UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Working Paper M08/03
Methodology

# Robust Designs For Binary Data: Applications Of Simulated Annealing

D. C. Woods

## Abstract

When the aim of an experiment is the estimation of a Generalised Linear Model (GLM), standard designs from linear model theory may prove inadequate. This paper describes a flexible approach for finding designs for experiments to estimate GLMs through the use of D-optimality and a simulated annealing algorithm. A variety of uncertainties in the model can be incorporated into the design search, including the form of the linear predictor, through use of a robust design selection criterion and a postulated model space. New methods appropriate for screening experiments and the incorporation of correlations between possible model parameters are described through examples. An updating formula for D-optimality under a GLM is presented which improves the computational efficiency of the search.

# Robust designs for binary data: applications of simulated annealing

D. C. WOODS*[1]
Southampton Statistical Sciences Research Institute,
University of Southampton,
Southampton, UK

**Abstract**

When the aim of an experiment is the estimation of a Generalised Linear Model (GLM), standard designs from linear model theory may prove inadequate. This paper describes a flexible approach for finding designs for experiments to estimate GLMs through the use of $D$-optimality and a simulated annealing algorithm. A variety of uncertainties in the model can be incorporated into the design search, including the form of the linear predictor, through use of a robust design selection criterion and a postulated model space. New methods appropriate for screening experiments and the incorporation of correlations between possible model parameters are described through examples. An updating formula for $D$-optimality under a GLM is presented which improves the computational efficiency of the search.

Keywords: Generalised linear models; Optimal design; Prior information; Screening experiments; Simulation

## 1. Introduction

Many experiments in science and industry involve the measurement of binary responses, such as success or failure of a drug (e.g. [1]) or a component under test (e.g. [2]). Often, generalised linear models (GLMs, see [3]) provide an appropriate model for such a response. Although these models have received considerable attention in the literature, the development of methods of designing efficient experiments to estimate GLMs has lagged behind. A difficulty is that, in common with nonlinear models generally, the information matrix for the parameter estimators depends on the unknown values of these parameters. This has led to the development of *locally* optimal designs which are optimal for *given* values of the model parameters [4]. Such designs, however, may have poor efficiency for other parameter values. The issue of parameter dependance has limited the type of design problems for GLMs that can be addressed.

Initial work to overcome the parameter dependence problem mainly concentrated on small experiments with one or two variables. There are three main approaches: sequential experimentation [5, 6], maximin designs [7] and Bayesian designs [8]. More recent work [9] has used computer intensive optimisation methods to find designs for several variables which are robust to misspecification of the parameter values. In addition, use of the optimum-in-average (or compromise) criterion by Woods et al. [10] has allowed designs to be found that are also robust to the choice of link function and form of the linear predictor (see also [11]).

In this paper we describe a simulated annealing algorithm for finding robust designs and introduce techniques to select designs for GLMs using an Information Capacity criterion and an approach which enables correlations between the possible values of the different parameters in a model to be taken into account in selecting a design. The performance of the designs obtained is illustrated through examples.

In Section 2, GLMs are introduced and the design selection criterion from [10] is briefly described. The simulated annealing algorithm is given in Section 3 and an updating procedure for the criterion's objective function is outlined in Section 4. The new techniques are evaluated in Section 5 via two examples. The performance and tuning of the simulated annealing algorithm is discussed in Section 6.

## 2. Generalised linear models and a design selection criterion

Consider an experiment involving $k$ variables and a design $\xi(N)$ composed of $N$ points, not necessarily distinct. Each design point is defined by a vector $\boldsymbol{x}_j = (x_{1j} \ldots, x_{kj})'$, where $x_{ij}$ holds the value taken by the $i$th variable in the $j$th run of the experiment where, after scaling if necessary, $-1 \leq x_{ij} \leq 1$ $(i = 1, \ldots, k;\ j = 1, \ldots, N)$. Thus, each $\boldsymbol{x}_j$ defines the treatment to be applied to the $j$th unit in the experiment. Suppose that the observation from the $j$th design point, $Y_j$, follows a distribution from the exponential family. Further, the experimental units are assumed to be exchangeable in the sense that the distribution of an observation from the $j$th design point depends only on the treatment applied to the $j$th unit.

A GLM to describe $Y_j$ has three components [3, p.27]:

1. a distribution for the response,
2. a *linear predictor* $\eta_j = f(\boldsymbol{x}_j)'\boldsymbol{\beta}$, where $f(\boldsymbol{x}_j)$ is a $p \times 1$ vector of known functions of the $k$ explanatory variables and $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown model parameters, and
3. a *link function* $g(\cdot)$ that relates the mean response from the $j$th design point to the linear predictor through $g(\mu_j) = \eta_j$.

Hence, for a given distribution, a GLM is completely specified by $s = (g, \eta, \boldsymbol{\beta})$. We shall refer to $s$ as a model.

For binary data, each $Y_j$ follows an independent Bernoulli distribution with success probability $\pi_j$. If there are $n$ distinct treatments to be applied to the $N$ units in the experiment, the exchangeability of the units allows the number of successes on the $t$th treatment to be described by a binomial$(m_t, \pi_t)$ distribution, where $m_t$ is the number of units to which the $t$th treatment is applied and $\pi_t$ is the probability of success induced by the $t$th treatment $(t = 1, \ldots, n)$. Appropriate link functions include the probit, the complementary log-log and the logit link.

The asymptotic variance-covariance matrix of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is given by the inverse of the Fisher Information matrix $M(\xi(N), s) = X'WX$, where $X$ is the model matrix and $W$ is a diagonal weight matrix with $j$th diagonal element $(\mathrm{d}\mu_j/\mathrm{d}\eta_j)^2(\mathrm{Var}(Y_j))^{-1}$ [3, p.119]. For example, for Binomial data and the logit link, $W$ has entries $\mu_j(1-\mu_j)$.

If all three components of $s$ are known, then the application of a standard optimality criterion, such as $D$-optimality (see, for example, [12]) within a search algorithm may be used to find a locally optimal design for several variables. As this knowledge is often unlikely to be available in practice, we follow the approach of Woods et al. [10] and represent uncertainty in the model $s = (g, \eta, \boldsymbol{\beta})$ through sets $\mathcal{G}$, $\mathcal{N}$ and $\mathcal{B}$ of possible link functions, linear predictors and model parameters, respectively. These sets may be incorporated into a criterion for design selection that maximizes an objective function $\Phi$ obtained by integrating a *local* objective function $\phi(\xi(N), s)$ across $\mathcal{M} = \mathcal{G} \times \mathcal{N} \times \mathcal{B}$ to give

$$\Phi\left[\xi(N)\right] = \int_{\mathcal{M}} \phi\left[\xi(N), s\right]\, dF(s)\,, \tag{1}$$

where $F$ is an appropriate cumulative distribution function. This criterion was first proposed for linear models in the seminal work of Läuter [13]. From a Bayesian perspective, (1) can be viewed as the pre-posterior loss from an asymptotic Normal approximation to the posterior distribution [8].

The expression (1) is computationally expensive to approximate within a design search. The computational burden can be reduced by finding model-robust (or compromise) designs through maximisation of the surrogate objective function

$$\Phi_s\left[\xi(N)\right] = \sum_{s\in\mathcal{S}} \phi\left[\xi(N), s\right] p(s)\,, \tag{2}$$

where $\mathcal{S} = \mathcal{G}_s \times \mathcal{N}_s \times \mathcal{B}_s$, with $\mathcal{G}_s$, $\mathcal{N}_s$ and $\mathcal{B}_s$ finite subsets of $\mathcal{G}$, $\mathcal{N}$ and $\mathcal{B}$, respectively [10], and where $p(s)$ is a probability mass functon. In this paper, we concentrate on finding compromise designs under $D$-optimality for $p(s)$ a uniform probability mass function, resulting in (2) having the form

$$\Phi_s^D\left[\xi(N)\right] = \frac{1}{|S|} \sum_{s\in\mathcal{S}} \log \phi_D\left[\xi(N), s\right]\,, \tag{3}$$

where $\phi_D[\xi(N), s] = (|X'WX|)^{1/p_s}$, $p_s$ is the number of parameters to be estimated in model $s$, and both $X$ and $W$ are dependent on the model $s$. Hence, in this context, a design $\xi(N)$ is robust to the choice of model $s$ if it achieves good average performance across $\mathcal{S}$ under $D$-optimality. A robust-optimal design $\xi^*(N)$ maximises the objective function (3).

## 3. Implementation

A simulated annealing algorithm [14, 15], modelled on the cooling of materials in metallurgy, may be used to solve the optimization problem. This is a probabilistic optimization technique, not a greedy algorithm, which means that changes to the current *state* (a design in our problem) are accepted according to a transition probability. In the original formulation for a minimization problem, transitions which decrease the *energy function* (objective function) are accepted with probability 1 but transitions that increase the energy function also have non-zero probability of being accepted, allowing the algorithm to move away from local optima in the search space. The probability of accepting a move that increases the value of the energy function is given by the Boltzmann-Gibbs distribution, as used in the Metropolis-Hastings algorithm [16]. Under this formulation, the probability of accepting an "uphill" transition decreases with the increase in energy between the initial state and the transition, and also decreases as the *temperature* of the system decreases. The initial temperature is user-controlled and often a geometric cooling scheme is employed with the temperature decreasing by a fixed proportion after a set number of iterations. A useful review of simulated annealing is given by Spall [17].

The annealing algorithm described below, for continuous variables, is similar in implementation to the algorithm for linear models [18]. A transition is via a random perturbation made to a design point $x_{ij}$, so that the new point $x_{ij}^{per}$ is given by

$$x_{ij}^{per} = \min\left\{1, \max\left[-1, x_{ij} + ud(T)\right]\right\}\,,$$

where $u$ is a realisation of a uniform random number from $[-1, 1]$ and $d(T)$ is the size of the maximum allowed perturbation at temperature $T$. The perturbed design, $\xi^{per}(N)$, is compared

to the original design, $\xi(N)$, via calculation of the value of the objective function (3) for each design. It is accepted with probability

$$\alpha_{ij} = \begin{cases} 1 & \text{if } \Phi_s^D[\xi^{per}(N)] > \Phi_s^D[\xi(N)] \\ \min\left\{1, \exp\left(\dfrac{\Phi_s^D[\xi^{per}(N)] - \Phi_s^D[\xi(N)]}{T}\right)\right\} & \text{if } \Phi_s^D[\xi^{per}(N)] \leq \Phi_s^D[\xi(N)] \end{cases}$$

for a perturbation of the value of the $i$th variable for the $j$th run ($i = 1, \ldots, k$; $j = 1, \ldots, N$). In general, any objective function $\Phi(\cdot)$ may be used instead of $\Phi_s^D(\cdot)$. Changes which improve the design are always accepted; changes which result in a singular design, that is, a design that does not allow the model to be estimated, are always rejected. As the system cools, i.e. the temperature parameter decreases, the probability of accepting a poor move also decreases. When the temperature is zero, a greedy algorithm results. Both the temperature $T$ and perturbation size $d(T)$ are decreased geometrically and the temperature is cooled continually. In order to reach an equilibrium state with respect to the Boltzmann-Gibbs distribution, $d(T)$ is only decreased when the average number of accepted moves, averaged across all $f \times N$ values of the variables in the design, lies between 0.4 and 0.6 [16]. The value of this average is assessed every $m$ iterations, where $m$ is user-specified; see Section 6.

## 4. Updating formula

An often quoted advantage of the $D$-optimality criterion for linear models is the availability of updating formulae ([19, p.162], [20]) for both the determinant and inverse of the information matrix. These formulae eliminate the need for costly evaluations of matrix determinants at every iteration of the optimization. In this section, we derive updating formulae for the $D$-optimality criterion for finding designs under a GLM. For simplicity of notation, the dependence of the information matrix $M\{\xi(N), s\}$ on the model $s$ is suppressed.

Assuming a fixed model $s$, consider a design $\xi(N+1)$ formed by the addition of a run $\boldsymbol{x}_{N+1}$ to a design $\xi(N)$. The additivity of information matrices implies that

$$M[\xi(N+1)] = X_N' W_N X_N + w_{N+1} f(\boldsymbol{x}_{N+1}) f(\boldsymbol{x}_{N+1})' ,$$

where $X_N$ and $W_N$ are the model and weight matrices, respectively, under design $\xi(N)$ and

$$w_{N+1} = (\mathrm{d}\mu_{N+1}/\mathrm{d}\eta_{N+1})^2 [\mathrm{Var}(Y_{N+1})]^{-1} .$$

For logistic regression, $w_{N+1} = \mu_{N+1}(1 - \mu_{N+1})$. Using results on partitioned matrices [21, p.183-184],

$$\begin{aligned} |M[\xi(N+1)]| &= |X_N' W_N X_N + w_{N+1} f(\boldsymbol{x}_{N+1}) f(\boldsymbol{x}_{N+1})'| \\ &= |M[\xi(N)]| \left\{1 + w_{N+1} f(\boldsymbol{x}_{N+1})' M^{-1}[\xi(N)] f(\boldsymbol{x}_{N+1})\right\} \end{aligned}$$

and

$$M^{-1}[\xi(N+1)] = M^{-1}[\xi(N)] - \frac{w_{N+1} M^{-1}[\xi(N)] f(\boldsymbol{x}_{N+1}) f(\boldsymbol{x}_{N+1})' M^{-1}[\xi(N)]}{1 + w_{N+1} f(\boldsymbol{x}_{N+1})' M^{-1}[\xi(N)] f(\boldsymbol{x}_{N+1})} .$$

Similar results can be derived for a design $\xi(N-1)$ formed by removing an arbitrary design point $\boldsymbol{x}_j$ from $\xi(N)$ $(j = 1, \ldots, N)$. Thus it is straightforward to obtain the following result for removing $\boldsymbol{x}_u$ from $\xi(N)$ and replacing it with point $\boldsymbol{x}_v$ to form a new design $\xi^e(N)$:

$$\begin{aligned} |M[\xi^e(N)]| &= |M[\xi(N)] - w_u f(\boldsymbol{x}_u) f(\boldsymbol{x}_u)' + w_v f(\boldsymbol{x}_v) f(\boldsymbol{x})'_v| \\ &= |M[\xi(N)]| \left(1 + \Delta(\boldsymbol{x}_u, \boldsymbol{x}_v, w_u, w_v)\right) , \end{aligned} \tag{4}$$

where $\Delta(\cdot)$ is a generalisation of Fedorov's delta function given by

$$\begin{aligned} \Delta(\boldsymbol{x}_u, \boldsymbol{x}_v, w_u, w_v) &= w_u f(\boldsymbol{x}_u)' M^{-1}[\xi(N)] f(\boldsymbol{x}_u) - w_v f(\boldsymbol{x}_v)' M^{-1}[\xi(N)] f(\boldsymbol{x}_v) \\ &\quad - w_u w_v f(\boldsymbol{x}_u) M^{-1}[\xi(N)] f(\boldsymbol{x}_u) f(\boldsymbol{x}_v) M^{-1}[\xi(N)] f(\boldsymbol{x}_v) \\ &\quad + w_u w_v \left\{ f(\boldsymbol{x}_v)' M^{-1}[\xi(N)] f(\boldsymbol{x}_u) \right\}^2 . \end{aligned}$$

As each transition in the annealing algorithm is made by perturbing a single design point, the objective function (3) is updated for each new design by repeated use of (4) for each model $s \in \mathcal{S}$. Use of the updating formula provides evaluations of the objective function that do not depend on $N$, and results in considerable computational savings for large experiment sizes as discussed in Section 6.

## 5. Applications

### 5.1. Information Capacity

In the early stages of experimentation, the aim is often to identify those variables that have a substantive impact on the response. For such screening experiments under linear models, a design is often chosen for which the designs formed from *projections* into subsets of the variables are as efficient as possible [see, for example, 22]. This ensures that *submodels* of a specified larger model can be estimated efficiently from data obtained using the given design. This idea has been formalised for linear models in the *Information Capacity* (IC) criterion [23, 25, 24], using a weighted average of the $D$ criterion across all submodels. This criterion can be viewed as a special case of (3) in which $\mathcal{S}$ is a set of all submodels of the full $k$ variable model.

As an example, suppose that an experiment has four variables and the response is described by a logistic regression with $\eta_j = \beta_0 + \sum_{i=1}^{4} \beta_i x_{ij}$. Then the 15 submodels, including the full model, can be written as

$$\eta_j^{(m)} = \beta_0 + \sum_{i=1}^{4} \beta_i x_{ij} I(m, i) \quad (m = 1 \ldots, 15;\ j = 1, \ldots, N) , \tag{5}$$

where $I(m, i) = 1$ if variable $i$ is in model $m$, and 0 otherwise. The submodels are indexed as shown in Table 1 with the linear predictors given in the notation of [26]. For example, submodel 5 is a logistic regression model with linear predictor $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

The simulated annealing algorithm was used to find a maximum IC design in 16 runs, $\xi_1(16)$, for the model space $\mathcal{S} = \{\text{logit}, \eta^{(m)}, \boldsymbol{\beta}_0;\ m = 1, \ldots, 15\}$, with $\boldsymbol{\beta}_0' = (0, 1, 0, 3, 0.5)$. The efficiency of the design for each of the 15 submodels is given in Table 1, where efficiency is defined relative to a locally optimal design $\xi^*$ as

Table 1: Submodel efficiencies for the maximum information capacity design in Section 5.1.

| Model | | Efficiency | Model | | Efficiency |
|---|---|---|---|---|---|
| 1 | $x_1$ | 0.94 | 9 | $x_2 + x_4$ | 0.93 |
| 2 | $x_2$ | 0.94 | 10 | $x_3 + x_4$ | 0.90 |
| 3 | $x_3$ | 0.88 | Average (2 factors) | | 0.92 |
| 4 | $x_4$ | 0.94 | 11 | $x_1 + x_2 + x_3$ | 0.91 |
| Average (1 factor) | | 0.93 | 12 | $x_1 + x_2 + x_4$ | 0.91 |
| 5 | $x_1 + x_2$ | 0.93 | 13 | $x_1 + x_3 + x_4$ | 0.91 |
| 6 | $x_1 + x_3$ | 0.91 | 14 | $x_2 + x_3 + x_4$ | 0.90 |
| 7 | $x_1 + x_4$ | 0.93 | Average (3 factors) | | 0.91 |
| 8 | $x_2 + x_3$ | 0.91 | 15 | $x_1 + x_2 + x_3 + x_4$ | 0.90 |

$$\frac{\phi_D(\xi_i(16), s)}{\phi(\xi^*(16), s)}, \quad s \in \mathcal{S}. \tag{6}$$

The maximum IC design is highly efficient for each of the models, with an overall average efficiency of 0.91. Model 3, which has a single variable, $x_3$, has the lowest efficiency of 0.88. Note that the coefficient $\beta_3$ of $x_3$ has the largest absolute value, and also that the other models which involve this factor also have slightly lower efficiencies than models which do not involve $x_3$.

We now compare $\xi_1$ with four further designs found for different parameter spaces, $\mathcal{B}$, for the full four variable model. The following independent intervals are used for each parameter, centered on the above values of $\beta_0, \ldots, \beta_4$:

$$\beta_i \in \begin{cases} (-3, 3) & i = 0 \\ (-2, 4) & i = 1 \\ (-3, 3) & i = 2 \\ (0, 6) & i = 3 \\ (-2.5, -3.5) & i = 4 \end{cases} \tag{7}$$

The product of these intervals defines a parameter space $\mathcal{B} \subset \mathbb{R}^5$. Four different designs, $\xi_2, \ldots, \xi_5$ were found, each with 16 runs for the four variable model and the logit link using the following four choices of $\mathcal{B}_s$:

- $\xi_2$: $\mathcal{B}_s$ is the centroid of $\mathcal{B}$
- $\xi_3$: $\mathcal{B}_s$ is a $2^{5-2}$ fractional factorial design where the levels of factor $i$ are the limits of the range of $\beta_i$, augmented by the centroid of $\mathcal{B}$
- $\xi_4$: $\mathcal{B}_s$ is a 9 point $U$-optimal set (as in SAS PROC OPTEX [27]) chosen from a candidate set of $6^5$ equally spaced points across $\mathcal{B}$
- $\xi_5$: $\mathcal{B}_s$ is a 9 point Latin Hypercube Sample (LHS), with uniform margins, selected from $\mathcal{B}$

Design $\xi_2$ is the locally optimal design for $\boldsymbol{\beta}_0' = (0, 1, 0, 3, 0.5)$; designs $\xi_3 - \xi_5$ are compromise designs across different sets of 9 models. Designs $\xi_2 - \xi_4$ were considered in [10]. Design $\xi_5$ employs a LHS as a designed sample of the parameter values from $\mathcal{B}$. A LHS is a commonly used

Table 2: Five number summaries (min., Q1, med., Q3, max.) for projections of five compromise designs onto 1, 2, 3 and 4 variables.

| Design | No. of variables | |
|---|---|---|
| | 1 | 2 |
| $\xi_1$. IC | (0.25, 0.64, 0.87, 1.00, 1.00) | (0.13, 0.49, 0.64, 0.78, 1.00) |
| $\xi_2$. Centroid | (0.44, 0.75, 0.84, 0.89, 0.97) | (0.10, 0.53, 0.63, 0.73, 0.91) |
| $\xi_3$. Fraction | (0.25, 0.65, 0.85, 1.00, 1.00) | (0.12, 0.50, 0.63, 0.78, 1.00) |
| $\xi_4$. $U$-opt. | (0.45, 0.70, 0.87, 0.97, 0.99) | (0.27, 0.55, 0.65, 0.77, 1.00) |
| $\xi_5$. LHS | (0.34, 0.71, 0.87, 0.97, 1.00) | (0.20, 0.54, 0.65, 0.77, 1.00) |
| | 3 | 4 |
| $\xi_1$. IC | (0.13, 0.40, 0.48, 0.60, 1.00) | (0.12, 0.33, 0.39, 0.47, 0.98) |
| $\xi_2$. Centroid | (0.11, 0.40, 0.49, 0.57, 0.86) | (0.08, 0.31, 0.39, 0.47, 0.79) |
| $\xi_3$. Fraction | (0.13, 0.41, 0.49, 0.61, 1.00) | (0.12, 0.35, 0.40, 0.47, 1.00) |
| $\xi_4$. $U$-opt. | (0.19, 0.45, 0.53, 0.62, 1.00) | (0.11, 0.38, 0.44, 0.51, 0.92) |
| $\xi_5$. LHS | (0.12, 0.44, 0.52, 0.62, 0.97) | (0.08, 0.35, 0.43, 0.51, 0.90) |

method in computer experiments (see, for example, [28]), when it is envisaged that only a subset of the variables will affect the response. Design $\xi_1$, found using the IC criterion, compromises across the possible submodels (projections) of the variables, whereas designs $\xi_3 - \xi_5$ compromise across the parameter space for the full model.

To assess the performance of these five designs, simulation studies were carried out. For each of $\beta_0, \ldots, \beta_4$, 1000 parameter vectors were generated from the ranges in (7) by quasi-random uniform sampling using a Sobol sequence [29, Ch.3]. Locally optimal exact designs $\xi^*(16)$ were found using simulated annealing for each parameter vector under each of the 15 models from (5). Designs $\xi_1 - \xi_5$ were then assessed by calculating the efficiency (6) for each model and parameter combination. Table 2 gives the five-number summaries for the efficiencies from these simulations for $k$-factor models ($k = 1, \ldots, 4$).

The design with the consistently best performance (highest efficiencies) is $\xi_4$, which compromises across the $U$-optimal set. In particular, it has the highest minimum efficiency for each size of submodel except for the full model with four variables. The locally optimal design ($\xi_2$) at the centroid of the 5-dimensional parameter space also performs well, as was found for the full model by [10]. Design $\xi_1$, found using the IC criterion, is the poorest design, especially for projections onto one and two variables. This study suggests that uncertainty in the values of the parameters has more impact on design performance than uncertainty in the numbers of variables.

### 5.2. Prior information for correlated parameters

An experiment from the food industry on protected atmosphere packing of potatoes was described in [10]. The experiment had three quantitative variables which could be varied and was performed using a 16 run central composite design (CCD). Due to separation in the data [30], a penalised likelihood method [31] was used to obtain parameter estimates and standard errors for three linear predictors, each with the logit link:

$$
\begin{aligned}
\eta_j^{(1)} &= \beta_0 + \sum_{i=1}^{3} \beta_i x_{ij} \,, \\
\eta_j^{(2)} &= \beta_0 + \sum_{i=1}^{3} \beta_i x_{ij} + \sum_{k=1}^{3} \sum_{l>k}^{3} \beta_{kl} x_{kj} x_{lj} \,, \\
\eta_j^{(3)} &= \beta_0 + \sum_{i=1}^{3} \beta_i x_{ij} + \sum_{k=1}^{3} \sum_{l>k}^{3} \beta_{kl} x_{kj} x_{lj} + \sum_{i=1}^{3} \beta_{ii} x_{ij}^2 \,.
\end{aligned}
$$

We further investigate this example by considering the impact on compromise design selection of (i) correlation between the parameters, and (ii) the size of the parameter space.

For $\eta^{(1)}$, $\eta^{(2)}$ and $\eta^{(3)}$, the maximum penalised likelihood estimates are

$$
\begin{aligned}
\eta^{(1)} &: \hat{\boldsymbol{\beta}}_1 = (-0.28, 0, -0.76, -1.15) \\
\eta^{(2)} &: \hat{\boldsymbol{\beta}}_2 = (-1.44, 0, -1.95, -2.36, 0, 0, -2.34) \\
\eta^{(3)} &: \hat{\boldsymbol{\beta}}_3 = (-2.93, 0, -0.52, -0.79, 0, 0, -0.66, 0.94, 0.79, 1.82) \,.
\end{aligned}
$$

The variance-covariance matrices for these estimators are given by

$$
\eta^{(1)} : \hat{V}_1 = \begin{pmatrix}
0.33 & 0 & 0.03 & 0.04 \\
 & 0.45 & 0 & 0 \\
 & & 0.52 & 0.14 \\
 & & & 0.56
\end{pmatrix}
$$

$$
\eta^{(2)} : \hat{V}_2 = \begin{pmatrix}
1.28 & 0 & 1.12 & 1.12 & 0 & 0 & 1.16 \\
 & 0.77 & 0 & 0 & 0.45 & 0.32 & 0 \\
 & & 1.73 & 1.29 & 0 & 0 & 1.36 \\
 & & & 1.91 & 0 & 0 & 1.53 \\
 & & & & 1.11 & 0.41 & 0 \\
 & & & & & 0.98 & 0 \\
 & & & & & & 2.17
\end{pmatrix}
$$

$$
\eta^{(3)} : \hat{V}_3 = \begin{pmatrix}
3.91 & 0 & 0.11 & 0.01 & 0 & 0 & -0.01 & -1.49 & -1.34 & -1.64 \\
 & 0.54 & 0 & 0 & 0.05 & -0.03 & 0 & 0 & 0 & 0 \\
 & & 0.58 & 0 & 0 & 0 & -0.04 & -0.08 & 0.11 & -0.08 \\
 & & & 0.51 & 0 & 0 & 0.04 & -0.01 & -0.01 & -0.03 \\
 & & & & 0.75 & 0.01 & 0 & 0 & 0 & 0 \\
 & & & & & 0.74 & 0 & 0 & 0 & 0 \\
 & & & & & & 0.75 & 0.01 & -0.01 & 0.01 \\
 & & & & & & & 1.39 & 0.15 & 0.37 \\
 & & & & & & & & 1.37 & 0.26 \\
 & & & & & & & & & 1.41
\end{pmatrix},
$$

where the rows and columns are ordered as intercept, linear terms, interactions, and squared terms. We represent uncertainty in the parameter values by a multivariate normal distirbution $N(\hat{\boldsymbol{\beta}}_i, \gamma^2 \hat{V}_i)$ for each $\eta^{(i)}$ $(i = 1, 2, 3)$. For $\gamma = 0.1, 0.5, 1$, the following 16-run designs are considered:

$\xi_1$: the CCD used in the original experiment, with 8 factorial points, 6 axial points with $\alpha = 1.2872$ and 2 centre points.

$\xi_{2i}$: the locally $D$-optimal design for $\hat{\boldsymbol{\beta}}_i$.

$\xi_{3i}^{\gamma}$: a $D$-optimal compromise design across $S = \{\text{logit}, \eta^{(i)}, \mathcal{B}_{s3}^{(i)};\ i = 1, 2, 3\}$, where $\mathcal{B}_{s3}^{(i)}$ is a set of 16 parameter vectors selected using a LHS from $N(\hat{\boldsymbol{\beta}}_i, \gamma^2 \hat{V}_i)$ generated by the method of Stein [32].

$\xi_{4i}^{\gamma}$: a $D$-optimal compromise design for $S = \{\text{logit}, \eta^{(i)}, \mathcal{B}_{s4}^{(i)};\ i = 1, 2, 3\}$, where $\mathcal{B}_{s4}^{(i)}$ is a set of 200 parameter vectors selected from $N(\hat{\boldsymbol{\beta}}_i, \gamma^2 \hat{V}_i)$ using a Sobol quasi-random sequence.

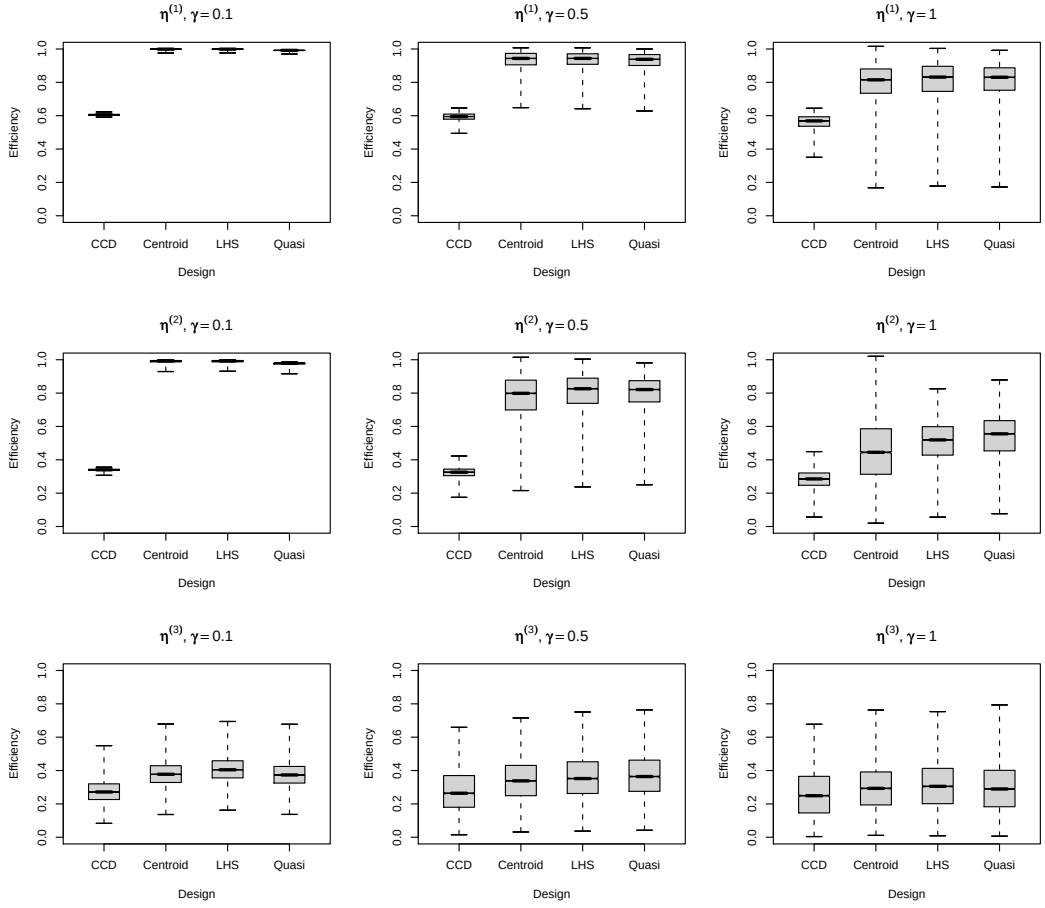


Figure 1: Design efficiencies for robust designs from Section 5.2.

For each of the nine combinations of $\eta^{(i)}$ and $\gamma$, the designs were assessed through simulation, with 1000 sets of parameter values drawn from $N(\hat{\boldsymbol{\beta}}_i, \gamma^2 \hat{V}_i)$ using transformed Sobol quasi-random sequences. Figure 1 shows the results of these simulation studies for each design as box plots of efficiency, defined in (6). In the figure, $\gamma$ increases with the column number and $\eta^{(i)}$ changes with the rows. Larger parameter spaces result from moving both right and down in the figure. Key points from the figure are (i) the poor performance of the CCD ($\xi_1$) for small parameter spaces (e.g. for small models or small $\gamma$); (ii) the strong performance of the

centroid designs ($\xi_{2i}$), which have similar efficiencies to the compromise designs $\xi_{3i}^{\gamma}$ and $\xi_{4i}^{\gamma}$ for each scenario; (iii) the comparable performance of $\xi_{3i}^{\gamma}$ and $\xi_{4i}^{\gamma}$, showing that the use of 200 models (parameter vectors) from $\mathcal{B}$ offers little or no improvement over the use of 16 models, whilst being considerably more computationally expensive; (iv) when there is less information on the model, i.e. for second-order models with large $\gamma$, the CCD ($\xi_1$) is competitive with the other, more tailored, designs.

## 6. Performance and tuning of the annealing algorithm

The performance of an annealing algorithm depends upon the choice of cooling schedule and the scheme chosen for transitions between states (designs). The initial temperature of the system is set empirically through the evaluation of 1000 perturbations of a randomly selected design. This temperature is determined such that the probability of accepting the single transition which produced the 950th largest difference between the objective function (3) for the original design and the perturbed design is equal to 0.5. The best design from this set of 1000 is used as the starting design for the annealing algorithm.

The algorithm is controlled by three tuning parameters, the geometric cooling rate ($\delta_t$), the geometric rate of decrease ($\delta_s$) in the size of the perturbation for $x_{ij}$, and the number of iterations ($m$) between potential changes in $d(T)$. Although theoretically guaranteed to converge to the optimal solution if cooled sufficiently slowly, in practice the annealing algorithm is a heuristic optimisation method. Hence the values of the tuning parameters are vital in controlling the trade-off between computational effort and quality of the final solution. The performance of the algorithm for 10 different random starts is summarised in Table 3 for design $\xi_3$ from Example 1, including indicative timings for 16 runs (as in the example) and also 24 runs. The results were obtained by running the algorithm on a 3.2Ghz Intel Pentium IV desktop PC with $\gamma_t = 0.9$ and $m = 20$. Empirical studies for this problem have shown that $\gamma_s$ and the updating formula have most impact on the computational cost of the design search. The computational savings from using the updating formula are clear, especially for the larger run size and for smaller $\gamma_s$. Also note that the use of larger $\gamma_s$, which results in a more rapid decrease in the sizes of the $x_{ij}$ perturbations, can still produce designs which perform well.

Table 3: Performance of the SA algorithm for Section 5.1, using 10 random starts.

| Runs | $\delta_s$ | Update | Average Run Time | Average Value of Objective Function |
|---|---|---|---|---|
| 16 | 1.1 | No | 1m 36s | 2.30 |
| 16 | 1.1 | Yes | 1m 26s | 2.24 |
| 16 | 1.01 | No | 9m 21s | 2.28 |
| 16 | 1.01 | Yes | 5m 29s | 2.26 |
| 24 | 1.1 | No | 3m 33s | 97.47 |
| 24 | 1.1 | Yes | 2m 33s | 97.09 |
| 24 | 1.01 | No | 22m 15s | 98.62 |
| 24 | 1.01 | Yes | 9m 46s | 98.72 |

In common with other heuristic algorithms, simulated annealing is not guaranteed to find

the optimal design after a finite number of iterations. Consequently it is important to run several design searches and use the best design found. It is easy to implement the algorithm in a simple parallel fashion, with each separate design search carried out on a different machine or on a different node of a computational cluster. All the designs in this paper were found using multiple starts of the algorithm on a Beowulf cluster.

## 7. Discussion

We have described a flexible algorithmic approach to finding designs for GLMs that can incorporate different numbers of runs and factors, and a variety of linear predictors. The resulting designs can substantially outperform standard designs from linear model theory and can incorporate uncertainty in all aspects of the model. Simulated annealing has some particular benefits for this design search problem. These include stochastic transitions between designs to avoid local optima, and a candidate-list free continuous search which can accommodate larger problems than an exchange algorithm. The provision of an updating formula for $D$-optimality under GLMs increases the efficiency of the design search, making the algorithm a practical design tool. For moderate-sized problems, the algorithm runs in "coffee break" time; for larger problems, an overnight search may be necessary to find the best design but highly efficient designs can be found in much less time.

Although the methods have been described and illustrated for binary data, they can be employed to find designs for models with other types of categorical or count data. For example, Poisson log-linear regression [33] and multinomial models [34, Ch.7]. For multinomial data, the response may be modelled using a multivariate GLM [35], with potentially different linear predictors used to describe the probability of success in each category, together with an identifiability constraint. The design problem is then a natural generalisation of that given in Section 2, using a suitably defined information matrix [36] which will depend on the parameters from each linear predictor.

The usefulness of the resulting designs for binary data has been demonstrated through new applications that have several striking features: the poor performance of the standard linear model designs for all but the weakest information on the model; the remarkable robustness of locally optimal designs for the centroid of model spaces; and the ability of designs for several variables to estimate efficiently submodels which contain subsets of the variables. The algorithm can also be used to search heuristically for approximate designs, where the the design is represented as a probability measure with finite support across the design region [37]. The algorithm is written in C++ using the Gnu Scientific Library [38] and source code is available, with documentation, at `http://www.soton.ac.uk/~davew/glm_alg`.

## Acknowledgements

This work was supported by the EPSRC grant EP/C008863/1.

## References

[1] D. Collett *Modelling Binary Data*, 2nd Chapman and Hall, Boca Raton, Fl, 2002.

[2] R.H. Myers, D.C. Montgomery, and G.G. Vining *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley, New York, 2002.

[3] P. McCullagh and J.A. Nelder *Generalized Linear Models*, 2nd Chapman and Hall, New York, 1989.

[4] A.C. Atkinson, *Generalized linear models and response transformation*, in *Response Surface Methodology*, A. I. Khuri ed., World Scientific, Singapore, 2006, pp. 173–202.

[5] K.M. Abdelbasit and R.L. Plackett, *Experimental Design for Binary Data*, Journal of the American Statistical Association 78 (1983), pp. 90–98.

[6] C.F.J. Wu, *Efficient Sequential Designs with Binary Data*, Journal of the American Statistical Association 80 (1985), pp. 974–984.

[7] J. King and W.K. Wong, *Minimax D-optimal Designs for the Logistic Model*, Biometrics 56 (2000), pp. 1263–1267.

[8] K. Chaloner and K. Larntz, *Optimal Bayesian Design Applied to Logistic Regression Experiments*, Journal of Statistical Planning and Inference 21 (1989), pp. 191–208.

[9] A.C. Atkinson, C.G.B. Demetrio, and S.S. Zocchi, *Optimum dose levels when males and females differ in response*, Journal of the Royal Statistical Society Series C 44 (1995), pp. 213–226.

[10] D.C. Woods et al., *Designs for generalized linear models with several variables and model uncertainty*, Technometrics 48 (2006), pp. 284–292.

[11] H.A. Dror and D.M. Steinberg, *Robust experimental design for multivariate generalized linear models*, Technometrics 48 (2006), pp. 520–529.

[12] V.V. Fedorov and P. Hackl *Model-Oriented Design of Experiments*, Springer, New York, 1997.

[13] E. Läuter, *Experimental Design in a Class of Models*, Mathematische Operationsforschung und Statistik 5 (1974), pp. 379–396.

[14] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, *Optimization by simulated annealing*, Science 220 (1983), pp. 671–680.

[15] V. Cerny, *A thermodynamic approach to the travelling salesman problem: an efficient simulation algorithm*, Journal of Optimization Theory and Algorithms 45 (1985), pp. 41–51.

[16] N. Metropolis et al., *Equation of state calculations by fast computing machines*, Journal of Chemical Physics 21 (1953), pp. 1087–1092.

[17] J.P. Spall *Introduction to Stochastic Search and Optimization*, Wiley, New York, 2003.

[18] L.M. Haines, *The application of the annealing algorithm to the construction of exact optimal designs for linear regression models*, Technometrics 29 (1987), pp. 439–447.

[19] V.V. Fedorov *Theory of Optimal Experiments*, Academic Press, New York, 1972.

[20] N.K. Nyugen and A.J. Miller, *A review of some exchange algorithms for constructing discrete D-optimal designs*, Computational Statistics and Data Analysis 14 (1992), pp. 489–498.

[21] F.A. Graybill *Matrice with Applications in Statistics*, Duxbury, Belmont, CA, 1983.

[22] C.S. Cheng, *Projection properties of factorial designs for factor screening*, in *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*, A. Dean and S. Lewis, eds., Springer, New York, 2006, pp. 156–168.

[23] D.X. Sun, *Estimation Capacity and Related Topics in Experimental Designs*, Dept. of Statistics and Actuarial Science, University of Waterloo, 1993.

[24] W. Li and C.J. Nachtsheim, *Model-robust factorial designs*, Technometrics 42 (2000), pp. 345–352.

[25] C.F.J. Wu, *Construction of supersaturated designs through partially aliased interactions*, Biometrika 80 (1993), pp. 661–669.

[26] G.N. Wilkinson and C.E. Rogers, *Symbolic description of factorial models for analysis of variance*, Journal of the Royal Statistical Society Series C 22 (1973), pp. 392–399.

[27] SAS, Chapter title. *SAS QC Software Vol 1: Useage and Reference*, SAS Institute, Inc., Cary, NC (1995), .

[28] K.T. Fang, R. Li, and A. Sudjianto *Design and Modeling for Computer Experiments*, Chapman and Hall/CRC, Boca Raton, 2006.

[29] J.E. Gentle *Random Number Generation and Monte Carlo Methods*, 2nd Springer, New York, 2003.

[30] M.J. Silvapulle, *On the Existence of Maximum Likelihood Estimators for the Binomial Response Model*, Journal of the Royal Statistical Society, Ser. B 43 (1981), pp. 310–313.

[31] D. Firth, *Bias Reduction of Maximum Likelihood Estimates*, Biometrika 80 (1993), pp. 27–38.

[32] M. Stein, *Large sample properties of simulations using Latin Hypercube Sampling*, Technometrics 29 (1987), pp. 143–151.

[33] K.G. Russell et al., *D-optimal designs for Poisson regression models*, Statistica Sinica in press (2008).

[34] A. Agresti *Categorical Data Analysis*, 2nd Wiley, New York, 2002.

[35] G.F.V. Glonek and P. McCullagh, *Multivariate logistic models*, Journal of the Royal Statistical Society Series B 57 (1995), pp. 533–546.

[36] S.S. Zocchi and A.C. Atkinson, *Optimum experimental designs for multinomial logistic models*, Biometrics 55 (1999), pp. 437–444.

[37] D.C. Woods and S.M. Lewis, *Continuous Optimal designs for generalised linear models under model uncertainty*, Tech. Rep. 389, University of Southampton, School of Mathematics, 2005.

[38] M. Galassi et al. *GNU Scientific Library Reference Manual*, 2nd Network Theory Ltd, Bristol, UK, 2005.