However, in the face of model misspecification the asymptotic covariance, $cov(\hat{\boldsymbol{\beta}})$, of the maximum likelihood estimator of the model parameters no longer equals the inverse of Fisher information—see White (1982) and also Fahrmeir (1990), who discusses the asymptotic properties of MLEs under a misspecified likelihood.

Suppose that data $\{\mathbf{x}_i, y_i\}$ are given, where the $\mathbf{x}_i$ are the design points chosen from $\mathscr{S}$ with n_i observations at $\mathbf{x}_i$ such that $\sum_{i=1}^{N} n_i = n$, and y_i is the proportion of successes at location $\mathbf{x}_i$. The asymptotic bias and covariance of the MLE $\hat{\boldsymbol{\beta}}$ are given in Theorem 1; see the Appendix for details of this and other proofs. The expressions for the asymptotic bias and covariance of the MLE $\hat{\boldsymbol{\beta}}$ are used in the derivation of the loss function in Corollary 2.

Theorem 1. *Define*

$$w_i = \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} = \mu_i(1-\mu_i) = \frac{1}{4}\mathrm{sech}^2\left(\frac{\mathbf{z}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\beta}_0}{2}\right), \tag{6}$$

and let $\mathbf{Z}$ *be the* $N \times p$ *matrix with rows* $\mathbf{z}^{\mathrm{T}}(\mathbf{x}_i)$. *Recall* (2) *and* (3); *let* $\boldsymbol{\gamma}$ *and* $\boldsymbol{\gamma}_T$ *be the* $N \times 1$ *vectors with elements* μ_i *and* $\mu_{T,i}$, *respectively. Let* $\mathbf{P}$, $\mathbf{W}$ *and* $\mathbf{W}_T$ *be the* $N \times N$ *diagonal matrices with diagonal elements* n_i/n, w_i *and* $w_{T,i} = \mu_{T,i}(1-\mu_{T,i})$, *respectively. Finally, define* $\mathbf{b} = \mathbf{Z}^{\mathrm{T}}\mathbf{P}(\boldsymbol{\gamma}_T - \boldsymbol{\gamma}), \mathbf{H}_n = \mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{Z}, \tilde{\mathbf{H}}_n = \mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{W}_T\mathbf{Z}$. *The asymptotic bias and asymptotic covariance matrix of the maximum likelihood estimator* $\hat{\boldsymbol{\beta}}$ *of the model parameter vector* $\boldsymbol{\beta}$ *from the misspecified model are*

$$\mathrm{bias}(\hat{\boldsymbol{\beta}}) = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \mathbf{H}_n^{-1}\mathbf{b} + \mathrm{o}(n^{-1/2}),$$

$$cov(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)) = \mathbf{H}_n^{-1}\tilde{\mathbf{H}}_n\mathbf{H}_n^{-1} + \mathrm{o}(1),$$

respectively.

Since the typical focus of logistic designs is prediction, we take as loss function the normalized average mean squared error (AMSE) I of the response prediction $\mu(\hat{\eta}_i)$, with $\hat{\eta}_i = \mathbf{z}^{\mathrm{T}}(\mathbf{x}_i)\hat{\boldsymbol{\beta}}$. This is given by

$$I \triangleq \frac{n}{N}\sum_{i=1}^{N} E[\{\mu(\hat{\eta}_i) - \mu(\eta_i + f(\mathbf{x}_i))\}^2].$$

Corollary 2. *The AMSE has the asymptotic approximation* $I = \mathscr{L}_I(\mathbf{P}, \mathbf{f}) + \mathrm{o}(1)$, *where*

$$\mathscr{L}_I(\mathbf{P}, \mathbf{f}) = \frac{1}{N}\{\mathrm{tr}[\mathbf{W}\mathbf{Z}\mathbf{H}_n^{-1}\tilde{\mathbf{H}}_n\mathbf{H}_n^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{W}] + n\|\mathbf{W}(\mathbf{Z}\mathbf{H}_n^{-1}\mathbf{b} - \mathbf{f})\|^2\} \tag{7}$$

for $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))^{\mathrm{T}}$.

By using the expressions for asymptotic bias and covariance given in Theorem 1, Corollary 2 expresses the AMSE as an explicit function of the design matrix $\mathbf{Z}$ and contamination vector $\mathbf{f}$. The first term in the loss function $\mathscr{L}_I$ corresponds to the average variance of the predictions and it depends on the contamination function $f(\mathbf{x})$ through the matrix $\tilde{\mathbf{H}}_n$. The second term in the expression for $\mathscr{L}_I$ is the average squared bias of the predictions, which depends on the contamination $f(\mathbf{x})$ through the contamination vector $\mathbf{f}$ and implicitly through the vector $\mathbf{b}$. Thus a design cannot minimize (7) directly without certain assumptions about the contamination $f(\mathbf{x})$.

Fang and Wiens (2000) constructed integer-valued designs for linear models, in the case of an unknown f, using a minimax approach. Their minimax criterion minimizes the maximum value of the loss function over f. They solve the design problem by minimizing the loss when the misspecification is the worst possible in the neighbourhood of interest.

Here, we take one of the two approaches depending on whether or not there are initial data. If we have initial data we represent the discrepancy between the true response and the assumed response, at a sampled location $\mathbf{x}$, by

$$d(\mathbf{x}) = \mu(\mathbf{z}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}_0 + f(\mathbf{x})) - \mu(\mathbf{z}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}_0),$$

and estimate this by the residual $\hat{d}(\mathbf{x}) = y(\mathbf{x}) - \mu(\mathbf{z}^{\mathrm{T}}(\mathbf{x})\hat{\boldsymbol{\beta}})$. A first order approximation is $d(\mathbf{x}) \approx (\mathrm{d}\mu/\mathrm{d}\eta)f(\mathbf{x})$, leading to $\hat{f}(\mathbf{x}) = \hat{d}(\mathbf{x})/(\mathrm{d}\mu/\mathrm{d}\eta_{|\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}})$. We smooth this estimated contamination over the entire design space—see Example 3 of Section 5 for an illustration. The resulting estimate $\hat{\mathbf{f}}$, together with $\hat{\boldsymbol{\beta}}$, is then substituted into the terms in (7), and we compute a design minimizing $\mathscr{L}_I(\mathbf{P}, \hat{\mathbf{f}})$ using the techniques outlined in Section 3.

If there are no initial data we propose to instead average $\mathscr{L}_I$ over $\mathscr{F}$ defined by (4) and (5). Our optimal design minimizes this average value. This criterion is in the spirit of Läuter (1974, 1976). Läuter's criterion optimizes the weighted average of the loss of a finite set of plausible models. Here we are instead faced with an infinite set of models indexed by $f \in \mathscr{F}$.

To carry out the averaging we begin as in Fang and Wiens (2000), with the singular value decomposition

$$\mathbf{Z} = \mathbf{U}_{N\times p}\boldsymbol{\Lambda}_{p\times p}\mathbf{V}^{\mathrm{T}}_{p\times p}, \tag{8}$$

with $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_p$ and $\boldsymbol{\Lambda}$ diagonal and invertible. We augment $\mathbf{U}$ by $\tilde{\mathbf{U}}_{N\times(N-p)}$ such that $[\mathbf{U}\dot{:}\tilde{\mathbf{U}}]_{N\times N}$ is orthogonal. Then by (4) and (5), we have that there is an $(N-p)\times 1$ vector $\mathbf{t}$, with $\|\mathbf{t}\| \leqslant 1$, satisfying

$$\mathbf{f}(=\mathbf{f}_{\mathbf{t}}) = \tau\sqrt{N}\tilde{\mathbf{U}}\mathbf{t}. \tag{9}$$

The average loss is taken to be the expected value of (7), as a function of $\mathbf{t}$, with respect to the uniform measure on the unit sphere and its interior in $\mathbb{R}^{N-p}$. This measure has density $p(\mathbf{t}) = (1/\kappa_{N,p})I(\|\mathbf{t}\| \leqslant 1)$, where $\kappa_{N,p} = \pi^{(N-p)/2}/\Gamma((N-p)/2+1)$ is the volume of the unit sphere. Theorem 3 handles the averaging of $\mathscr{L}_I$. The importance of this theorem is in its elimination of the dependency of our design criterion on the unknown contamination function.

Theorem 3. *The average loss over the misspecification neighbourhood $\mathscr{F}$ is, apart from terms which are* o(1)*, given by*

$$\begin{aligned}\mathscr{L}_{I,\mathrm{ave}}(\mathbf{P},\rho) &\triangleq \int \mathscr{L}_I(\mathbf{P},\mathbf{f}_{\mathbf{t}})p(\mathbf{t})\,\mathrm{d}\mathbf{t}\\ &= \frac{1}{N}\mathrm{tr}[(\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{U})^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{W}^2\mathbf{U})] + \frac{\rho}{N-p+2}\mathrm{tr}[\mathbf{W}(\mathbf{R}-\mathbf{I}_N)(\mathbf{R}^{\mathrm{T}}-\mathbf{I}_N)\mathbf{W}],\end{aligned} \tag{10}$$

where $\rho = n\tau^2$ and $\mathbf{R}_{N\times N} = \mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}$. For numerical work it is more efficient to compute the second trace in (10) *as*

$$\mathrm{tr}[\mathbf{W}(\mathbf{R}-\mathbf{I}_N)(\mathbf{R}^{\mathrm{T}}-\mathbf{I}_N)\mathbf{W}] = \sum_{i=1}^{N} w_i^2\|\tilde{\mathbf{r}}_i\|^2,$$

where $\tilde{\mathbf{r}}_i^{\mathrm{T}}$ is the ith row of $\mathbf{R}-\mathbf{I}_N$.

The dependency of the design criterion on the unknown contamination has now been represented by a design parameter ρ, which can be chosen by the experimenter. This parameter can be interpreted as a measure of departure of the true model from the fitted model. In other words, it is a measure of the experimenter's lack of confidence in the validity of the model that he fits. If he believes that this assumed model is exactly correct, he chooses $\rho = 0$ corresponding to the classical I-optimal design. On the other hand, if the experimenter believes that the assumed model is highly uncertain, he chooses a large value of ρ for his design. Designs corresponding to a large value of ρ are dominated by the bias component of the loss.

Our design criterion (10) remains dependent on the model parameter vector $\boldsymbol{\beta}_0$, as is the case in the general nonlinear design problems, through the weights, as at (6). In the examples of the next section we handle this dependency by either taking a guess (locally optimal designs) or assuming a prior distribution, say $\pi(\boldsymbol{\beta}_0)$, on $\boldsymbol{\beta}_0$ (Bayesian designs). The loss function $\mathscr{L}_{I,\mathrm{ave}}$ is modified as $\int \mathscr{L}_{I,\mathrm{ave}}(\boldsymbol{\beta})\pi(\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta}$ in the case of a Bayesian construction.

3. Designs—algorithm and examples

3.1. Simulated annealing

We consider problems with polynomial predictors, viz. $\eta = \mathbf{z}^{\mathrm{T}}(x)\boldsymbol{\beta}$ with $\mathbf{z}(x) = (1, x, x^2, \ldots, x^{p-1})^{\mathrm{T}}$. We take equally spaced design points $\{x_i\}_{i=1}^{N}$ in the interval $\mathscr{S}$. Our design minimizes the relevant loss function through the matrix $\mathbf{P} = \mathrm{diag}(n_1/n, \ldots, n_N/n)$. This is a nonlinear integer optimization problem for which there is no analytic solution, and for which we employ simulated annealing to search for the optimal design.

The simulated annealing algorithm is a direct search random walk optimization algorithm which has been quite successful at finding global extrema of non-smooth functions and/or functions with many local extrema. The algorithm consists of three steps, each of which must be well adapted to the problem of interest for the algorithm to be successful. The first step is a specification of the initial state of the process. In this step an initial design has to be specified, say $\mathbf{P}_0$. The second is a specification of a scheme by which a new design $\mathbf{P}_1$ is chosen from the optimization space. The last step is a prescription of the basis of acceptance or rejection: an acceptance with probability 1 if $\mathscr{L}_{I,\mathrm{ave}}(\mathbf{P}_1) < \mathscr{L}_{I,\mathrm{ave}}(\mathbf{P}_0)$, otherwise acceptance with probability $\exp\{-(\mathscr{L}_{I,\mathrm{ave}}(\mathbf{P}_1) - \mathscr{L}_{I,\mathrm{ave}}(\mathbf{P}_0))/T\}$, where T is a tuning parameter. The tuning parameter is usually decreased as the iterations proceed. After a large number of iterations between the second and third steps the loss function is expected to converge to its (near) minimum value. Simulated annealing has been used for design problems by, among others, Meyer and Nachtsheim (1988), Fang and Wiens (2000) and Adewale and Wiens (2006).

A very simple and general approach that we considered for choosing the initial design is to randomly select p points from $\{x_i\}_{i=1}^{N}$ and randomly allocate the observations to these points such that the total number of observations is n. Fang and Wiens (2000) used a different approach which assumes that one of (n, N) is a multiple of the other. For any (n, N) combination they chose the initial design to be as uniform as possible. We applied this approach as well but found that the two approaches are equally efficient. For generating a new design we adopted the perturbation scheme of Fang and Wiens (2000). The turning parameter in the third step was chosen initially such that the acceptance rate is in the range 70% and 95%. We decrease T by a factor of .95 after each 20 iterations. In the examples below we run the algorithm several times with varying turning parameter specification and reduction rate in order to satisfy ourselves that the resulting design has the least loss possible under the relevant circumstances of each example. In Fig. 1 we present the simulated annealing trajectory for one of the cases presented in Example 1. It took 83 s for the algorithm to complete the preset maximum number of iterations (12 000, for this case) and the minimum loss was attained just before the 9000th iteration.

3.2. Examples

Example 1 (*No contamination*). As a benchmark we first consider the logistic regression model with a single predictor: $p = 2$, $\mathbf{z}(x) = (1, x)^{\mathrm{T}}$, $x \in \mathscr{S} = [-1, 1]$, $\boldsymbol{\beta} = (1, 3)^{\mathrm{T}}$, and no contamination: $\rho = 0$. We initially took $n = 20$, $N = 40$ and considered designs minimizing $\mathscr{L}_I$. The annealing algorithm converged to the design placing 10 of the 20 observations at each of the points $-.744$ and .128. This design is therefore the classical I-optimal design minimizing the integrated variance of the predictions over $\mathscr{S}$. There is evidently no previous theory that applies to this case. However, using a model that is a reparameterization of ours, and a continuous design space $[-1, 1]$, King and Wong (2000) showed the locally D-optimal design to be the design that is equally supported at $-.848$ and .181. For the sake of comparison, we sought an equivalent design using our finite design space and the algorithm described above. The resulting design places 10 of 20 observations at each of $-.846$ and .180. Thus, our algorithm attains the closest approximation to King and Wong's solution in that the points $-.846$ and .180 are nearest, in our design space, to $-.848$ and .181. Unlike designs for linear models, the optimal designs in this case do not necessarily place observations at the extreme points of the design space. This phenomenon is due to the curvature introduced by the link function and the resulting nonlinear relationship between the mean response and x.

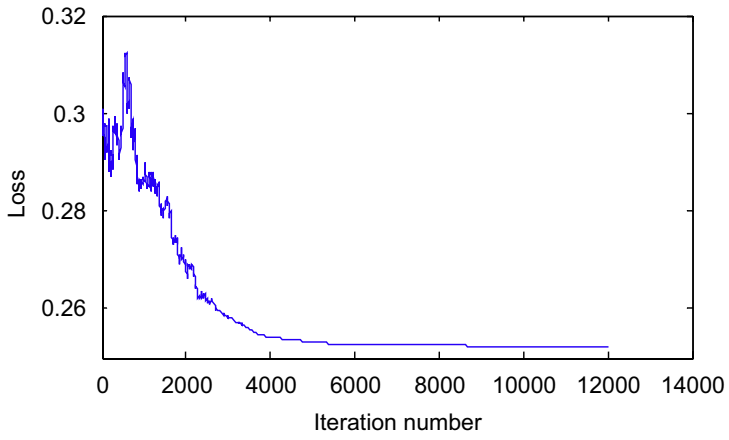

Fig. 1. Simulated annealing trajectory for logistic design with $\eta = 1 + 3x$, $x \in \mathscr{S} = [-1, 1]$, $\rho = 0$ and $(N = 40, n = 200)$.

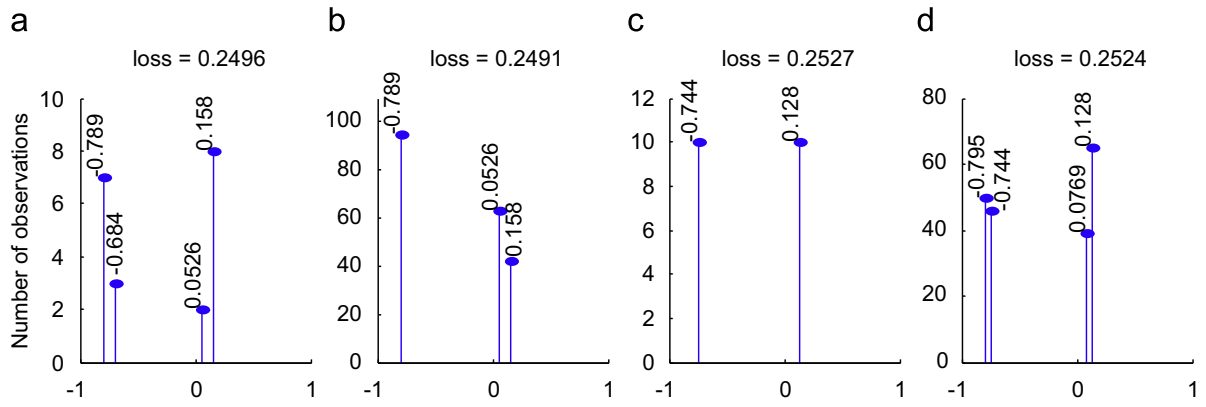

Fig. 2. Locally optimal designs minimizing $\mathscr{L}_{I,\text{ave}}$ when $\rho = 0$ (no contamination) with (a) $N = 20, n = 20$; (b) $N = 20, n = 200$; (c) $N = 40, n = 20$; (d) $N = 40, n = 200$.

Table 1
Comparing restricted designs with unrestricted design; $\rho = 0$

(N, n)	Restricted design (two-point)		Unrestricted design	
	Design points	Loss	Design points[a]	Loss
(20, 20)	−.789(9), .053(11)	.250	−.789(7), −.684(3), .0526(2), .158(8)	.250
(20, 200)	−.789(94), .053(106)	.250	−.789(95), .0526(63), .158(42)	.249
(40, 20)	−.744(10), .128(10)	.2527	−.744(10), .128(10)	.2527
(40, 200)	−.744(97), .128(103)	.2525	−.795(49), −.744(47), .0769(39), .128(65)	.2524

[a]Number of observations in parentheses.

Our numerical results further revealed that the designs depend on the number of points in the design space and the number of observations the experimenter is willing to take. For this "no-contamination" case, we investigated designs for various combinations of N and n. Some of these designs are presented in Fig. 2. The number of distinct design points varies from 2 to 4. We found this somewhat surprising, in light of the fact that all D-optimum designs for the two parameter logistic model in the literature are two-point designs. Presumably this is explained through our use of a finite design space, and/or our use of average loss rather than that based on the determinant of the information matrix.

To check that this phenomenon was not merely an artefact due to a lack of convergence, we modified our algorithm to obtain "restricted" designs—restricted to two support points only. The results for the same values of N and n as in Fig. 2 are presented in Table 1. The loss for the unrestricted design is less than or equal to that for the corresponding restricted design in all cases considered.

In the examples that follow we limit discussion to the case $N = 40$, $n = 200$.

Example 2 (*Example 1 continued*). In this example, which we include largely for illustrative purposes, the form of the contamination is known. Suppose that the experimenter anticipates fitting a simple logistic model, while wishing protection against a range of logistic models with quadratic predictor: $\eta(x) = \mathbf{z}^{\mathrm{T}}(x)\boldsymbol{\beta} + f(x)$, where $\mathbf{z}^{\mathrm{T}}(x)$ and $\boldsymbol{\beta}$ are as in Example 1, and $f(x) = \beta_2(x^2 - \mu_2)/\sqrt{\mu_4 - \mu_2^2}$, for $\mu_k = N^{-1}\sum x_i^k$ (=0 if k is odd). The contaminant $f(x)$ is an omitted quadratic term, translated and scaled to ensure the orthogonality condition (5); (4) becomes $|\beta_2| \leqslant \tau$. We obtained optimal designs for various values of the quadratic coefficient β_2. The resulting designs and the corresponding values of the loss function are presented in Table 2. In the range of values of β_2 considered, we found that the number of distinct points varied from 3 to 6. The spread of the design over the design space tended to increase as the magnitude of the omitted quadratic term increases. We computed the premium paid for robustness and the gain due to robustness for each design presented as

$$\text{Premium} = \left(\frac{\mathscr{L}_I(\mathbf{P}_{\text{opt}}, \mathbf{f} = \mathbf{0})}{\mathscr{L}_I(\mathbf{P}_{\text{classical}}, \mathbf{f} = \mathbf{0})} - 1 \right) \times 100\% \tag{11}$$

and

$$\text{Gain} = \left(1 - \frac{\mathscr{L}_I(\mathbf{P}_{\text{opt}}, \mathbf{f})}{\mathscr{L}_I(\mathbf{P}_{\text{classical}}, \mathbf{f})} \right) \times 100\%. \tag{12}$$

Table 2
Designs for simple logistic model when the true model has a quadratic term

β_2	Design points (number of observations)	$\mathscr{L}_I(\mathbf{P},\mathbf{f})$	Premium	Gain
−10	−1(42), −.180(42), −.128(96), −.077(12), −.026(2), .077(6)	3.500	35.0%	34.8%
−3	−1(48), −.282(26), −.231(64), .282(62)	.5020	10.9%	17.0%
−1	−.949(42), −.590(34), −.539(29), .128(95)	.2756	2.1%	.5%
0	−.795(49), −.744(47), .077(39), .128(65)	.2524	0	0
1	−.641(100), .128(22), .180(78)	.3080	1.9%	4.0%
3	−.692(57), −.641(29), −.077(39), −.026(44), .795(31)	.6073	11.9%	19.9%
10	−1(11), −.590(51), −.539(27), −.231(30), −.180(40), .949(41)	3.679	39.0%	42.9%

Table 3
Experimental design and response values

Design point	−1	$-\frac{7}{9}$	$-\frac{5}{9}$	$-\frac{3}{9}$	$-\frac{1}{9}$	$\frac{1}{9}$	$\frac{3}{9}$	$\frac{5}{9}$	$\frac{7}{9}$	1
No. of observations	20	20	20	20	20	20	20	20	20	20
No. of successes	8	6	7	13	17	18	18	19	20	20

The gain measure is the percentage reduction in loss due to the use of a robust design as opposed to a (non-robust) classical design which assumes the fitted model to be exactly correct. The premium measure is the percentage increase in loss as a result of not using the classical design if in actual fact the assumed model is correct. The application of the premium and/or gain measure depends on the amount of confidence the experimenter has in his knowledge of the true model. In this example, since the assumption is that the experimenter knows that the model with a linear predictor involving the quadratic term is a more appropriate model, the relevant measure would be the gain. Nevertheless, both measures are reported in Table 2. The value of a design from our robust procedure increases with increasing magnitude of the quadratic parameter. On the other hand, the experimenter has to be aware of the increasing premium when his knowledge of the true model is not accurate. The premium paid for robustness also increases with the magnitude of the quadratic parameter.

Of course this example is artificial, assuming as it does that the true form of the predictor is known to be quadratic, with parameter $\beta_2 = 3$, say. If one did indeed possess this knowledge then the classically optimal—i.e., variance minimizing—design would be $-1(49), -.282(2), -.231(91), .436(9), .487(49)$. The premium figures in Table 2 would rise appreciably—to typical values of 100% or more—since the robust design would be protecting against bias, known not to be present.

Example 3 (*Designing when there are initial data to estimate contamination*). Table 3 shows simulated data ("# of successes") from a logistic regression model with the predictor $\eta(x) = 1 + 3x + f(x)$, the model of the previous example; the quadratic parameter was $\beta_2 = 3$. The data were simulated using a uniform design over equally spaced points in $[-1, 1]$. Having simulated the data, we suppose the contamination function $f(x)$ to be unknown. We proceed using the procedure described in Section 2 for estimation and eventual smoothing of the contamination. A plot of the estimated contamination with its loess smooth $\hat{f}(x)$ over the design space is presented in Fig. 3.

We plugged the smoothed contamination values into the loss function (7), and used simulated annealing to obtain the design. Our design places 34, 82, and 84 of the 200 observations at −.641, −.590, and .180, respectively. For this design the premium for robustness is 5.0% and the gain is 60.0%. This example indicates that when there are initial data, it is expedient to incorporate the information from the data into the design procedure. The resulting design can lead to substantial gain at a reduced premium.

Example 4 (*Unknown contamination*). Consider the logistic model with predictor

$$\eta(x) = \beta_0 + \beta_1 x + f(x). \tag{13}$$

In this example—as in Example 3—we assume that f is an unknown member of the class $\mathscr{F}$ defined by (4) and (5). In Fig. 4 we exhibit designs minimizing the averaged loss (10) for various values of ρ, β_0 and β_1. We observed a progression of the dispersion of the design points over the design space with increasing ρ. The pattern of the dispersion is, however, modified by the curvature indexed by β_0 and β_1 through the nonlinear mean response. For small ρ our robust designs can be described as taking clusters of observations at neighbouring locations rather than replications at only a few distinct sites; this was noticed for linear models by Fang and Wiens (2000). However, here there is always a pattern to the clusters of observation to be taken depending on the values of the model parameters. Large values of ρ denote large departures from the assumed model and an extremely large ρ value corresponds to the all-bias design. Even though the all-bias design is spread over the entire design space the frequencies of observations are different and these frequencies are prescribed by the curvature of the mean response as determined by the

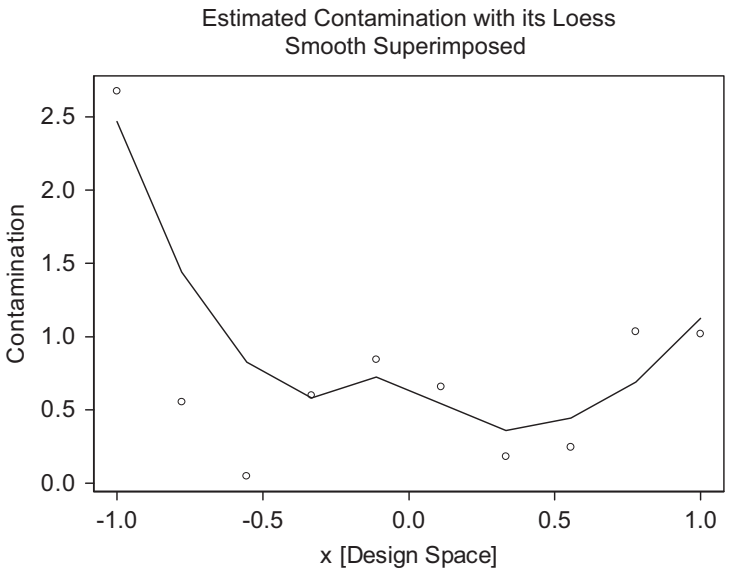

Fig. 3. Estimated contamination plot for Example 3. True (but unknown) form of contamination is quadratic.

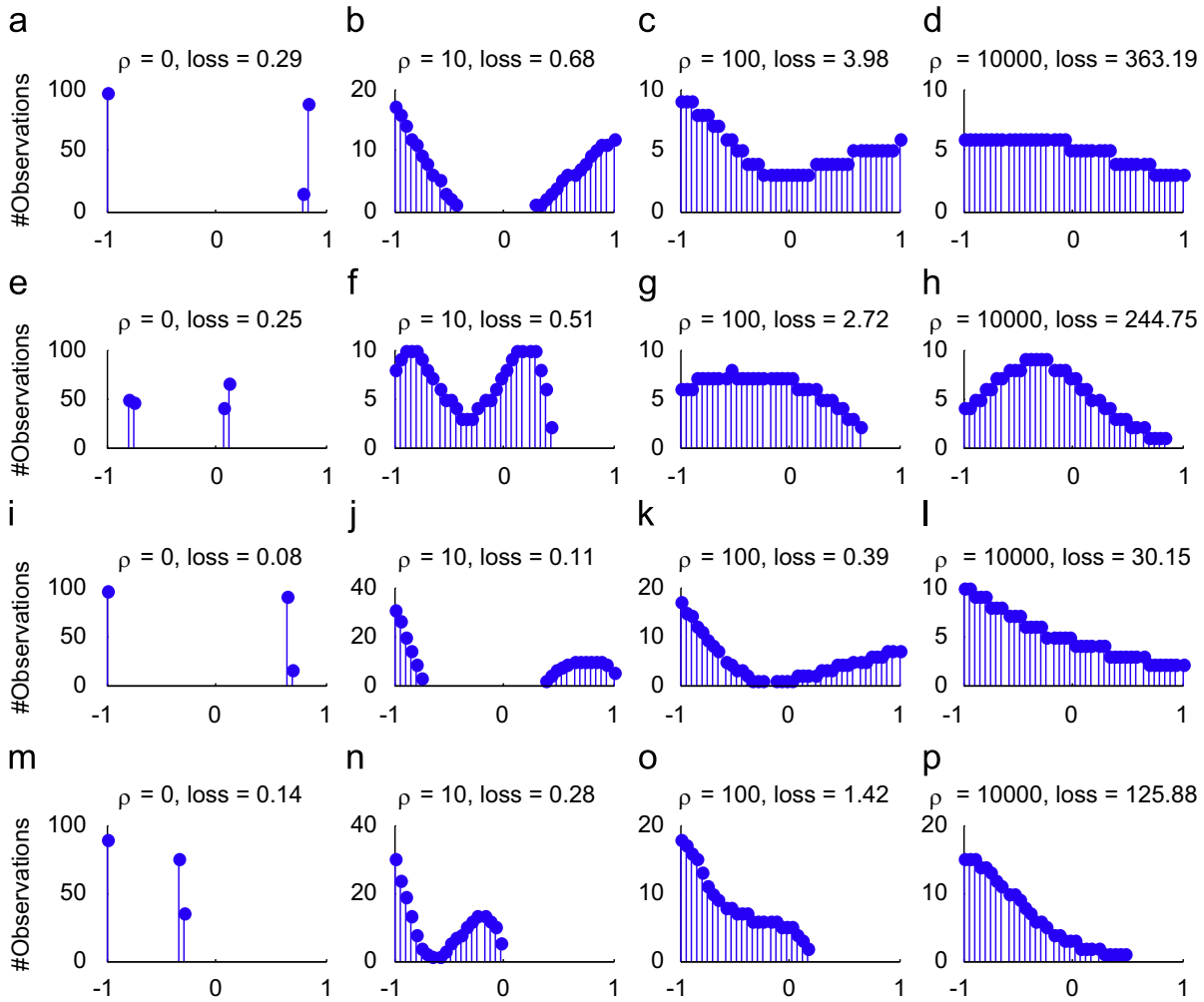

Fig. 4. Locally optimal designs in Example 4: (a)–(d) $(\beta_0, \beta_1) = (1, 1)$; (e)–(h) $(\beta_0, \beta_1) = (1, 3)$; (i)–(l) $(\beta_0, \beta_1) = (3, 1)$; (m)–(p) $(\beta_0, \beta_1) = (3, 3)$.

Table 4
Design for unknown contamination with $\beta_0 = 1$ and $\beta_1 = 3$

ρ	$\mathscr{L}_{I,\text{ave}}(\mathbf{P}, \rho)$	Premium	Gain
0	.2524	0	0
1	.2809	.62%	2.01%
10	.5090	3.06%	14.33%
100	2.7204	7.44%	25.87%
1000	24.7294	11.29%	28.16%
10 000	244.7545	11.97%	28.43%

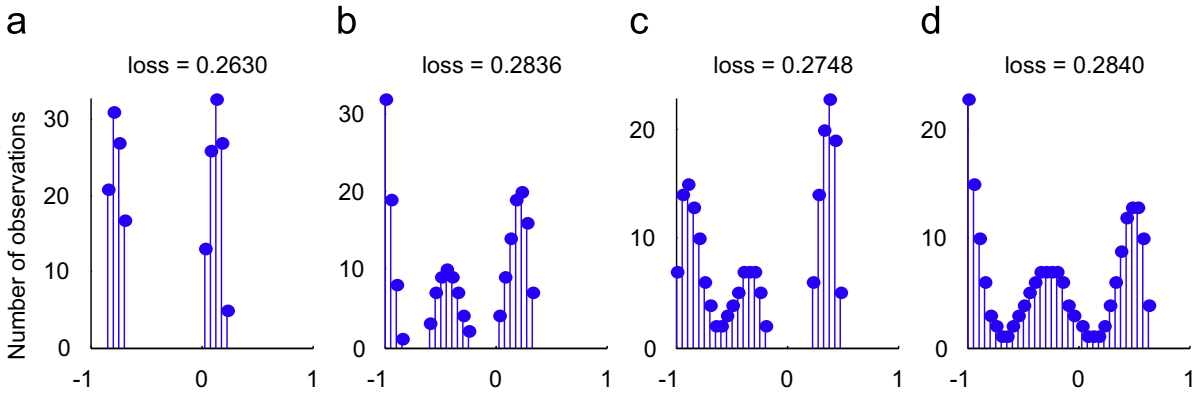

Fig. 5. Robust Bayesian optimal design in Example 5 with $\rho = .25$ and parameters β_0 and β_1 having independent uniform priors over (a) $[.5, 1.5] \times [2.5, 3.5]$, (b) $[.5, 1.5] \times [1, 5]$, (c) $[-1, 3] \times [2.5, 3.5]$, (d) $[-1, 3] \times [1, 5]$.

model parameters. In Table 4 we present the values of the premium paid and the gain in robustness for designs corresponding to different values of ρ for the particular case of $(\beta_0, \beta_1) = (1, 3)$. The gain in robustness, measured by (12), exceeds the premium paid, as measured by (11), for each design. Increasing robustness, however, comes with increasing premium; the experimenter would thus have to choose his level of comfort.

Thus far, the examples we have presented have been locally optimal, hence have assumed good parameter guesses for unknown model parameters. In the absence of a reliable best guess for model parameters, Sitter (1992) and King and Wong (2000) considered minimax *D*-optimality, a procedure which assumes the knowledge of a prior range for each of the parameters. We consider a Bayesian paradigm to be in the same spirit as averaging the contamination function over the specified misspecification neighbourhood, and take independent uniform prior distributions over the range of each model parameter. Our design criteria then becomes the expected loss, $E(\mathscr{L}_{I,\text{ave}}(\mathbf{P}, \rho))$, with the expectation taken with respect to these priors. The dependency of our design criteria on the model parameters is through the weights w_i, and we do not have analytic expressions for the resulting integrals. In the examples that follow we employ number-theoretic methods for numerical evaluation of multiple integrals as discussed in Fang and Wang (1994). This approach is based on generating quasi-random points in the domain of definition of the integrand, and averaging the values of the loss over the sample of points.

Example 5 (*Robust Bayesian design*). In this example we consider the following ranges of parameter values: (a) $[.5, 1.5] \times [2.5, 3.5]$, (b) $[.5, 1.5] \times [1, 5]$, (c) $[-1, 3] \times [2.5, 3.5]$, (d) $[-1, 3] \times [1, 5]$, all with centre point $(1, 3)$ but with coverage areas 1, 4, 4, and 16, respectively. As described above, the robust design for each range of parameter values is the design that minimizes the expected average loss with respect to uniform distributions on the specified ranges of parameter values. For each of the designs—see Fig. 5—we take $\rho = .25$. We observed an increasing spread over the design space with increasing uncertainty in model parameters, as measured by the coverage area of the priors. This is consistent with previous work in optimal Bayesian design—see, for example, Chaloner and Larntz (1989)—which suggests increasing number of distinct design points with increasing uncertainty in the specified prior distributions. Comparing the design plots in panels (b) and (c) of Fig. 5, we see that there is more sensitivity to uncertainty in the intercept parameter than the slope parameter.

4. Case study: beetle mortality data

Bliss (1935) reported the numbers of beetles dead after 5 h exposure to gaseous carbon disulphide at various concentrations. The doses are given in Table 5; to facilitate our discussion we have linearly transformed these to the range $[0, 1]$. Note that the original design is then very nearly uniform on the eight equally spaced points $0(\frac{1}{7})1$.

Table 5
Beetle mortality data

Dose, x_i ($\log_{10}$ CS_2 mg l^{-1})	1.69	1.72	1.75	1.78	1.81	1.84	1.86	1.88
Number of beetles, n_i	59	60	62	56	63	59	62	60
Number killed, $n_i y_i$	6	13	18	28	52	53	61	60

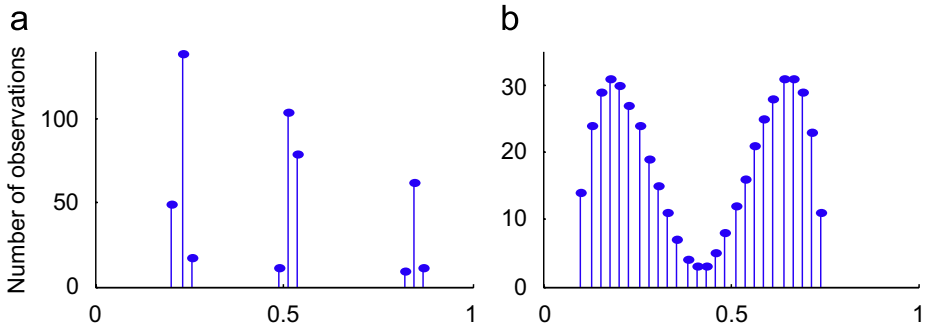

Fig. 6. (a) Prediction design when contamination is estimated from initial data. (b) Robust Bayesian prediction design with multivariate normal prior and $\rho = 5$.

We first fitted the logistic model with the linear predictor $\eta^{(1)} = \beta_0^{(1)} + \beta_1^{(1)}x$, and obtained the estimates $\hat{\beta}_0^{(1)} = -2.777$ and $\hat{\beta}_1^{(1)} = 6.621$ with the estimated variance–covariance matrix

$$\Sigma^{(1)} = \begin{pmatrix} .082 & -.144 \\ -.144 & .317 \end{pmatrix}$$

and deviance =11.232 (df = 6). The corresponding estimates for the logistic model with the linear predictor $\eta^{(2)} = \beta_0^{(2)} + \beta_1^{(2)}x + \beta_2^{(2)}x^2$ are $\hat{\beta}_0^{(2)} = -2.00$, $\hat{\beta}_1^{(2)} = 1.60$, $\hat{\beta}_2^{(2)} = 5.84$ and

$$\Sigma^{(2)} = \begin{pmatrix} .124 & -.522 & .489 \\ -.522 & 3.252 & -3.690 \\ .489 & -3.690 & 4.665 \end{pmatrix}$$

with deviance =3.195 (df = 5). The deviances and a plot (not presented here) of proportions of beetles killed against dose levels with the estimated proportions from each model superimposed suggest that the model with the quadratic term is a significantly better fit for these data. Suppose the experimenter is inclined to use the simple logistic fit for future data for ease of interpretation and model simplicity or that the adequacy of the model with the quadratic term is itself in doubt. We proceed by estimating the contamination and then smoothing over the design space as discussed in Section 2. The resulting design, obtained using the parameter estimates $\hat{\beta}_0^{(1)}$ and $\hat{\beta}_1^{(1)}$ as initial guesses, with total number of observations $n = 481$ over an equally spaced grid of $N = 40$ points in $[0, 1]$ is presented in panel (a) of Fig. 6. This would be the design of choice if the experimenter were interested in prediction but contemplated the superiority of the model with quadratic term. However, the experimenter can ensure robustness against a broader set of alternatives by taking the contamination to belong to the class $\mathcal{F}$ while assuming an initial multivariate normal prior on the parameter, with mean vector $(\hat{\beta}_0^{(1)}, \hat{\beta}_1^{(1)})^{\mathrm{T}}$ and variance–covariance matrix $\Sigma^{(1)}$, and then using the Bayesian paradigm as in Example 5. The loss function becomes the expected value of (10), with expectation taken with respect to the multivariate normal prior. The numerical implementation of expectation is done using a quasi-Monte Carlo sampling approach. The design plot is given in Fig. 6(b).

5. Conclusions

We have investigated integer-valued designs for logistic regression models, using polynomial predictors as specific examples. Our designs are robust against misspecification in the predictor. We have addressed both known and unknown contamination. Previous robustness work done for logistic models has concentrated on the uncertainty of model parameters; in this contribution we have gone further to investigate specific violations in the form of the assumed linear (in the parameters) predictor.

Designs for a specific alternative, for example quadratic versus linear in the independent variable, are quite different from those for broad classes of alternatives. The number of distinct design points is usually not as large in the former case as in the latter. In fact, when the magnitude of the misspecification is minimal the resulting robust design could have about the same number of distinct observation points as its classical counterpart. Nevertheless, the gain in robustness often exceeds the premium paid for robustness—see Table 1.

Designs for a very specific alternative may, however, suffer the same fate as designs assuming the correctness of the fitted model when the alternative itself is not valid. Both take replicates of observations at only a few distinct points, especially when the magnitude of the departure is small. However, when there is a higher degree of certainty in the alternative, these designs could result in substantial gain in robustness. An example of this would be when the experimenter is aware of a more appropriate model but seeks a design that allows for the fitting of a more parsimonious model. Also, designs when there are data to estimate model contamination are quite similar to designs when the exact form of the contamination is known (single alternative). When the information in the initial data is incorporated into the design procedure, as seen in Example 3 above, the robustness of the resulting design could come at a very reduced premium.

In general, we have found there to be increasing numbers of distinct observation sites with increasing model uncertainty. The overall message is consistent with that reported in the model robust design literature for linear models—robust designs can be approximated by placing clusters of observations about the support points for classical designs. However, the nonlinearity of the mean response in logistic design adds a slight twist to the overall message, in that the clusters of observation come with patterns that are determined by the curvature prescribed by the model parameters. More striking is the fact that the all-bias design is non-uniform in logistic regression models—even though the recommended design points are spread over the entire design space, the frequencies of observations vary due to the curvature.

Overall, the design that protects against uncertainty in model parameters (via a Bayesian paradigm) and that which protects against uncertainty in assumed model form could be described as taking observations in clusters. These clusters often come in interesting patterns of curvature prescribed by the nonlinearity of the model—see examples in the previous section. Further work would be required to obtain analytic descriptions of the effect of curvature in this robust approach, or even for the all-bias designs for logistic models. While the focus of the model misspecification reported here is exclusively on linear predictor misspecification, we are currently investigating other forms of misspecification in designing for the broader class of generalized linear models, of which the logistic model is but a special case.

Acknowledgements

The research of both authors is supported by the Natural Sciences and Engineering Research Council of Canada. We appreciate helpful comments from an anonymous referee.

Appendix A. Derivations

Proof of Theorem 1. Under conditions as in Fahrmeir (1990) the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ exists and is consistent, and $\partial l(\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}$ is $o_p(n^{-1/2})$. The log-likelihood l, the score function and -1 times the second derivative according to the assumed model are

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N}\left\{n_i\left[y_i\log\left(\frac{\mu_i}{1-\mu_i}\right)+\log(1-\mu_i)\right]+\log\binom{n_i}{n_i y_i}\right\},$$

$$\frac{\partial l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{N} n_i(y_i-\mu_i)\mathbf{z}(\mathbf{x}_i), \quad -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} = \sum_{i=1}^{N} n_i w_i \mathbf{z}(\mathbf{x}_i)\mathbf{z}^{\mathrm{T}}(\mathbf{x}_i).$$

An expansion of $\partial l(\boldsymbol{\beta})/\partial\beta_j$ around $\boldsymbol{\beta}_0$ gives

$$\frac{\partial l(\boldsymbol{\beta})}{\partial\beta_j} = \frac{\partial l(\boldsymbol{\beta}_0)}{\partial\beta_j} + \sum_k(\beta_k-\beta_{0,k})\frac{\partial^2 l(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_k} + \frac{1}{2}\sum_k\sum_l(\beta_k-\beta_{0,k})(\beta_l-\beta_{0,l})\frac{\partial^3 l(\boldsymbol{\beta}_*)}{\partial\beta_j\partial\beta_k\partial\beta_l},$$

where β_j and $\beta_{0,j}$ are the jth terms of the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$, respectively, and $\boldsymbol{\beta}_*$ is a point on the line segment connecting $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. If we replace $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ in this expansion, we obtain

$$\sqrt{n}\sum_k(\hat{\beta}_k-\beta_{0,k})\left[\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_k}+\frac{1}{2n}\sum_l(\hat{\beta}_l-\beta_{0,l})\frac{\partial^3 l(\boldsymbol{\beta}_*)}{\partial\beta_j\partial\beta_k\partial\beta_l}\right] = -\frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\beta}_0)}{\partial\beta_j}.$$

For the logistic likelihood the $\partial^3 l(\boldsymbol{\beta}_*)/\partial\beta_j\partial\beta_k\partial\beta_l$ are bounded, and so, using the consistency of $\hat{\boldsymbol{\beta}}$, we have that

$$\left[\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_k}+\frac{1}{2n}\sum(\hat{\beta}_l-\beta_{0,l})\frac{\partial^3 l(\boldsymbol{\beta}_*)}{\partial\beta_j\partial\beta_k\partial\beta_l}\right]\xrightarrow{\mathrm{p}} -H_{jk},$$

where H_{jk} is the (j,k)th element of the matrix $\mathbf{H}_n = -(1/n)\partial^2 l(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}} = \mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{Z}$. Thus the limit distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ is that of the solution of the equations $\sum H_{jk}\sqrt{n}(\hat{\beta}_k-\beta_{0,k}) = (1/\sqrt{n})\partial l(\boldsymbol{\beta}_0)/\partial\beta_j$, i.e., is the limit distribution of $\mathbf{H}_n^{-1}(1/\sqrt{n})\partial l(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}$.

Using the central limit theorem for independent not identically distributed random variables we have that $(1/\sqrt{n})\partial l(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}$ has a multivariate normal limit distribution with asymptotic mean $(1/\sqrt{n})\sum_{i=1}^{N} n_i E[y_i-\mu_i(\boldsymbol{\beta}_0)]\mathbf{z}(\mathbf{x}_i)=\sqrt{n}\mathbf{b}$ and asymptotic covariance matrix $\tilde{\mathbf{H}}_n=\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{W}_T\mathbf{Z}$. From this it follows that $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ is AN$(\sqrt{n}\mathbf{H}_n^{-1}\mathbf{b}, \mathbf{H}_n^{-1}\tilde{\mathbf{H}}_n\mathbf{H}_n^{-1})$, as required. □

Proof of Corollary 2. First write

$$I=\frac{1}{N}\sum_{i=1}^{N} var[\sqrt{n}\mu(\hat{\eta}_i)]+\frac{1}{N}\sum_{i=1}^{N}\{E[\sqrt{n}\mu(\hat{\eta}_i)]-\sqrt{n}\mu(\eta_i+f(\mathbf{x}_i))\}^2.$$

By the δ-method, the first sum is, up to terms which are o(1),

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i}\right)^2 var[\sqrt{n}\hat{\eta}_i] &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathrm{d}\mu}{\mathrm{d}\eta_i}\right)^2 \mathbf{z}^{\mathrm{T}}(\mathbf{x}_i)\mathbf{H}_n^{-1}\tilde{\mathbf{H}}_n\mathbf{H}_n^{-1}\mathbf{z}(\mathbf{x}_i)\\ &= \frac{1}{N}\mathrm{tr}[\mathbf{W}\mathbf{Z}\mathbf{H}_n^{-1}\tilde{\mathbf{H}}_n\mathbf{H}_n^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{W}].\end{aligned}$$

Also, on expanding $\mu(\hat{\eta}_i)$ and $\mu(\eta_i+f(\mathbf{x}_i))$ around η_i, we have

$$E[\sqrt{n}\mu(\hat{\eta}_i)]=\sqrt{n}\mu(\eta_i)+E\left[\sqrt{n}\frac{\mathrm{d}\mu}{\mathrm{d}\eta_i}(\hat{\eta}_i-\eta_i)+\mathrm{o}(\sqrt{n}(\hat{\eta}_i-\eta_i))\right],$$

and

$$\sqrt{n}\mu(\eta_i+f(\mathbf{x}_i))=\sqrt{n}\mu(\eta_i)+\sqrt{n}\frac{\mathrm{d}\mu}{\mathrm{d}\eta_i}f(\mathbf{x}_i)+\mathrm{o}(\sqrt{n}f(\mathbf{x}_i)).$$

Using an argument similar to that in the proof of Theorem 1, we have

$$E[\sqrt{n}\mu(\hat{\eta}_i)]=\sqrt{n}\mu(\eta_i)+\sqrt{n}\frac{\mathrm{d}\mu}{\mathrm{d}\eta_i}E(\hat{\eta}_i-\eta_i)+\mathrm{o}(1).$$

Thus, the second sum in the expression of I is, up to terms which are o(1),

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\{E[\sqrt{n}\mu(\hat{\eta}_i)]-\sqrt{n}\mu(\eta_i+f(\mathbf{x}_i))\}^2\\ &\quad=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathrm{d}\mu}{\mathrm{d}\eta_i}\right)^2\{n\cdot\mathrm{bias}^{\mathrm{T}}(\hat{\boldsymbol{\beta}})\mathbf{z}(\mathbf{x}_i)\mathbf{z}^{\mathrm{T}}(\mathbf{x}_i)\,\mathrm{bias}(\hat{\boldsymbol{\beta}})+nf^2(\mathbf{x}_i)\}\\ &\quad=\frac{1}{N}\{n\cdot\mathbf{b}^{\mathrm{T}}\mathbf{H}_n^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{W}^2\mathbf{Z}\mathbf{H}_n^{-1}\mathbf{b}-2n\mathbf{f}^{\mathrm{T}}\mathbf{W}^2\mathbf{Z}\mathbf{H}_n^{-1}\mathbf{b}+n\cdot\mathbf{f}^{\mathrm{T}}\mathbf{W}^2\mathbf{f}\},\end{aligned}$$

reducing to $(n/N)\|\mathbf{W}(\mathbf{Z}\mathbf{H}_n^{-1}\mathbf{b}-\mathbf{f})\|^2$. □

Proof of Theorem 3. Here and elsewhere, in the averaging we will use the identity $\int \mathbf{t}^{\mathrm{T}}\mathbf{t}p(\mathbf{t})\,\mathrm{d}\mathbf{t}=(N-p)/(N-p+2)$, which implies that

$$\int \mathbf{t}\mathbf{t}^{\mathrm{T}}p(\mathbf{t})\,\mathrm{d}\mathbf{t}=\frac{1}{N-p+2}\mathbf{I}_{N-p}.$$

First use (8) and (9) to write (7) explicitly in terms of $\mathbf{t}$:

$$\mathscr{L}_I(\mathbf{P},\mathbf{f})=\frac{1}{N}\left\{\begin{array}{l}\mathrm{tr}[(\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{U})^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}_T(\mathbf{t})\mathbf{U})(\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{W}^2\mathbf{U}]\\ \quad+n\|\mathbf{W}(\mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{P}(\boldsymbol{\gamma}_T(\mathbf{t})-\boldsymbol{\gamma})-\tau\sqrt{N}\tilde{\mathbf{U}}\mathbf{t})\|^2\end{array}\right\}. \tag{A.1}$$

Using (9) again we have $\mathbf{W}_T(\mathbf{t})=\mathbf{W}+\dot{\mathbf{W}}\tau\sqrt{N}\tilde{\mathbf{U}}\mathbf{t}+\mathrm{O}(\tau^2)$, where $\dot{\mathbf{W}}=\mathrm{diag}(w'(\eta_1),\ldots,w'(\eta_N))$. Since $\tau^2=\mathrm{O}(n^{-1})$ we obtain $\int \mathbf{W}_T(\mathbf{t})p(\mathbf{t})\,\mathrm{d}\mathbf{t}=\mathbf{W}+\mathrm{O}(n^{-1})$, and so

$$\int \mathrm{tr}[(\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{U})^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}_T(\mathbf{t})\mathbf{U})(\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{W}^2\mathbf{U}]p(\mathbf{t})\,\mathrm{d}\mathbf{t}=\mathrm{tr}[(\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{U})^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{W}^2\mathbf{U})].$$

Similarly, we have $\boldsymbol{\gamma}_T(\mathbf{t})-\boldsymbol{\gamma}=\tau\sqrt{N}\mathbf{W}\tilde{\mathbf{U}}\mathbf{t}+\mathrm{O}(\tau^2)$, and so

$$n\|\mathbf{W}(\mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{P}(\boldsymbol{\gamma}_T(\mathbf{t})-\boldsymbol{\gamma})-\tau\sqrt{N}\tilde{\mathbf{U}}\mathbf{t})\|^2=n\tau^2N\|\mathbf{W}(\mathbf{R}-\mathbf{I})\tilde{\mathbf{U}}\mathbf{t}\|^2+\mathrm{O}(n^{-1/2}),$$

with

$$\int n\|\mathbf{W}(\mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{P}(\boldsymbol{\gamma}_T(\mathbf{t})-\boldsymbol{\gamma})-\tau\sqrt{N}\tilde{\mathbf{U}}\mathbf{t})\|^2 p(\mathbf{t})\,\mathrm{d}\mathbf{t} = \frac{n\tau^2 N\cdot \mathrm{tr}[\mathbf{W}(\mathbf{R}-\mathbf{I})\tilde{\mathbf{U}}\tilde{\mathbf{U}}^{\mathrm{T}}(\mathbf{R}-\mathbf{I})^{\mathrm{T}}\mathbf{W}]}{N-p+2}.$$

The result follows upon noting that $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^{\mathrm{T}} = \mathbf{I} - \mathbf{U}\mathbf{U}^{\mathrm{T}}$ and $(\mathbf{R}-\mathbf{I})\mathbf{U} = \mathbf{0}$, and then substituting these integrals into (A.1) and simplifying. □

References

Abdelbasit, K.M., Plackett, R.L., 1983. Experimental design for binary data. J. Amer. Statist. Assoc. 8, 90–98.
Adewale, A., Wiens, D.P., 2006. New criteria for robust integer-valued designs in linear models. Comput. Statist. Data Anal. 51, 723–736.
Atkinson, A.C., Haines, L.M., 1996. Designs for nonlinear and generalized linear models. In: Ghosh, S., Rao, C.R. (Eds.), Handbook of Statistics, vol. 13. pp. 437–475.
Bliss, C.I., 1935. The calculation of the dose-mortality curve. Ann. Appl. Biol. 22, 134–167.
Box, G.E.P., Draper, N.R., 1959. A basis for the selection of a response surface design. J. Amer. Statist. Assoc. 54, 622–654.
Burridge, J., Sebastiani, P., 1994. *D*-optimal designs for generalized linear models with variance proportional to the square of the mean. Biometrika 81, 295–304.
Chaloner, K., Larntz, K., 1989. Optimal Bayesian design applied to logistic regression experiments. J. Statist. Plann. Inference 21, 191–208.
Chaudhuri, P., Mykland, P., 1993. Nonlinear experiments: optimal design and inference based on likelihood. J. Amer. Statist. Assoc. 88, 538–546.
Chernoff, H., 1953. Locally optimal designs for estimating parameters. Ann. Math. Statist. 24, 586–602.
Dette, H., Wong, W.K., 1996. Optimal Bayesian designs for models with partially specified heteroscedastic structure. Ann. Statist. 24, 2108–2127.
Dette, H., Haines, L., Imhof, L., 2003. Bayesian and maximin optimal designs for heteroscedastic regression models. Canad. J. Statist. 33, 221–241.
Fahrmeir, L., 1990. Maximum likelihood estimation in misspecified generalized linear models. Statistics 21, 487–502.
Fang, K.-T., Wang, Y., 1994. Number-Theoretic Methods in Statistics. Chapman & Hall, London.
Fang, Z., Wiens, D.P., 2000. Integer-valued, minimax robust designs for estimation and extrapolation in heteroscedastic, approximately linear models. J. Amer. Statist. Assoc. 95, 807–818.
Fedorov, V.V., 1972. Theory of Optimal Experiments. Academic Press, New York.
Ford, I., Silvey, S.D., 1980. A sequentially constructed design for estimating a nonlinear parametric function. Biometrika 67, 381–388.
Ford, I., Titterington, D.M., Kitsos, C.P., 1989. Recent advances in nonlinear experimental design. Technometrics 31, 49–60.
Ford, I., Torsney, B., Wu, C.F.J., 1992. The use of a canonical form in the construction of locally optimal designs for nonlinear problems. J. Roy. Statist. Soc. B 54, 569–583.
King, J., Wong, W.-K., 2000. Minimax *D*-Optimal designs for the logistic model. Biometrics 56, 1263–1267.
Läuter, E., 1974. Experimental design in a class of models. Math. Operationsforschung Statist. 5, 379–396.
Läuter, E., 1976. Optimal multipurpose designs for regression models. Math. Operationsforschung Statist. 7, 51–68.
Li, K.C., Notz, W., 1982. Robust designs for nearly linear regression. J. Statist. Plann. Inference 6, 135–151.
Marcus, M.B., Sacks, J., 1976. Robust designs for regression problems. In: Gupta, S.S., Moore, D.S. (Eds.), Statistical Theory and Related Topics II. Academic Press, New York, pp. 245–268.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall, CRC, London, Boca Raton, FL.
Meyer, R.K., Nachtsheim, C.J., 1988. Constructing exact *D*-optimal experimental designs by simulated annealing. Amer. J. Math. Management Sci. 3 & 4, 329–359.
Minkin, S., 1987. Optimal design for binary data. J. Amer. Statist. Assoc. 82, 1098–1103.
Sinha, S., Wiens, D.P., 2002. Robust sequential designs for nonlinear regression. Canad. J. Statist. 30, 601–618.
Sitter, R.R., 1992. Robust designs for binary data. Biometrics 48, 1145–1155.
White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–25.
Wiens, D.P., 1992. Minimax designs for approximately linear regression. J. Statist. Plann. Inference 31, 353–371.
Wiens, D.P., Zhou, J., 1999. Minimax designs for approximately linear models with AR(1) errors. Canad. J. Statist. 27, 781–794.