

Designs for Generalized Linear Models

Anthony C. Atkinson*

David C. Woods†

*London School of Economics and Political Science, UK

†University of Southampton, UK

December 9, 2013

*Email: a.c.atkinson@lse.ac.uk

†Email: d.woods@southampton.ac.uk

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 1 |
| 2 | Generalized linear models | 2 |
| 2.1 | The family of models | 2 |
| 2.2 | The normal distribution | 3 |
| 2.3 | Binomial data | 3 |
| 2.4 | Poisson data | 4 |
| 2.5 | Gamma data | 5 |
| 3 | Optimal experimental designs | 6 |
| 3.1 | Theory | 6 |
| 3.2 | D-optimal designs | 9 |
| 3.3 | Design efficiency | 9 |
| 3.4 | Parameter uncertainty | 10 |
| 3.4.1 | Bayesian designs | 10 |
| 3.4.2 | Sequential designs | 10 |
| 3.4.3 | Minimax and maximin designs | 11 |
| 3.4.4 | Cluster designs | 12 |
| 3.5 | Small effects | 12 |
| 4 | Locally optimal designs | 12 |
| 4.1 | Binomial data: logistic regression in one variable | 12 |
| 5 | Optimal designs for two or more explanatory variables | 16 |
| 5.1 | Binomial data | 16 |
| 5.1.1 | Induced design region | 16 |
| 5.1.2 | First-order models | 17 |
| 5.1.3 | Induced design region for the logistic link | 19 |
| 5.1.4 | Theoretical results for a first-order predictor | 22 |
| 5.1.5 | A second-order response surface | 24 |
| 5.1.6 | Bayesian D-optimal designs | 27 |
| 5.2 | Gamma data | 28 |
| 5.2.1 | Theoretical results for first-order designs with the power link | 28 |
| 5.2.2 | Examples of first-order designs with the power link | 31 |
| 5.2.3 | Second-order response surface with the power link | 33 |
| 5.2.4 | Efficient standard designs for gamma models | 34 |
| 5.2.5 | Designs for the power link and for the Box-Cox link | 35 |
| 5.3 | Poisson data | 36 |

| | | |
|----------|---|-----------|
| 6 | Designs with dependent data | 41 |
| 6.1 | Random intercept model | 42 |
| 6.2 | Continuous block designs | 43 |
| 6.3 | Information matrix for a random intercept model | 43 |
| 6.4 | Approximating the information matrix using estimating equations | 44 |
| 6.5 | Comparison of approximations | 45 |
| 7 | Extensions and further reading | 47 |

1 Introduction

The methods of experimental design described in the majority of the earlier chapters are appropriate if the continuous response, perhaps after transformation, has independent errors with a variance that is known up to a multiplicative constant. (An exception is Chapter 5, which describes designs for correlated errors). However, this is not a characteristic of the Poisson and binomial distributions, where there is a strong relationship for these discrete random variables between mean and variance. The main emphasis of this chapter is on designs for generalized linear models appropriate for data from these and other distributions.

The classic account of generalized linear models is McCullagh and Nelder (1989). Issues in the design of experiments for these models are reviewed by Khuri et al. (2006); in addition to the methods of optimal experimental design, they consider stochastic approximation (Robbins and Monro, 1951) and adaptations of response surface methodology (Box and Draper, 1963, and Chapter 10 of this Handbook). Their emphasis is mainly on models with a single explanatory variable. On the other hand, the review of Atkinson (2006) focuses on optimal design and models with several explanatory variables as, to some extent, does the more recent review of Stufken and Yang (2012), where the emphasis is towards analytical results. Here we follow the approach of Atkinson, but focus on more recent results and on computational methods.

The assumptions of normality and constancy of variance for regression models enter the criteria of optimal design through the form of the information matrix $\mathbf{X}'\mathbf{X}$, where, as in other chapters, \mathbf{X} is the $n \times p$ model, or extended design, matrix. Other forms of information matrix arise from other distributions. See Atkinson et al. (2014). Given the appropriate information matrix, the principles of optimal design are the same as those described in earlier chapters. In designs for generalized linear models, the asymptotic covariance matrix of the parameters of the linear model is of the form $\mathbf{X}'\mathbf{U}\mathbf{X}$, where the $n \times n$ diagonal matrix of weights \mathbf{U} depends on the parameters of the linear predictor, on the error distribution and on the link between them. The dependence of the designs on parameters whose values are unknown prior to experimentation means that, in general, designs for generalized linear models require some specification of prior information.

In §2, we briefly review the class of generalized linear models, with particular emphasis on models for binomial, Poisson and gamma data. Some fundamental ideas in optimal design of experiments are reviewed in §3, and the optimality criteria employed in this chapter are introduced. We emphasize the reliance of optimal designs on the unknown parameter values, and discuss methods of overcoming this dependence. Locally optimal designs are

introduced in §4 for logistic models for binomial data with a single explanatory variable. In §5 we move on to designs for binomial, Poisson and gamma data with several explanatory variables. This latter section also includes results on Bayesian designs. In §6, we discuss designs for dependent non-normal data, for example, arising from blocked experiments, and demonstrate optimal design for generalized linear mixed models through an example. In §7, we give some extensions and suggestions for further reading.

2 Generalized linear models

2.1 The family of models

The family of generalized linear models extends normal theory regression to any distribution belonging to the one-parameter exponential family. As well as the normal (with known variance), this includes the gamma, Poisson and binomial distributions, all of which are important in the analysis of data. The three components of a generalized linear model are:

1. A distribution for the univariate response y with mean μ .
2. A linear predictor $\eta = \mathbf{f}'(\mathbf{x})\boldsymbol{\theta}$ where $\mathbf{f}(\mathbf{x})$ is a p -vector of known functions of the k explanatory variables \mathbf{x} , and $\boldsymbol{\theta}$ is a p -vector of unknown model parameters.
3. A link function $g(\mu) = \eta$, relating \mathbf{x} to the mean μ .

The distribution of y determines the relationship between the mean and the variance of the observations. The variance is of the form

$$\text{var}(y) = \phi V(\mu), \quad (1)$$

where ϕ is a ‘dispersion parameter’, equal to σ^2 for the normal distribution and equal to one for the binomial and Poisson distributions. The variance function $V(\mu)$ is specific to the error distribution.

The information matrix for a single observation at a point \mathbf{x} is

$$\mathbf{M}(\mathbf{x}; \boldsymbol{\theta}) = u(\mathbf{x}) \mathbf{f}(\mathbf{x}) \mathbf{f}'(\mathbf{x}), \quad (2)$$

with the weights for individual observations given by

$$u(\mathbf{x}) = V^{-1}(\mu) \left(\frac{d\mu}{d\eta} \right)^2. \quad (3)$$

These weights depend both on the distribution of y and on the link function.

2.2 The normal distribution

The linear multiple regression model can be written as

$$E(y) = \mu = \eta = \mathbf{f}'(\mathbf{x})\boldsymbol{\theta}, \quad (4)$$

where μ , the mean of y for given \mathbf{x} , is equal to the linear predictor η . In (1), $V(\mu) = u(\mathbf{x}) = 1$ and, in this simple case, $\phi = \sigma^2$.

Important design problems arise with extensions to this model, particularly those in which the variance is parameterized through a link function and linear predictor, that may include parameters in common with the linear predictor for the mean (Muirhead, 1982; Magnus and Neudecker, 1988; Atkinson and Cook, 1995 and Fedorov and Leonov, 2014, §6.3.1). Some references for designs related to the extension of the model to include transformation of the response and to transformation of both sides of the model are in §7.

2.3 Binomial data

For the binomial distribution, the variance function (1) is

$$V(\mu) = \mu(1 - \mu). \quad (5)$$

In models for binomial data with R successes in n observations, the response y is defined to be R/n_i . The link function should be such that, however the values of \mathbf{x} and $\boldsymbol{\theta}$ vary, the mean μ satisfies the physically meaningful constraint that $0 \leq \mu \leq 1$. We list four link functions that have been found useful in the analysis of data.

1. **Logistic:**

$$\eta = \log \left(\frac{\mu}{1 - \mu} \right). \quad (6)$$

The ratio $\mu/(1 - \mu)$ is the odds that $y = 1$. In the logistic model the log odds is therefore equal to the linear predictor. Apart from a change of sign, the model is unaltered if “success” is replaced with “failure”. For this canonical link, it follows from calculation of $d\eta/d\mu$ that

$$u(\mathbf{x}) = \mu(1 - \mu), \quad (7)$$

a simpler form than for the other three link functions we shall discuss.

2. **Probit:**

$$\eta = \Phi^{-1}(\mu), \quad (8)$$

where Φ is the normal cumulative distribution function. This link has very similar properties to those of the logistic link. In this case

$$u(\mathbf{x}) = \frac{\phi^2(\eta)}{\Phi(\eta)\{1 - \Phi(\eta)\}}. \quad (9)$$

Here ϕ is the standard normal density.

3. **Complementary log-log:**

$$\eta = \log\{-\log(1 - \mu)\}, \quad (10)$$

for which

$$u(\mathbf{x}) = \frac{(1 - \mu)}{\mu} \{\log(1 - \mu)\}^2. \quad (11)$$

The complementary log-log link is not symmetric in success and failure, so providing a model with properties distinct from those of the logistic and probit links. Interchanging success and failure gives the following log-log link.

4. **Log-log:**

$$\eta = \log(-\log \mu). \quad (12)$$

A plot of these four link functions is in Figure 4.1 of McCullagh and Nelder (1989). Atkinson and Riani (2000, §6.18) describe a fifth link, the arcsine link, which has some desirable robustness properties for binary data. In our examples we only calculate designs for the logistic link. Atkinson et al. (2007, §22.4.3) compare designs for logistic and complementary log-log links when there is a single explanatory variable.

2.4 Poisson data

For Poisson data, where $V(\mu) = \mu$, we require that $\mu > 0$. The log link, $\log \mu = \eta$, is standard for the analysis of Poisson data in areas such as medicine, social science and economics; see, for example, Agresti (2002, ch. 9), Winkelmann (2008), and von Eye et al. (2011). This link leads to models with $\mu = \exp \eta$, which satisfy the constraint on values of μ , and weights $u(\mathbf{x}) = \mu$.

2.5 Gamma data

The gamma family is one in which the correct link is often in doubt. The physical requirement is again that $\mu > 0$. A useful, flexible family of links that obeys this constraint is the Box and Cox family, in which

$$g(\mu) = \begin{cases} (\mu^\lambda - 1)/\lambda & (\lambda \neq 0) \\ \log \mu & (\lambda = 0) \end{cases}. \quad (13)$$

See Box and Cox (1964) for the use of this function in the transformation of data.

Differentiation of (13) combined with (1) shows that the weight for the gamma distribution with this link family is

$$u(\mathbf{x}) = \mu^{-2\lambda}. \quad (14)$$

When $\lambda = 0$, that is for the log link, the weights in (14) are equal to one. It therefore follows that optimal designs for gamma models with this link and $\lambda = 0$ are identical to optimal designs for regression models with the same linear predictors. Unlike designs for binomial and Poisson generalized linear models, the designs when $\lambda = 0$ do not depend on the parameter $\boldsymbol{\theta}$.

This link is seemingly equivalent to the power family of links

$$g(\mu) = \mu^\kappa, \quad (15)$$

for which differentiation shows that

$$u(\mathbf{x}) = \frac{1}{\kappa^2} \frac{1}{\mu^{2\kappa}} = \frac{1}{\kappa^2} \frac{1}{\eta^2}. \quad (16)$$

Since $1/\kappa^2$ is constant over \mathcal{X} , the optimal design depends on η , but not on κ . The situation is similar to that for normal theory regression, in which D-optimal designs are independent of the value of the variance σ^2 , although the information matrix is a function of that parameter. In §5.2.5 we discuss the relationship of the designs produced by these two links.

The gamma model is often an alternative to response transformation (§7). In particular, with a log link, it may be hard to distinguish the gamma from a linear regression model with logged response. A discussion is in §§8.1 and 8.3.4 of McCullagh and Nelder (1989) with examples of data analyses in §7.5 of Myers et al. (2010).

3 Optimal experimental designs

3.1 Theory

The first exposition of optimal design in its modern form is Kiefer (1959), although the subject goes back to Smith (1918) (see Chapter 1, especially §1.9.3). Book length treatments include Fedorov (1972), Pázman (1986) and Pukelsheim (1993). The focus of Silvey (1980) and Fedorov and Hackl (1997) is on the mathematical theory; Atkinson, Donev, and Tobias (2007) and Berger and Wong (2009) are more concerned with applications, whereas Goos and Jones (2011) introduce theory and examples via the JMP software. Pronzato and Pázman (2013) present the theory of optimal design for non-linear models, while Fedorov and Leonov (2014) illustrate their theory with pharmaceutical applications, particularly dose finding.

As in §3.1.2 of Chapter 2, where interest was in linear models, we let an experimental design ξ place a fraction w_i of the experimental trials at the conditions \mathbf{x}_i . A design with t points of support is written as

$$\xi = \left\{ \begin{array}{c} \mathbf{x}_1 \quad \mathbf{x}_2 \dots \mathbf{x}_t \\ w_1 \quad w_2 \dots w_t \end{array} \right\}, \quad (17)$$

where $w_i > 0$ and $\sum_{i=1}^t w_i = 1$. There are thus two sets of weights: w_i which determine the proportion of experimentation at \mathbf{x}_i and $u(\mathbf{x}_i)$ which, from (2), partially determine the amount of information from observations at that point. Any realisable experimental design for a total of n trials will require that the weights are ratios of integers, that is $w_i = r_i/n$, where r_i is the number of replicates at condition \mathbf{x}_i . Such designs are called *exact* and are labeled d_n . The mathematics of finding optimal experimental designs and demonstrating their properties is greatly simplified by consideration of *continuous* designs in which the integer restriction is ignored.

The resulting design is a list of conditions under which observations are to be taken. The order in which the observations are made is usually also important; for example, the order should be randomized subject to any restrictions imposed by, for example, blocking factors (§6.2).

Optimal experimental designs minimize some measure of uncertainty of the parameter estimators, typically a function of the information matrix. They require the specification of

1. A model, or set of models, of interest. For generalized linear models the specifications will include a set of parameter values $\boldsymbol{\theta} \in \Theta$, perhaps with an accompanying prior distribution $p(\boldsymbol{\theta})$.

2. A design criterion; for example, the minimization of a function of one or more information matrices.
3. A design region $\mathcal{X} \subseteq \mathbb{R}^k$ to which the \mathbf{x}_i belong.

The information matrix for the design ξ with t support points is, from (2),

$$\mathbf{M}(\xi; \boldsymbol{\theta}) = \sum_{i=1}^t w_i \mathbf{M}(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{i=1}^t w_i u(\mathbf{x}_i) \mathbf{f}(\mathbf{x}_i) \mathbf{f}'(\mathbf{x}_i). \quad (18)$$

As we are concerned with generalized *linear* models, the parameter values enter only through the GLM weights $u(\mathbf{x}_i)$.

For continuous designs we consider minimization of the general measure of imprecision $\Psi\{\mathbf{M}(\xi; \boldsymbol{\theta})\}$. Under very mild assumptions, the most important of which are the compactness of \mathcal{X} and the convexity and differentiability of Ψ , designs that minimize Ψ also satisfy a second criterion. The relationship between these two provides a ‘general equivalence theorem’, one form of which was introduced by Kiefer and Wolfowitz (1960). See §2.3 of this book for a discussion of such theorems for linear models.

Let the measure $\bar{\xi}$ put unit mass at the point \mathbf{x} and let the measure ξ^α be given by

$$\xi^\alpha = (1 - \alpha)\xi + \alpha\bar{\xi} \quad (0 \leq \alpha \leq 1). \quad (19)$$

Then, from (18),

$$\mathbf{M}(\xi^\alpha; \boldsymbol{\theta}) = (1 - \alpha)\mathbf{M}(\xi; \boldsymbol{\theta}) + \alpha\mathbf{M}(\bar{\xi}; \boldsymbol{\theta}). \quad (20)$$

Accordingly, the derivative of Ψ in the direction $\bar{\xi}$ is

$$\psi(\mathbf{x}, \xi; \boldsymbol{\theta}) = \lim_{\alpha \rightarrow 0^+} \frac{1}{\alpha} [\Psi\{(1 - \alpha)\mathbf{M}(\xi; \boldsymbol{\theta}) + \alpha\mathbf{M}(\bar{\xi}; \boldsymbol{\theta})\} - \Psi\{\mathbf{M}(\xi; \boldsymbol{\theta})\}]. \quad (21)$$

The values of $\mathbf{M}(\xi; \boldsymbol{\theta})$ and of the Fréchet derivative $\psi(\mathbf{x}, \xi; \boldsymbol{\theta})$ depend on the parameter value $\boldsymbol{\theta}$, as will the design ξ^* minimizing $\Psi\{\mathbf{M}(\xi; \boldsymbol{\theta})\}$. Variation of ξ^* with $\boldsymbol{\theta}$ is one of the themes of this chapter. In order to incorporate uncertainty in $\boldsymbol{\theta}$ we define (pseudo) Bayesian criteria and state the equivalence theorem in a form that incorporates the prior distribution $p(\boldsymbol{\theta})$ through use of

$$\Psi\{\xi\} = E_{\boldsymbol{\theta}} \Psi\{\mathbf{M}(\xi; \boldsymbol{\theta})\} = \int_{\Theta} \Psi\{\mathbf{M}(\xi; \boldsymbol{\theta})\} p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (22)$$

and

$$\psi(\mathbf{x}, \xi) = \mathbf{E}_{\boldsymbol{\theta}} \psi(\mathbf{x}, \xi; \boldsymbol{\theta}). \quad (23)$$

The General Equivalence Theorem states the equivalence of the following three conditions on ξ^* :

1. The design ξ^* minimizes $\Psi(\xi)$.
2. The design ξ^* maximizes the minimum over \mathcal{X} of $\psi(\mathbf{x}, \xi)$.
3. The minimum over \mathcal{X} of $\psi(\mathbf{x}, \xi^*)$ is equal to zero, this minimum occurring at the points of support of the design.

As a consequence of 3, we obtain the further condition:

4. For any non-optimal design ξ the minimum over \mathcal{X} of $\psi(\mathbf{x}, \xi) < 0$.

The proof of this theorem follows Whittle (1973). See Chaloner and Larntz (1989) and Woods and Lewis (2011) for general equivalence theorems developed for, and applied to, optimal designs for generalized linear models.

As we illustrate later, the theorem provides methods for the construction and checking of designs. However, it says nothing about t , the number of support points of the design. If $p(\boldsymbol{\theta})$ is a point prior, putting all mass at a single value $\boldsymbol{\theta}_0$, the designs are called *locally optimal*. A bound on t can then be obtained from the nature of $\mathbf{M}(\xi; \boldsymbol{\theta}_0)$, which is a symmetric $p \times p$ matrix. Due to the additive nature of information matrices (18), it follows from Carathéodory's Theorem that the information matrix of a design can be represented as a weighted sum of, at most, $p(p+1)/2+1$ rank-one information matrices. The maximum number of support points is therefore $p(p+1)/2+1$. A careful discussion is given by Pukelsheim (1993, §8.3), with a shorter treatment by Fedorov and Leonov (2014, §2.4.1) who state a simple and usually satisfied condition under which the maximum number of support points reduces to $p(p+1)/2$. In the examples that follow the number of support points of optimal designs is usually appreciably less than this; often as few as p is optimal. Of course, such designs with minimum support provide no means of model checking (see §7). For more general priors $p(\boldsymbol{\theta})$, the number of support points may be larger, increasing with the variance of the prior. Atkinson et al. (2007, p. 300) give examples for one-parameter nonlinear normal models in which the optimal designs have up to five support points.

3.2 D-optimal designs

The most widely used design criterion is D-optimality (see Chapter 2) in which $\Psi\{\mathbf{M}(\xi; \boldsymbol{\theta})\} = -\log |\mathbf{M}(\xi; \boldsymbol{\theta})|$, so that the log determinant of the information matrix is maximized. Then (22) becomes

$$\Psi(\xi) = -E_{\boldsymbol{\theta}} \log |\mathbf{M}(\xi; \boldsymbol{\theta})|. \quad (25)$$

This (pseudo) Bayesian D-optimality criterion has been used to find designs for generalized linear models by Chaloner and Larntz (1989) and Woods et al. (2006), amongst others. See §3.4.1 for a comment on the distinction between such criteria and truly Bayesian designs.

Fedorov and Leonov (2014, p. 68 and §10.1) provide the mathematical results, including matrix differentiation, required for calculation of the Fréchet derivative (21). The expected derivative (23) then becomes

$$\begin{aligned} \psi(\mathbf{x}, \xi) &= E_{\boldsymbol{\theta}} \{p - u(\mathbf{x}) \mathbf{f}'(\mathbf{x}) \mathbf{M}^{-1}(\xi; \boldsymbol{\theta}) \mathbf{f}(\mathbf{x})\} \\ &= p - E_{\boldsymbol{\theta}} \{u(\mathbf{x}) \mathbf{f}'(\mathbf{x}) \mathbf{M}^{-1}(\xi; \boldsymbol{\theta}) \mathbf{f}(\mathbf{x})\}. \end{aligned} \quad (26)$$

For locally D-optimal designs the number of support points t may be between p and $p(p+1)/2 + 1$. If $t = p$, the designs weights are $w_i = 1/p$.

3.3 Design efficiency

Efficiencies of designs can be compared through the values of the objective function (22). If ξ_0^* is the optimal design for the prior $p_0(\boldsymbol{\theta})$, comparison is of the values of $\Psi\{\xi_0^*\}$ and of $\Psi\{\xi\}$, where ξ is some other design to be compared and both expectations are taken over the prior $p_0(\boldsymbol{\theta})$.

There is a particularly satisfying form of efficiency for D-optimality when $p(\boldsymbol{\theta})$ is a point prior. Then from (25) the locally D-optimal design maximizes $|\mathbf{M}(\xi; \boldsymbol{\theta}_0)|$ and the efficiency of the design ξ is

$$\text{Eff}_D = \left\{ \frac{|\mathbf{M}(\xi; \boldsymbol{\theta}_0)|}{|\mathbf{M}(\xi_0^*; \boldsymbol{\theta}_0)|} \right\}^{1/p}. \quad (27)$$

Raising the ratio of determinants to the power $1/p$ results in an efficiency measure which has the dimensions of variance; a design with an efficiency of 50% requires twice as many trials as the D-optimal design to give the same precision of estimation. If only one parameter is of interest, (27) reduces to comparison of the variances of the estimated parameters under different designs (see §§9.1 and 9.2 of Chapter 1 for a discussion of efficiency for

wider classes of designs and for functions of parameter estimates including contrasts).

If the prior is not concentrated on a single point the optimal design has to be found by taking the expectation of the determinant over the prior distribution. Usually this requires a numerical approximation to the value. The efficiency in (27) in addition requires calculation of the expectation of the determinant for the design ξ . An informative alternative, which we illustrate in §5.1.6, is to look instead at the distribution of efficiencies found by simulation from the prior distribution. This procedure avoids taking expectations since we calculate (27) for each sampled value of $\boldsymbol{\theta}_0$.

3.4 Parameter uncertainty

We illustrate the dependence of the locally optimal design on the value of $\boldsymbol{\theta}$ for logistic regression with a single explanatory variable in §4.1. An example for two-variable logistic regression is in §5.1.5 and for two variable gamma regression in §5.2.3. In addition to the Bayesian design of §5.1.6, we here list some other approaches to parameter uncertainty.

3.4.1 Bayesian designs

Bayesian designs are found to maximize expectations such as (22). The ease of calculation depends on the form of the prior $p(\boldsymbol{\theta})$ and of $\Psi(\cdot)$ as well as, often crucially, on the region of integration Θ . The easiest case is when the $\boldsymbol{\theta}$ have a multivariate normal distribution over \mathbb{R}^p , although numerical approximation is still needed. Sometimes a transformation of the parameters is required to achieve this simple structure for multivariate priors. For less amenable cases, a standard solution is to sample from the prior distribution and to use an average objective function. An example for a nonlinear model is in §18.5 of Atkinson et al. (2007).

Designs maximizing expectations such as (22) ignore the additional effect of the prior information about $\boldsymbol{\theta}$ on the information matrix and make no assumption of a Bayesian analysis. The designs are accordingly sometimes called pseudo-Bayesian. A discussion of Bayesian experimental design is given by Chaloner and Verdinelli (1995).

3.4.2 Sequential designs

Where possible, sequential experimentation can provide an efficient strategy in the absence of knowledge of plausible parameter values. The usual steps are:

1. Start with some preliminary information providing an estimate, or guess, of the parameter values. This may lead to either a point prior $\boldsymbol{\theta}_0$ or a prior distribution $p(\boldsymbol{\theta})$.
2. One or a few trials of the optimal design are executed and analysed. If the new estimate of $\boldsymbol{\theta}$ is sufficiently accurate, the process stops. Otherwise, step 2 is repeated for the new estimate and the process continued until sufficient accuracy is obtained or the experimental resources are exhausted.

An early example, for nonlinear regression, is Box and Hunter (1965), extended by Atkinson et al. (2007, §17.7). Dror and Steinberg (2008) developed a Bayesian sequential design methodology for generalized linear models.

3.4.3 Minimax and maximin designs

The minimax approach overcomes dependence of designs on the unknown value of $\boldsymbol{\theta}$ by finding the best design for the worst case when the parameter $\boldsymbol{\theta}$ belongs to a set Θ . A design ξ^* is found for which

$$\Psi\{\mathbf{M}(\xi^*)\} = \min_{\xi} \max_{\boldsymbol{\theta} \in \Theta} \Psi\{\mathbf{M}(\xi; \boldsymbol{\theta})\}. \quad (28)$$

In (28) the criterion $\Psi(\cdot)$ needs to be chosen with care. Suppose interest is in D-optimality, with $\xi^*(\boldsymbol{\theta}')$ the locally D-optimal design for parameter value $\boldsymbol{\theta}'$. Unlike with linear models, the value of $|\mathbf{M}\{\xi^*(\boldsymbol{\theta}'); \boldsymbol{\theta}'\}|$ depends on the value of $\boldsymbol{\theta}'$. Accordingly, the related design efficiencies are often used as a criterion, when the maximin design ξ^* is found to maximize the minimum efficiency:

$$\text{Eff}_D(\xi^*) = \max_{\xi} \min_{\boldsymbol{\theta}' \in \Theta} \text{Eff}_D(\xi; \boldsymbol{\theta}').$$

A potential objection to these designs is that the minimax or maximin design is often close to a combination of locally optimal designs for values of $\boldsymbol{\theta}$ at the edges of the parameter space. If a prior distribution is available, such points may have a very low probability; their importance in the design criterion may therefore be over-emphasized by the minimax criterion. Providing adequate performance in these unlikely worst case scenarios may greatly affect overall design performance.

A computational difficulty is that such designs can be hard to find. Numerical procedures are described by Nyquist (2013), Fedorov and Leonov (2014, pp. 82 and p.130) and, in greatest detail, by Pronzato and Pázman (2013, §9.3). Minimax designs for generalized linear models have been found by Sitter (1992) and King and Wong (2000). A practical difficulty is that the

designs may have an appreciable number of support points, some with very low weights. Of course, approximations to the optimal design with fewer support points can always be evaluated, provided the optimal design has been found.

3.4.4 Cluster designs

Some of the computational issues associated with finding Bayesian or minimax designs can be alleviated through the use of clustering of design points, or other less formal techniques, to find designs that incorporate the overall structure of the set of locally optimal designs. Cluster designs are found by (i) sampling parameter vectors $\boldsymbol{\theta}$ from a prior distribution; (ii) finding a locally optimal design for each sampled vector; (iii) clustering the support points of these designs; (iv) forming a new, robust design having equally weighted support points that are the cluster centroids. See Dror and Steinberg (2006). Such methods are particularly effective in reducing computation when the locally optimal designs are available analytically (Russell et al., 2009).

3.5 Small effects

If the effects of the factors are slight, the mean μ_i of each observation will be similar, and so will each corresponding model weight $u(\mathbf{x}_i)$. The information matrix will then, apart from a scaling factor, be close to the unweighted information matrix $\mathbf{X}'\mathbf{X}$ and the optimal designs for the weighted and unweighted matrices will be close. Since the designs minimizing functions of $\mathbf{X}'\mathbf{X}$ are the optimal regression designs, these will be optimal, or near optimal, for generalized linear models with small effects (Cox, 1988). This is in addition to the result of §2.5 that regression designs are optimal for the gamma model with log link.

4 Locally optimal designs

4.1 Binomial data: logistic regression in one variable

The logistic model is widely used for dose-response data when the response is binomial. For example, Bliss (1935) gives data from subjecting groups of around 60 beetles to eight different levels of an insecticide. The response is the number of beetles killed. The data are reported in numerous textbooks, including Agresti (2002), Collett (2002), Dobson (2001) and Morgan (1992). The original analysis used a probit model, with more recent analyses preferring a logistic link. Atkinson and Riani (2000, §6.14) use a goodness of link

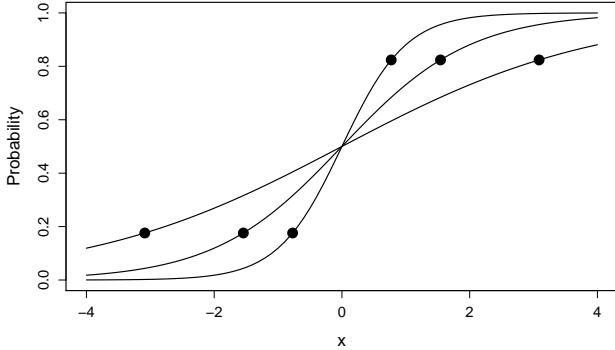


Figure 1: Logistic regression on a single explanatory variable. Response functions for $\theta_1 = 0.5, 1$ and 2 (the steepest curve). The D-optimal design points \bullet are clearly highly dependent on the value of θ_1 .

test to argue that a complementary log-log link is preferable; Atkinson et al. (2007, §22.4) present locally optimal designs for both models.

With a single explanatory variable and the logistic link

$$\log\{\mu/(1-\mu)\} = \eta = \theta_0 + \theta_1 x. \quad (29)$$

Figure 1 shows these response curves for $\theta_0 = 0$ and $\theta_1 = 0.5, 1$ and 2 . As θ_1 increases, so does the rate of increase of the response at $x = 0$.

It is clear that optimal designs for the three sets of parameter values will be rather different: experiments for values of x for which the responses are virtually all 0 or 1 will be uninformative about the values of the parameters. This intuitive result also follows from the weight $u(x)$ (7) which goes to zero for extreme values of μ . This result is also related to the phenomenon of separation in the analysis of binary data (see Firth, 1993 and Woods et al., 2006), where a hyperplane of the design space separates the observed data into zeroes and ones.

We start by finding the locally D-optimal design for the *canonical* case of $\theta = (0, 1)'$. From §3.2 we know that the D-optimal design will have either two or three support points. A standard way to proceed is to assume that there are two, when, again from §3.2, each $w_i = 0.5$. Further, with a response symmetrical about zero, the design can also be assumed symmetrical about zero. It is thus simple to calculate the D-optimal design within this class. The equivalence theorem is then used to check whether the design is indeed optimal. In this case, see Figure 2, this is the correct form and the D-optimal

Table 1: D-efficiencies (%), of designs for one-variable logistic regression as θ_1 varies with $\theta_0 = 0$. The design in column i is optimal for the parameter value in row j (100% efficiency when $i = j$).

| θ_1 | Design point x^* (31) | | |
|------------|-------------------------|--------|--------|
| | 3.086 | 1.5434 | 0.7717 |
| 0.5 | 100 | 74.52 | 41.52 |
| 1 | 57.56 | 100 | 74.52 |
| 2 | 5.72 | 57.56 | 100 |

design for a sufficiently large design region \mathcal{X} is

$$\xi^* = \begin{Bmatrix} -1.5434 & 1.5434 \\ 0.5 & 0.5 \end{Bmatrix}, \quad (30)$$

that is equal weights at two points symmetrical about $x = 0$. At these support points $\mu = 0.176$ and $1 - 0.176 = 0.824$.

Although designs for other values of the parameters can likewise be found numerically, design problems for a single x can often be solved in a canonical form, yielding a structure for the designs independent of the particular parameter values (see §5.1.1). The translation into a design for other parameter values depends, of course, on the particular $\boldsymbol{\theta}$.

For the upper design point in (30) the linear predictor $\eta = 0 + 1 \times x$ has the value 1.5434, which is the value we need for the optimal design whatever the parameterization. If we solve (29) for the η giving this value, the upper support point of the design is given by

$$x^* = (1.5434 - \theta_0)/\theta_1. \quad (31)$$

For linear regression the D-optimal design puts half of the design weight at the two extreme values of \mathcal{X} , whereas, for logistic regression, the design does not span \mathcal{X} , provided the region is sufficiently large. Note that as $\theta_1 \rightarrow 0$, the value of x^* increases without limit. This is an example of the result of Cox (1988) mentioned above that, for small effects of the variables, the design tends to that for homoscedastic regression. In practice the design region will not be unlimited and, depending on $\boldsymbol{\theta}$, the optimal design may put equal weight on one boundary point and an internal point, or on the two boundary points of \mathcal{X} .

In addition to the plots of the probability of success μ against x for three values of θ_1 , 0.5, 1 and 2, Figure 1 also shows the D-optimal design points.

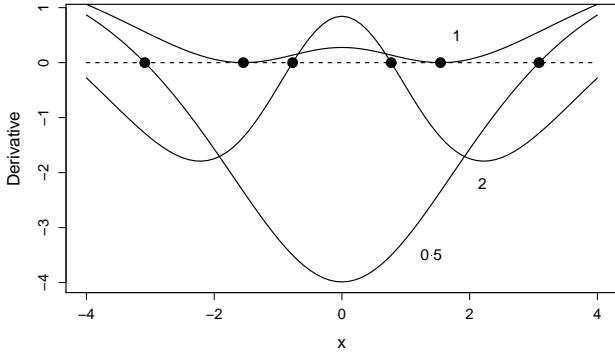


Figure 2: Logistic regression on a single explanatory variable. Equivalence theorem: derivative functions for three D-optimal designs for different values of θ_1 assessed for $\theta_1 = 1$. Curves labelled by value of θ_1 ; • design points.

For the three parameter values we obtain design points of ± 3.0863 , ± 1.5434 and ± 0.7717 . The D-efficiencies of each for the other set of parameter values are in Table 1. The most striking feature of the table is that the design at ± 3.0863 , that is for $\theta_1 = 0.5$ has an efficiency of only 5.72% when $\theta_1 = 2$. The explanation of this low efficiency is clear from Figure 1; the design point is so extreme that the value of μ when $\theta_1 = 2$ is virtually 0 or 1, so that the experiment yields little information. Clearly we need a compromise design when the parameters are not approximately known. Note that this comparison does not include the value of θ_0 , changes in which will make the design asymmetrical about zero. A design robust to the values of θ typically involves more support points than does the locally optimal design. We give an example in §5.1.6.

We conclude by showing how condition 3 of the Equivalence Theorem (24) can be used both to check the optimality of designs and to suggest improvements to non-optimal designs.

Figure 2 plots the derivative functions $\psi(x, \xi)$ (26) for the three designs, assessed for $\theta_1 = 1$. The results therefore relate to the second line of Table 1. For $\theta_1 = 1$ the minimum of the function is at zero, the minimum occurring at the points of support of the design. The design is therefore D-optimal.

For the other two designs the minima are both well below zero. The most central set of design points are for $\theta_1 = 2$. Here the minima are around ± 2 , indicating that the design points should be more extreme than they are. Likewise, for $\theta_1 = 0.5$, the design points are too extreme and the minimum

of the function at the centre of the design region indicates that the design should be shrunk. An important feature of the plot is that, for all three designs, $\psi(x, \xi) = 0$ at the points of support of the design. In order to check the optimality of designs it is necessary to search over \mathcal{X} and determine the minimum of the function, rather than just checking at the design points.

Although we have illustrated the use of $\psi(x, \xi)$ in the assessment of designs, it can also be used in their construction. Sequential addition of design points at the minimum of $\psi(x, \xi)$ leads to the D-optimal design. An example of such sequential design construction is in §11.2 of Atkinson et al. (2007). A straightforward extension is to the adaptive construction of designs mentioned in §3.4.2 where the parameters are re-estimated as observations become available and one, or a few, trials added at the point maximizing the criterion function. For D-optimality and addition of single observations, this is achieved after N trials by addition of the point minimizing $\psi(x, \xi_N)$ (Box and Hunter, 1965).

5 Optimal designs for two or more explanatory variables

5.1 Binomial data

5.1.1 Induced design region

As the information matrix (18) is of a weighted form, design for the additive linear predictor

$$\eta(\mathbf{x}) = \theta_0 + \sum_{j=1}^k \theta_j x_j, \quad (32)$$

is equivalent to (unweighted) design for the linear model

$$E(y) = \theta_0 \sqrt{u(\mathbf{x})} + \sum_{j=1}^k \theta_j \sqrt{u(\mathbf{x})} x_j, = \theta_0 z_0 + \sum_{j=1}^k \theta_j z_j. \quad (33)$$

Hence the original design region \mathcal{X} can be transformed to the induced design region \mathcal{Z} for the induced variables z_0, \dots, z_k . Clearly, \mathcal{Z} depends on both \mathcal{X} and $\boldsymbol{\theta}$.

Ford et al. (1992) used this relationship with linear model design to provide geometric insight into the structure of designs for single variable generalized linear models. With one explanatory variable, \mathcal{X} is a line segment $a \leq x \leq b$. However, because in (33) $z_0 = \sqrt{u(x)}$, \mathcal{Z} is of dimension 2.

For the models for binomial data of §2.3 and $x \in \mathbb{R}$, \mathcal{Z} is a closed curve similar to the “design locus” in Figure 2 of Box and Lucas (1959), the exact shape of which will vary with $\boldsymbol{\theta}$. The results of Ford et al. (1992) require the careful enumeration of a large number of criteria and special cases. For D-optimality they use results on the relationship between D-optimal designs and the ellipse of minimum volume centered at the origin that contains \mathcal{Z} (Silvey and Titterington, 1973; Sibson, 1974; Silvey, 1980).

Ford et al. (1992) are concerned with exponential family linear models with a single explanatory variable. Mathematical results for the linear predictor (33) with more than one explanatory variable are not generally available. An important limitation on the class of models for which results can be expected comes from the form of (33) which excludes interaction and higher-order polynomial terms. We have written $z_j = \sqrt{u(\mathbf{x})}x_j$ and so $z_k = \sqrt{u(\mathbf{x})}x_k$. But $z_j z_k \neq \sqrt{u(\mathbf{x})}x_j x_k$.

In §5.1.2 we compute locally optimal designs for binomial data with linear predictor (32) for two variables. Views of the designs in \mathcal{X} and in \mathcal{Z} are quite distinct, but both are informative about the structure of the designs. Those in \mathcal{X} relate the design points to the values of the response, whereas those in \mathcal{Z} show the design points at the extremes of the region and, for second-order models, near centres of edges and at the centre of the region. The relationship with response surface designs is clear, even if the mirror is distorted.

5.1.2 First-order models

The properties of designs for response surface models, that is with two or more continuous explanatory variables, depend much more on the experimental region than those where there is only one factor.

Although it was assumed in the previous section that the experimental region \mathcal{X} was effectively unbounded, the design was constrained by the weight u to lie in a region in which μ was not too close to zero or one. But with more than one explanatory variable, constraints on the region are necessary. For example, for the two-variable first-order model

$$\log\{\mu/(1-\mu)\} = \eta = \theta_0 + \theta_1 x_1 + \theta_2 x_2, \quad (34)$$

with $\boldsymbol{\theta}' = (0, \gamma, \gamma)$, all design points for which $x_1 + x_2 = 0$ yield a value of 0.5 for μ , however extreme the values of x .

We now explore designs for the linear predictor (34) with the logistic link for a variety of parameter values and $\mathcal{X} = [-1, 1]^2$. These and succeeding designs were found by a numerical search with a quasi-Newton algorithm. The constraints to ensure that $\mathbf{x} \in \mathcal{X}$ and on the design weights were enforced using the trigonometric transformations listed in Atkinson et al. (2007, §9.5).

Table 2: Sets of parameter values for first-order linear predictors in two variables with the logistic link.

| Set | θ_0 | θ_1 | θ_2 |
|-----|------------|------------|------------|
| B1 | 0 | 1 | 1 |
| B2 | 0 | 2 | 2 |
| B3 | 2 | 2 | 2 |
| B4 | 2.5 | 2 | 2 |

Table 3: D-optimal designs for logistic models with the sets of parameter values B1 and B2 of Table 2; w_i design weights

| i | Design for B1 | | | | Design for B2 | | | |
|-----|---------------|----------|-------|---------|---------------|----------|-------------|-------------|
| | x_{1i} | x_{2i} | w_i | μ_i | x_{1i} | x_{2i} | $w_i^{(1)}$ | $w_i^{(2)}$ |
| 1 | -1 | -1 | 0.204 | 0.119 | 0.1178 | -1.0000 | | 0.240 |
| 2 | | | | | 1.0000 | -0.1178 | | 0.240 |
| 3 | 1 | -1 | 0.296 | 0.500 | 1.0000 | -1.0000 | 0.327 | 0.193 |
| 4 | -1 | 1 | 0.296 | 0.500 | -1.0000 | 1.0000 | 0.193 | 0.327 |
| 5 | | | | | -1.0000 | 0.1178 | 0.240 | |
| 6 | 1 | 1 | 0.204 | 0.881 | -0.1178 | 1.0000 | 0.240 | 0.854 |

Four sets of parameter values are given in Table 2. D-optimal designs for the sets B1 and B2 are listed in Table 3. The parameter values of B1 (0, 1, 1) are closest to zero. The table shows that the design has support at the points of the 2^2 factorial, although the design weights are not quite equal. They are so for the normal theory model and become so for the logistic model as θ_1 and $\theta_2 \rightarrow 0$ with $\theta_0 > 0$. At those factorial points for which $x_1 + x_2 = 0$, $\mu = 0.5$ since $\theta_1 = \theta_2$. At the other design points $\mu = 0.119$ and 0.881, slightly more extreme values than the values of 0.176 and 0.824 for the experiment with a single variable.

An interesting feature of our example is that the number of support points of the design depends upon the values of the parameter $\boldsymbol{\theta}$. From Carathéodory's Theorem discussed in §3.1, the maximum number of support points required by an optimal design is usually $p(p+1)/2$. Our second set of parameters, B2 in which $\boldsymbol{\theta}' = (0, 2, 2)$, gives two four-point optimal designs,

Table 4: D-optimal designs for logistic models with the parameter values B3 and B4 of Table 2; w_i design weights

| i | Parameter set B3 | | | | Parameter set B4 | | | |
|-----|------------------|----------|-------|---------|------------------|----------|-------|---------|
| | x_{1i} | x_{2i} | w_i | μ_i | x_{1i} | x_{2i} | w_i | μ_i |
| 1 | -1.0000 | -0.7370 | 0.169 | 0.186 | -1.0000 | 0.5309 | 0.333 | 0.827 |
| 2 | -1.0000 | 0.7370 | 0.331 | 0.814 | -1.0000 | -1.0000 | 0.333 | 0.182 |
| 3 | -0.7370 | -1.0000 | 0.169 | 0.186 | 0.5309 | -1.0000 | 0.333 | 0.827 |
| 4 | 0.7370 | -1.0000 | 0.331 | 0.814 | | | | |

with weights given by $w_i^{(1)}$ and $w_i^{(2)}$ in Table 3 and support points where $\mu = 0.146$, 0.5 and 0.854 . Any convex combination of these two designs, $\alpha w_i^{(1)} + (1 - \alpha)w_i^{(2)}$ with $0 \leq \alpha \leq 1$, will also be optimal, and will have six support points, which is the value of the Carathéodory bound when $p = 3$. These two component designs arise from the symmetry of the design problem; not only does $\theta_1 = \theta_2$ but also the design region is symmetrical in x_1 and x_2 .

The D-optimal designs for the two remaining sets of parameters in Table 2 are given in Table 4. These designs have respectively 4 and 3 points of support. When $\boldsymbol{\theta}' = (2, 2, 2)$, the design points are where $\mu = 0.186$ and 0.814 . For $\boldsymbol{\theta}' = (2.5, 2, 2)$ the minimum value of μ , 0.182 at $(-1, -1)$, is sufficiently high that there are no experimental conditions for which $\mu = 0.15$. The experimental values of μ are 0.182 and 0.827 . For this three-point design for a three parameter model, the design weights $w_i = 1/3$.

The relationship between the design points and the values of μ are shown, for parameter values B2 and B3, in Figure 3. For $\boldsymbol{\theta}' = (0, 2, 2)$, one four-point design for B2 is depicted by open circles, the other by filled circles; the symmetry of the designs is evident. For $\boldsymbol{\theta}' = (2, 2, 2)$ there are again four support points of the design, which now lie somewhat away from the boundaries of the regions of high and low values of μ .

5.1.3 Induced design region for the logistic link

For the first-order model in $k = 2$ factors, with $p = 3$, the induced design space \mathcal{Z} is of dimension three. Two examples, projected onto z_1 and z_2 and thus ignoring $z_0 = \sqrt{w}$, are given in Figure 4 for \mathcal{X} the unit square. In the left-hand panel of the figure, $\boldsymbol{\theta}' = (0, 2, 2)$ so that at the corner of \mathcal{X}

for which $x_1 = x_2 = 1$, $\eta = 4$ and $\mu = 0.982$. This is well beyond the range for informative experiments and the projection of the induced design space appears to be folded over. As a consequence, experiments at extreme positions in \mathcal{Z} are not at extreme points in \mathcal{X} . The results in the other panel for $\boldsymbol{\theta}' = (2, 2, 2)$ are similar, but more extreme. For both sets of parameter values the design points lie, as they should, on the boundary of the induced design region.

These examples show the importance of both the design region and the value of μ in determining the optimal design. In order to reveal the structure of the designs as clearly as possible, the designs considered have all had $\theta_1 = \theta_2$, and so are symmetrical in x_1 and x_2 . When $\theta_1 \neq \theta_2$, both the design region and the values of μ are important in the resulting asymmetrical designs. Asymmetrical designs also arise when the log-log and complementary log-log links are used, since these links are not symmetrical functions.

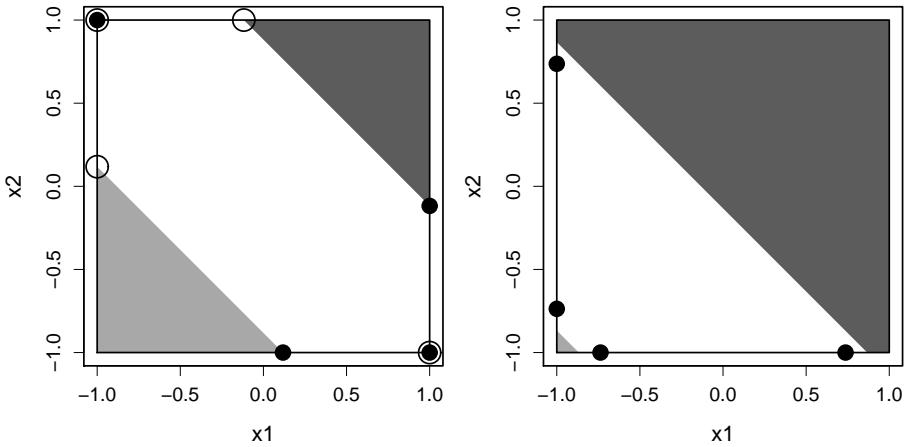


Figure 3: Support points for D-optimal designs for logistic models with parameter values B2 and B3 in Table 2. In the lightly shaded area $\mu \leq 0.15$, whereas, in the darker region, $\mu \geq 0.85$. In the left-hand panel, one four-point design for B2 is depicted by open circles, the other by filled circles. The symmetry of the designs is evident.

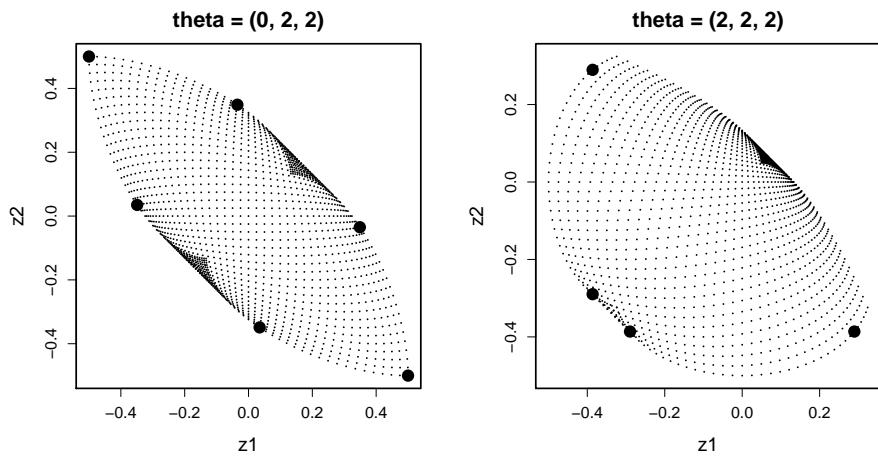


Figure 4: Support points for D-optimal designs for logistic models with parameter values B_2 (left) and B_3 (right) of Table 2 in the induced design region \mathcal{Z} . For these first-order models all design points lie on the boundary of \mathcal{Z} .

5.1.4 Theoretical results for a first-order predictor

Various authors have derived optimal designs for $k > 1$ for first-order linear predictor (32) for some special cases; see Sitter and Torsney (1995a), Sitter and Torsney (1995b) and Torsney and Gunduz (2001). Yang, Zhang, and Huang (2011) considered the case when the assumption of a bounded design region \mathcal{X} is relaxed by allowing one variable, say x_k , to take any value in \mathfrak{R} . As the D-optimality objective function is still bounded in this case, the authors were able to provide a methodology to obtain a locally D-optimal design analytically. We restate their corollary 1 for the special case of logistic regression and D-optimality.

Theorem 1: (Yang et al., 2011) *For the logistic model with linear predictor (32) and $\mathcal{X} = [-1, 1]^{k-1} \times \mathfrak{R}$, a D-optimal design has 2^k equally weighted support points,*

$$\mathbf{x}_l^* = \begin{cases} (x_{1l}, \dots, x_{(k-1)l}, a_l^*) & \text{for } l = 1, \dots, 2^{k-1} \\ (x_{1l}, \dots, x_{(k-1)l}, -a_l^*) & \text{for } l = 2^{k-1} + 1, \dots, 2^k, \end{cases}$$

where

$$x_{jl} = \begin{cases} -1 & \text{if } \lceil \frac{l}{2^{k-1-j}} \rceil \text{ is odd} \\ 1 & \text{if } \lceil \frac{l}{2^{k-1-j}} \rceil \text{ is even} \end{cases} \quad j = 1, \dots, k-1,$$

$\lceil a \rceil$ is the smallest integer greater than or equal to a , the numerator of the fractions is l and $a_l^* = \eta^* - \eta_l^*(k)$. Here, η^* maximises

$$\eta^2 \left\{ \frac{dh(\eta)/d\eta}{h(\eta)[1-h(\eta)]} \right\}^{k+1},$$

where $h = g^{-1}$, the inverse of the logistic link function, and $\eta_l^*(k) = \theta_0 + \sum_{j=1}^{k-1} \theta_j x_{jl}$.

This result has a fairly straightforward interpretation. If we fix the values of $k-1$ variables in the support points at the level combinations of a 2^{k-1} factorial design, then the selection of an optimal design reduces to a one variable problem (the choice of the values to be taken by x_k). Note that \mathcal{X} is such that each variable lies in an interval, rather than just taking the two values ± 1 .

To illustrate this result, Table 5 gives D-optimal designs on $[-1, 1] \times \mathfrak{R}$ for the sets of parameter values in Table 2. These designs are quite different from the designs in Tables 3 and 4, where the restriction of x_2 to the interval $[-1, 1]$ results in the designs having different numbers of support points, different values for the supports points and different weights.

Table 5: D-optimal support points from Theorem 1 (Yang et al., 2011) for parameter values in Table 2. For each design, the support points are equally weighted.

| Parameters | Support points | |
|------------|----------------|---------|
| | x_1 | x_2 |
| B1 | -1 | -2.2229 |
| | -1 | 2.2229 |
| | +1 | -0.2229 |
| | +1 | 0.2229 |
| B2 | -1 | 1.6115 |
| | -1 | -1.6115 |
| | +1 | -0.3886 |
| | +1 | 0.3886 |
| B3 | -1 | 0.6115 |
| | -1 | -0.6115 |
| | +1 | -1.3886 |
| | +1 | 1.3886 |
| B4 | -1 | -0.3615 |
| | -1 | 0.3615 |
| | +1 | -1.6386 |
| | +1 | 1.6386 |

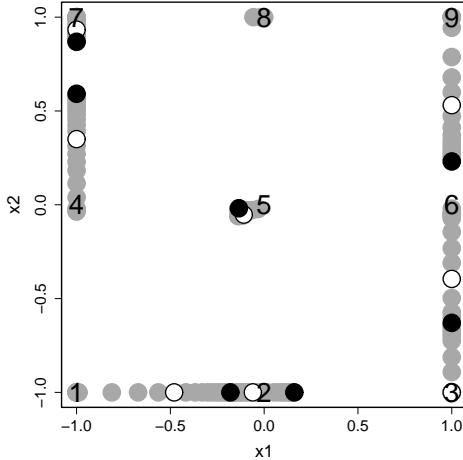


Figure 5: D-optimal designs for second-order logistic model as γ varies ($0 \leq \gamma \leq 2$). Support points: numbers, $\gamma = 0$; circles, $\gamma = 1$; black dots $\gamma = 2$; and grey dots, intermediate values.

5.1.5 A second-order response surface

This section extends the results of earlier sections to the second-order response surface model, again with two factors and again with the logistic link. The purpose is to show the relationship with, and differences from, designs for regression models. The D-optimal designs are found, as before, by minimizing $-\log |\mathbf{M}(\xi; \boldsymbol{\theta})|$. With $n = 9$ and two-dimensional x , the numerical search is in 26 dimensions. However, once the structure of the designs is established, with 8 of the design points on the edges of \mathcal{X} (see Figure 5) the search can be reduced to 18 dimensions, with the equivalence theorem providing a check on this reduction.

To explore how the design changes with the parameters of the model we look at a series of designs for the family of linear predictors

$$\eta = \theta_0 + \gamma(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2 + \theta_{11} x_1^2 + \theta_{22} x_2^2) \quad \text{with } \gamma \geq 0, \quad (35)$$

and design region $\mathcal{X} = [-1, 1]^2$. The parameter γ provides a family of similarly shaped linear predictors which increasingly depart, in a proportional way, from a constant value as γ increases. When $\gamma = 0$ the result of Cox (1988) shows that the design is the D-optimal design for the second-order regression model, the unequally weighted 3^2 factorial design given in §5.2.3.

For numerical exploration we take $\theta_0 = 1$, $\theta_1 = 2$, $\theta_2 = 2$, $\theta_{12} = -1$, $\theta_{11} = -1.5$ and $\theta_{22} = 1.5$. As γ varies from 0 to 2, the shape of the response

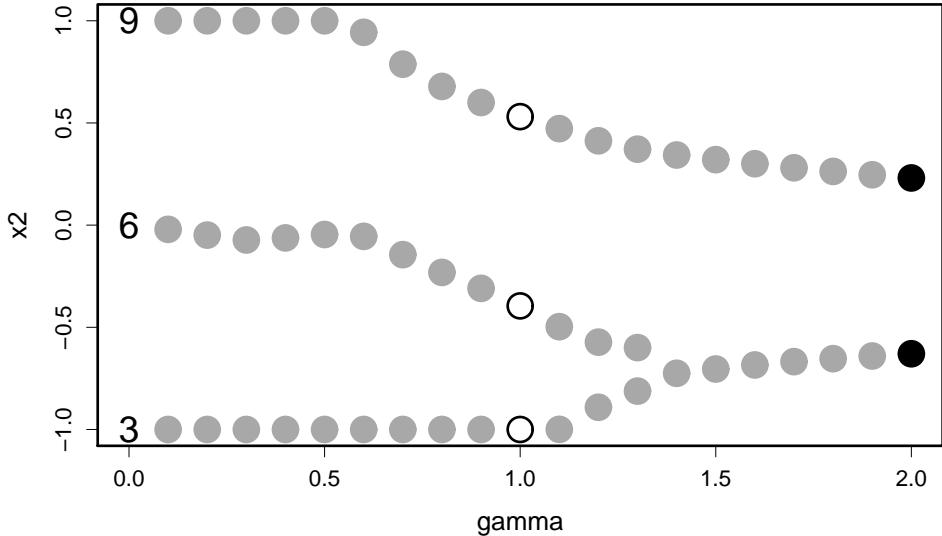


Figure 6: D-optimal designs for second-order logistic model. The values of x_2 for support points 3, 6 and 9 of Figure 5 as γ varies between zero and two. The coding of the symbols is the same as in the earlier figure.

surface becomes increasingly complicated.

Figure 5 shows the support points of the D-optimal designs as γ increases from zero in steps of 0.1. The design points are labelled, for $\gamma = 0$, in standard order for the 3^2 factorial, with x_1 changing more frequently. The figure shows how all but one of the design points stay on the boundary of the design region; the circles and black dots are the support points for $\gamma = 1$ and 2, respectively, with the grey dots indicating intermediate values. There is little change in the location of the centre point, point 5, during these changes. Initially the design has nine points, but the weight on point 8 decreases to zero when $\gamma = 0.3$. Thereafter, the design has eight support points until $\gamma = 1.4$ when the weight on observation 6 becomes zero.

Figure 6 serves to help interpret the behaviour of the design as γ increases, showing the values of x_2 for the three design points (3, 6 and 9) in Figure 5 for which $x_1 = 1$. Initially the values of x_2 are those for the 3^2 factorial and they remain virtually so until $\gamma = 0.6$. Thereafter, they gradually converge towards three more central values.

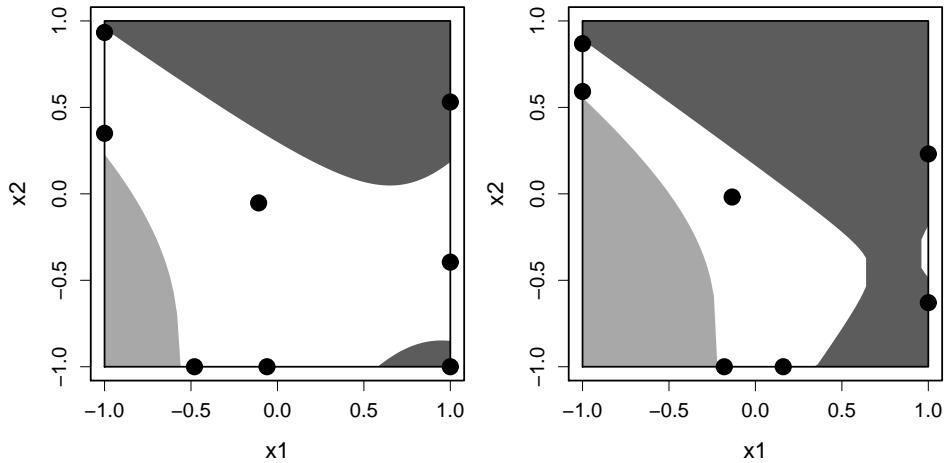


Figure 7: Support points for D-optimal designs for the second-order logistic model. Left-hand panel $\gamma = 1$, right-hand panel $\gamma = 2$. In the lightly shaded area $\mu \leq 0.15$, whereas, in the darker region, $\mu \geq 0.85$.

The relationship between the support points of the design and the values of μ is highlighted in Figure 7 where, as in Figure 3, the pale areas are regions in which $\mu \leq 0.15$, with the dark regions the complementary ones where $\mu \geq 0.85$. The left-hand panel of Figure 7, for $\gamma = 1$, shows that the 8-point design is a distortion of a standard response surface design, with most points in the white area and the remainder on the boundary of the design region, some close to the contours of $\mu = 0.15$ or 0.85 . In the numbering of Figure 5, points 2 and 6 are on the edge of the design region where μ is close to 0.5. Points 3 and 9 are at higher values of μ .

A similar pattern is clear in the seven-point design for $\gamma = 2$ in the right-hand panel of the figure; four of the seven points are on the edges of the white region, one is in the centre and only points 3 and 9 are at more extreme values of μ .

The two panels of Figure 7 taken together explain the trajectories of the points in Figure 5 as γ varies. For example, points 1 and 4 move away from $(-1, -1)$ as the value of μ at that point decreases, point 3 remains at $(1, -1)$

until γ is close to one and point 8 is rapidly eliminated from the design as the value of μ near $(0, 1)$ increases with γ .

Further insight into the structure of the designs can be obtained from consideration of the induced design region introduced in §5.1.3. Although, as stated earlier, extension of the procedure based on (33) to second-order models such as (35) is not obvious, it is still informative to look at the plot the designs in \mathcal{Z} space. The left-hand panel of Figure 22.8 of Atkinson et al. (2007) shows the eight-point design for $\gamma = 1$ plotted against z_1 and z_2 ; seven points lie on the edge of this region, well spaced and far from the centre, which is where the eighth point is. The right-hand panel for $\gamma = 2$ shows six points similarly on the edge of \mathcal{Z} ; the centre point is hidden under the seemingly folded-over region near the origin.

In the induced design region these designs are reminiscent of response surface designs, with a support point at the centre of the region and others at remote points. However the form of \mathcal{Z} depends on the unknown parameters of the linear predictor, so this description is not helpful in constructing designs. In the original space \mathcal{X} we have described the designs for this second-order model as a series of progressive distortions of designs with support at the points of the 3^2 factorial. For small values of γ the unweighted 3^2 factorial provides an efficient design, with a D-efficiency of 97.4% when $\gamma = 0$. However, the efficiency of this design declines steadily with γ , being 74.2% for $\gamma = 1$ and a low 38.0% when $\gamma = 2$. If appreciable effects of the factors are expected, the special experimental design methods of this section need to be used. Further discussion of designs for two variable logistic models is given by Sitter and Torsney (1995a) with particular emphasis on the structure of \mathcal{Z} .

5.1.6 Bayesian D-optimal designs

We can also find Bayesian D-optimal designs, maximizing (25), for response surface designs and binomial data. Motivated by a food technology example, Woods et al. (2006) found designs for logistic regression with linear predictor

$$\eta(\mathbf{x}) = \theta_0 + \sum_{i=1}^3 \theta_i x_i + \sum_{i=1}^3 \sum_{j \geq i}^3 \theta_{ij} x_i x_j,$$

with $x_i \in [-1.2782, 1.2782]$. Here we find a Bayesian D-optimal design assuming independent uniform prior densities for the parameters in $\boldsymbol{\theta}$, defined on the support

$$\theta_1, \theta_2 \in [2, 6], \quad \theta_0, \theta_3, \theta_{ij} \in [-2, 2] \text{ for } i, j = 1, 2, 3.$$

We approximate the expectation (25) by the sample average across a 20 run Latin Hypercube sample (see, for example, Chapter 19 and Santner et al., 2003, Ch. 6). Gotwalt et al. (2009) and Woods and van de Ven (2011) discuss and compare some alternative approaches for this approximation.

For this example, a simulated annealing algorithm (Woods, 2010) was used to find an exact design, d_{16} , with $n = 16$ points. The design is given in Figure 8 and, in fact, has $t = n = 16$ support points and no replication. For reference, Figure 8 also gives the design points of a 16 run central composite design (CCD, see Chapter 10), d_{ccd} , with 8 factorial points, 6 axial points with $x_j = \pm 1.2782$ and two centre points; see, for example, Box and Draper (2007). This design is standard for normal theory response surface studies and is an obvious comparator for the Bayesian GLM design. In fact, a CCD had been employed for the food technology example in previous experimentation.

While there are some familiar features to the Bayesian D-optimal design, including (near) centre points, there are also some distinct differences from the CCD. These include the presence of extreme corner points in the D-optimal design and, for x_1 and x_2 , fewer distinct levels (~ 3 for each of these two variables).

The relative performance of the designs was assessed via a simulation study. A sample of 1000 parameter vectors, $\boldsymbol{\theta}^{(l)}$, was drawn from the prior distribution for $\boldsymbol{\theta}$. For each vector, we calculated the relative D-efficiency

$$\text{Eff}_D(l) = \left\{ \frac{|\mathbf{M}(d_{16}); \boldsymbol{\theta}^{(l)}|}{|\mathbf{M}(d_{ccd}); \boldsymbol{\theta}^{(l)}|} \right\}^{\frac{1}{10}}, \quad l = 1, \dots, 1000. \quad (36)$$

The empirical cumulative distribution of the relative efficiency, Figure 9, shows a dramatic difference in performance between the two designs. The Bayesian D-optimal design is more efficient than the CCD for about 85% of the sampled parameter vectors, and has efficiency of 1.75 or more for 50% of the sampled vectors. The maximum relative efficiency is close to 7.

5.2 Gamma data

5.2.1 Theoretical results for first-order designs with the power link

We use numerical examples to illustrate the great difference between designs for gamma models and those for binomial responses of §5.1.2. The examples are calculated using the power link. In this introductory section we present the theoretical results of Burridge and Sebastiani (1994) for first-order mod-

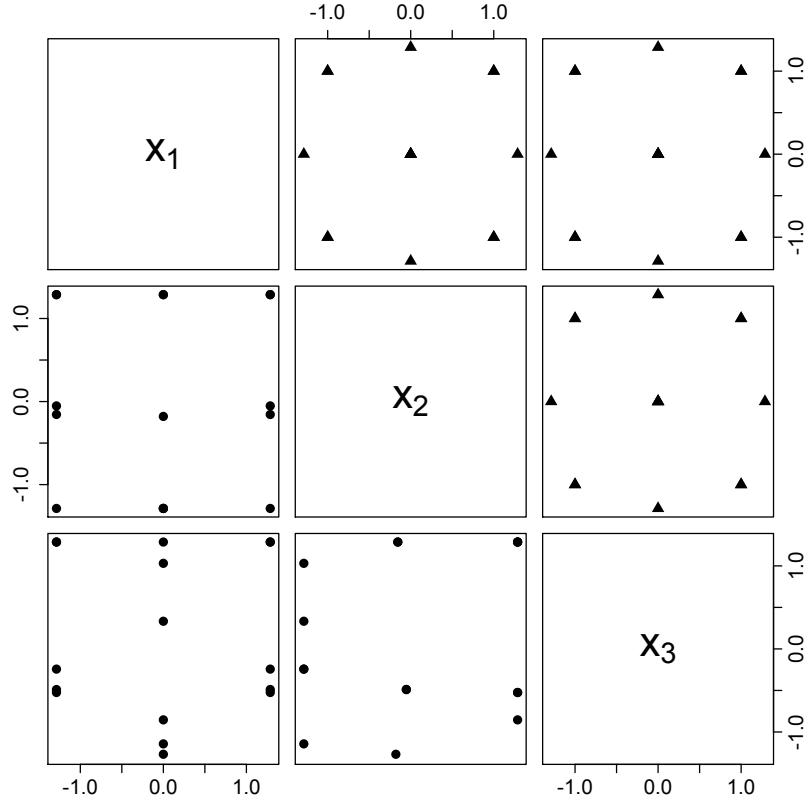


Figure 8: Two-dimensional projections of the Bayesian D-optimal design for a logistic regression model (lower diagonal) and a central composite design (upper diagonal).

els. Our numerical example of a first-order model is in §5.2.2 with an example for a response-surface model in §5.2.3. We conclude with a brief comparison of designs for the power link with those for the Box-Cox link.

If the design region for uncoded variables is of the form $a_j \leq x_j^u \leq b_j$, the region can be coded so that $0 \leq x_j \leq 1$ for $j = 1, \dots, k$. The requirement that $\mu > 0$ for all non-negative x leads to a canonical form of the original problem with $\theta_j \geq 0$ for all $j = 0, \dots, k$, with at least one inequality. Since the weights (16) are monotonic in η , the support points of D-optimal designs as $\boldsymbol{\theta}$ varies must then be some of the points of the 2^k factorial. Which points have non-zero weights depends on the values of the θ_j . For effects large relative to θ_0 , Burridge and Sebastiani (1994) provide the following theorem.

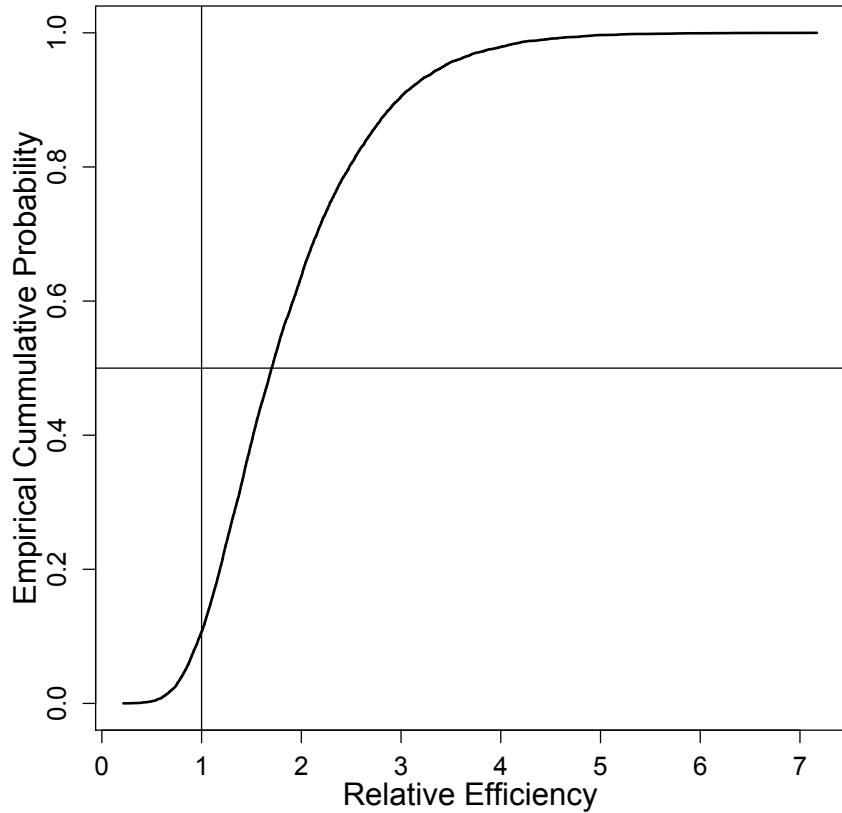


Figure 9: Empirical cumulative distribution function of the relative D-efficiency of the Bayesian D-optimal design compared to the central composite design.

Theorem 2. (Burridge and Sebastiani, 1994) *For the coded design variables x_j , the design which puts weights $1/(k+1)$ at each of the $k+1$ points*

$$(0, \dots, 0)', (1, 0, \dots, 0)', (0, 1, 0, \dots, 0)', \dots, (0, \dots, 0, 1)'$$

is D-optimal for gamma regression, the power link and a first-order linear predictor if, and only if, for all $i, j = 1, \dots, k$,

$$\theta_0^2 \leq \theta_i \theta_j.$$

Thus, for ‘large’ effects a ‘one-factor-at-a-time’ approach is optimal. However, as the effects become smaller, the design approaches the 2^k factorial in line with the result of Cox (1988) discussed in §3.5. Of course, the weights

Table 6: D-optimal designs for two-variable first-order model for gamma responses with the power link; $\boldsymbol{\theta} = (1, \chi, \chi)'$.

| | | Design weights w_i | | |
|-------|-------|----------------------|--------------|------------|
| x_1 | x_2 | $\chi = 0.1$ | $\chi = 0.5$ | $\chi = 1$ |
| 0 | 0 | 0.271 | 0.313 | 1/3 |
| 0 | 1 | 0.252 | 0.281 | 1/3 |
| 1 | 0 | 0.252 | 0.281 | 1/3 |
| 1 | 1 | 0.225 | 0.125 | 0 |

w_i have to be found numerically. However, the numerical search is greatly simplified by being restricted to the support of the 2^k factorial. It is also a great simplification that, as we showed in §2.5, the designs do not depend on the value of κ .

5.2.2 Examples of first-order designs with the power link

In both examples there are two explanatory variables. We only look at symmetrical designs generated with $\theta_1 = \theta_2$ having the three values 0.1, 0.5 and 1. In all calculations $\theta_0 = 1$. The resulting designs are in Table 6.

These results nicely illustrate the theoretical results of §5.2.1. We have parameterized the problem with $\theta_1 = \theta_2 = \chi$. For $\chi = 0.1$, that is with small effects, the design is virtually the 2^2 factorial, with weights close to 1/4 ranging from 0.225 to 0.271. Increasing χ to 0.5 leaves the support points unchanged, but now the weights range, symmetrically of course in x_1 and x_2 , from 0.125 to 0.3125, with the lowest weight on (1,1). When $\chi = 1$ we have $\theta_0^2 = \theta_1 \theta_2$, so that we are at the lowest value of χ for which we obtain a design with three support points. All weights are, of course, equal.

These designs were found numerically using a Quasi-Newton algorithm combined with the transformations given in §9.5 of Atkinson et al. (2007). The general equivalence theorem was used to check the designs by evaluation of the derivative function $\psi(\mathbf{x}, \xi)$ over a fine grid in \mathcal{X} . One point in the construction of these designs is that with $\chi = 1$ the optimization algorithm had not quite converged to the theoretical value after the default limit of 100 iterations, whereas around 10 iterations were needed for the other values of χ . The effect on the minimum value of $\psi(\mathbf{x}, \xi)$ was negligible. A second point is that, to five significant values, the weights for $\chi = 0.5$ were exactly 5/16, 9/32, 9/32 and 1/8. Such simple weights can be an indication that theoretical results are possible. See Atkinson (2010) and Dette et al. (2012)

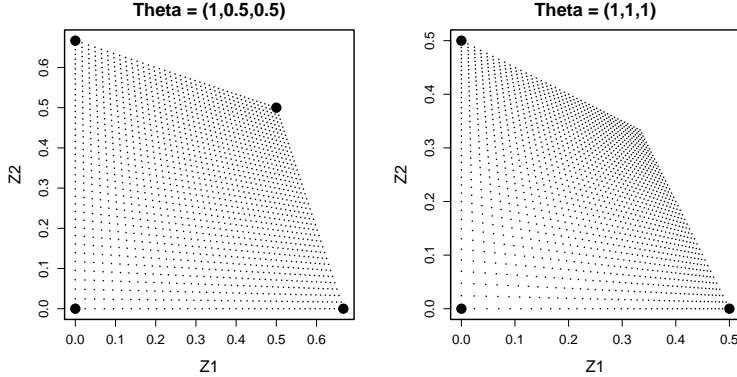


Figure 10: Support points in the induced design region \mathcal{Z} of D-optimal designs for two-variable first-order model for gamma responses with the power link; \bullet design points

for an example in discrimination between polynomial regression models.

To conclude our discussion of this example, we look at the plots of the design in the induced design region \mathcal{Z} defined in (33). As the results of Burridge and Sebastiani (1994) show, the boundary of \mathcal{Z} is formed, for $k = 2$, by straight line segments, as is shown in Figure 10. There is none of the curving over of the space of \mathcal{X} that is caused by the nonlinear nature of the GLM weights $u(\mathbf{x}_i)$ for the binomial distribution that is evident in Figure 4.

For small χ the induced design region is virtually square, becoming less so as χ increases. The left-hand panel of Figure 10 is for $\chi = 0.5$, for which the weight at $\mathbf{x} = (1, 1)'$ is $1/8$. As χ increases, the weight on this value decreases. Insight about the case $\chi = 1$ comes from the results of Silvey and Titterington (1973) relating D-optimality to minimum volume ellipsoids enclosing design regions. It follows that when $\chi = 1$ the values of z_1 and z_2 at $\mathbf{x} = (1, 1)'$, say $z_1(1, 1)$ and $z_2(1, 1)$, must be the same distance from the origin as $z_1(1, 0)$ and $z_2(1, 0)$ (or $z_1(0, 1)$ and $z_2(0, 1)$). Hence $z_1(1, 1) = z_2(1, 1) = \sqrt{2}/4$. For larger values of χ these values lie inside the circle and \mathcal{Z} becomes increasingly triangular. The design does not change as χ increases above 1.

In these calculations we have taken $\kappa = 1$. From the form of $u(\mathbf{x})$ in (16), other values of κ lead to identical figures, but with different numerical values on the axes.

Table 7: Parameter values for linear predictor (37) with $\kappa = 0.5$

| Parameter set | θ_0 | θ_1 | θ_2 | θ_{11} | θ_{22} | θ_{12} |
|---------------|------------|------------|------------|---------------|---------------|---------------|
| G1 | 3.7 | -0.46 | -0.65 | -0.19 | -0.45 | -0.57 |
| G2 | 3.7 | -0.23 | -0.325 | -0.095 | -0.225 | -0.285 |

5.2.3 Second-order response surface with the power link

Atkinson and Riani (2000, §6.9) use a gamma model to analyse data from Nelson (1981) on the degradation of insulation due to elevated temperature at a series of times. A second-order model is required in the two continuous variables and a gamma model fits well with a power link with $\kappa = 0.5$. We scale the variables to obtain design region $\mathcal{X} = [-1, 1]^2$. The linear predictor is the quadratic

$$\eta = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_{11} x_1^2 + \theta_{22} x_2^2 + \theta_{12} x_1 x_2, \quad (37)$$

that is, (35) with $\gamma = 1$. Then the standard D-optimal design for the normal theory regression model has unequally weighted support at the points of the 3^2 factorial: weight 0.1458 at the four corners of the design region, 0.0802 at the centre points of the sides and 0.0960 at the centre of the region. This design is optimal for the gamma model with log link and for the model with the power link as the parameters in (37), apart from θ_0 , become small. We take $\boldsymbol{\theta}$ to have the values given in Table 7, G1 being rounded from an analysis of Nelson's data.

The exact optimal 9-trial design for G1, found by searching over a grid of candidates with steps of 0.01 in x_1 and x_2 , is in Table 8. This shows that, at the points of the design, the minimum value of μ is 1.90 and the maximum 14.59. The parameter values are thus such that we satisfy the requirement $\mu > 0$.

As the left-hand half of Table 8 shows, the design has seven support points. The points of the 2^2 factorial are in the upper part of the table. All are included in the design, two being replicated. The other three support points are slight distortions of some remaining points of the support of the 3^2 factorial. Figure 11 makes clear the virtually symmetrical nature of the design, although the parameters are not quite symmetrical in value for x_1 and x_2 .

Table 8: Exact D-optimal designs for the parameter sets G1 and G2 of Table 7

| i | Design for G1 | | | | Design for G2 | | | |
|-----|---------------|----------|-------|---------|---------------|----------|-------|---------|
| | x_{1i} | x_{2i} | n_i | μ_i | x_{1i} | x_{2i} | n_i | μ_i |
| 1 | -1.00 | -1.00 | 1 | 12.96 | -1.00 | -1.00 | 1 | 13.32 |
| 2 | -1.00 | 1.00 | 2 | 11.83 | -1.00 | 1.00 | 1 | 12.74 |
| 3 | 1.00 | -1.00 | 2 | 14.59 | 1.00 | -1.00 | 1 | 14.14 |
| 4 | 1.00 | 1.00 | 1 | 1.90 | 1.00 | 1.00 | 1 | 6.45 |
| 5 | 0.11 | 0.15 | 1 | 12.46 | -1.00 | 0.00 | 1 | 14.71 |
| 6 | 0.26 | 1.00 | 1 | 5.38 | -0.01 | -1.00 | 1 | 14.44 |
| 7 | 1.00 | 0.29 | 1 | 7.07 | 0.07 | 0.09 | 1 | 13.33 |
| 8 | | | | | 0.08 | 1.00 | 1 | 9.66 |
| 9 | | | | | 1.00 | 0.09 | 1 | 11.01 |

To illustrate the approach of the design to the 3^2 factorial as the parameter values decrease we also found the D-optimal 9-point design for the set of parameter values G2 in Table 7 in which all parameters, other than θ_0 , have half the values they have for design G1. As Table 8 shows, the range of means at the design points is now 6.45 to 14.71, an appreciable reduction in the ratio of largest to smallest response. The support points of the design for G2 are shown in Figure 11 by the symbol X. There are now nine distinct support points close to those of the 3^2 factorial. For G2 the three design points in the lower half of the Table for G1 are moved in the direction of the full factorial design. For linear regression the unweighted 3^2 factorial is D-optimal.

5.2.4 Efficient standard designs for gamma models

We conclude our analysis of designs for the gamma model with second-order predictor and a power link by briefly exploring how well the unweighted 3^2 factorial performs when there are nine trials by comparing both it and the design for G2 with that for G1.

The D-optimal design for the less extreme parameter set G2 of Table 7 has efficiency 97.32%, while the equi-replicated 3^2 factorial has efficiency 96.35%. The main feature of these designs is how efficient they are for the gamma model, both efficiencies being greater than 95%. The design for parameters

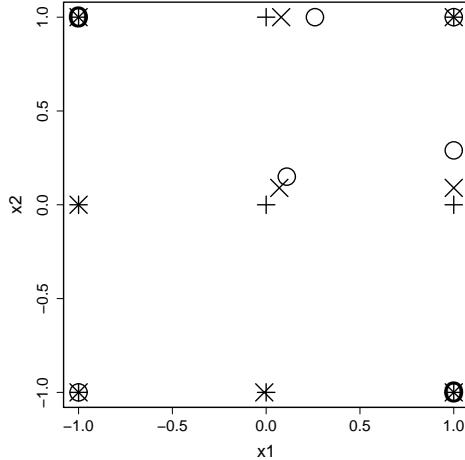


Figure 11: Points for D-optimal 9-point designs for gamma models in Table 8: +, the points of the 3^2 factorial; o, G1 and x, G2. Points for G1 which are replicated twice are darker.

G2 is for a model with smaller effects than G1, so that the design and its efficiency are between those for G1 and the factorial design.

An indication of this example with a gamma response is that standard designs may be satisfactory for second-order response surfaces. However, Burridge and Sebastiani (1994) show that, for first-order models, full 2^2 factorial designs, or their regular fractions (see Chapter 7), can be very inefficient when the effects are strong and the optimal designs have only $k + 1$ points of support.

5.2.5 Designs for the power link and for the Box-Cox link

We now briefly explore the relationship between designs for the power link with weights given by (16) and those for the Box-Cox link (14). We relate the two through their dependence on the linear predictor η .

From (16) the weights for the power link can be written

$$\{u(\mathbf{x})\}^{-0.5} = \eta, \quad (38)$$

since the constant value of κ is ignored. For the Box-Cox link, on the other hand,

$$\{u(\mathbf{x})\}^{-0.5} = \mu^\lambda. \quad (39)$$

However, from (13)

$$\mu = (1 + \lambda\eta)^{1/\lambda},$$

so that, for the Box-Cox link,

$$\{u(\mathbf{x})\}^{-0.5} = (1 + \lambda\eta) = 1 + \lambda(\theta_0 + \theta_1x_1 + \dots + \theta_kx_k).$$

The condition in Theorem 2 of §5.2.1 that the one-factor-at-a-time design is optimal therefore becomes

$$(1 + \lambda\theta_0)^2 \leq \lambda^2\theta_i\theta_j. \quad (40)$$

An advantage of the Box-Cox link in data analysis is that it is continuous at $\lambda = 0$, becoming the log link. The search over suitable links to describe the data therefore does not contain any discontinuity. In designing experiments, on the other hand, a locally optimal design will be selected for a particular λ . The results of §2.5 show that, if the power link is used, a value of κ does not have to be stated *a priori*. However, prior values will be required for $\boldsymbol{\theta}$. These will typically be ascertained from guessed responses as the factors vary. Despite the absence of explicit reference to κ in the design criterion, the value will enter implicitly through the relationship between μ and η (15). Finally, (40) shows that, as $\lambda \rightarrow 0$, the one-factor-at-a-time design will not be optimal. Further, from the (39), it follows that, under these conditions the weights $u(\mathbf{x}) \rightarrow 1$ and the design will tend to the 2^k factorial, even for large values of the θ_j .

5.3 Poisson data

D-optimal designs for Poisson regression with the log link, $\log \mu = \eta$, share some similarities with the Gamma designs in §5.2. In particular, for log linear models with a first-order linear predictor,

$$\log \mu = \eta = \theta_0 + \sum_{j=1}^k \theta_j x_j, \quad (41)$$

the optimal design has a similar structure to those from Theorem 2.

There is only a moderate number of results on designs for Poisson regression in the literature. For (41) and $k = 1$, Minkin (1993) found locally optimal designs for estimating θ_1 ; see also Chapter 14 and the references therein for more general results on models with one variable. For $k = 1, 2$, Wang, Myers, Smith, and Ye (2006) investigated the dependence of locally optimal designs on functions of the parameter values and Wang, Smith, and

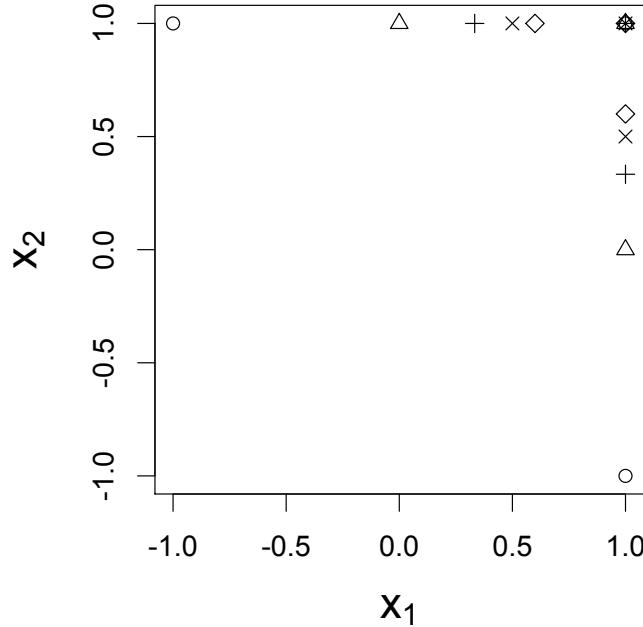


Figure 12: Support points for locally D-optimal designs for Poisson regression with (41) and $\boldsymbol{\theta} = (0, r, r)$. Key: \circ $r = 1$; \triangle $r = 2$; $+$ $r = 3$; \times $r = 4$; \diamond $r = 5$. All designs include the point $(1, 1)$.

Ye (2006) developed sequential designs. For a single variable, Ford et al. (1992) used a transformation of the design space to a canonical form, together with geometrical arguments along the lines of §5.1.1, to find locally optimal designs for a class of nonlinear models that included Poisson regression.

Russell et al. (2009) addressed the problem of D-optimal design for (41) with $k \geq 1$ and provided the following theorem.

Theorem 3. (Russell et al., 2009) *A D-optimal design for Poisson regression (41) with $l_i \leq x_{ij} \leq u_i$ and $|\theta_i(u_i - l_i)| \geq 2$ ($i = 1, \dots, k$) has the $k + 1$ equally weighted support points*

$$\begin{aligned}\mathbf{x}_i &= \mathbf{c} - \frac{2}{\theta_i} \mathbf{e}_i, \quad i = 1, \dots, k \\ \mathbf{x}_{k+1} &= \mathbf{c},\end{aligned}\tag{42}$$

for \mathbf{e}_i the i th column vector of the $k \times k$ identity matrix, $i = 1, \dots, k$, and $\mathbf{c} = (c_1, \dots, c_k)'$, where $c_i = u_i$ if $\theta_i > 0$ and $c_i = l_i$ if $\theta_i < 0$.

The proof of Theorem 3 is via a canonical transformation and an application of the general equivalence theorem. Note that the D-optimal design does not depend on the value of the intercept, θ_0 , and is invariant to permutation of the factor labels. The requirement $|\theta_i(u_i - l_i)| \geq 2$ is not overly restrictive in practice; $\mathcal{X} = [-1, 1]^k$ requires $|\theta_i| \geq 1$, $i = 1, \dots, k$. In Figure 12, we give the support points for $k = 2$ and a number of example parameter vectors, $\boldsymbol{\theta} = (0, r, r)'$. Notice the one-factor-at-a-time structure of the design and how the support points tend towards $(1, 1)$ as r increases. Figure 13 gives the support points in the induced design space \mathcal{Z} , projected into z_1, z_2 and defined from equation (33). The optimal support points lie on the boundary of the induced space. Not only do the values of the z_j increase with r , but the induced design region itself becomes more elongated as r increases.

For $|\theta_i(u_i - l_i)| < 2$, numerical investigation has found that both the optimal support points and weights depend on θ_0 , in addition to the other parameters. As expected, as $|\theta_i/\theta_0|$ tends to zero, for $i = 1, \dots, k$, the D-optimal design tends to the equally weighted factorial design.

For more general linear predictors, for example containing interactions or quadratic terms, numerical search is required to find optimal designs. This is also the case for Bayesian D-optimal designs except for the special case of minimally-supported designs for (41), that is, designs with $t = k + 1$ support points. McGree and Eccleston (2012) provided theoretical results for minimally-supported designs robust to a set of models of the form (41) defined through a discrete set of parameter vectors. We extend their result to Bayesian minimally supported D-optimal designs.

Theorem 4: *Assume a Poisson regression model with linear predictor (41). The Bayesian D-optimal design among the class of minimally supported designs, minimising (25), is the locally D-optimal design (42) for the parameter vector $\boldsymbol{\theta}^* = E(\boldsymbol{\theta})$ provided $|E(\theta_i)(u_i - l_i)| \geq 2$.*

Proof: *For a minimally-supported design, the model matrix*

$$\mathbf{X} = [\mathbf{f}(\mathbf{x}_1), \dots, \mathbf{f}(\mathbf{x}_p)]'$$

is $p \times p$. Now the objective function (25) can be written as

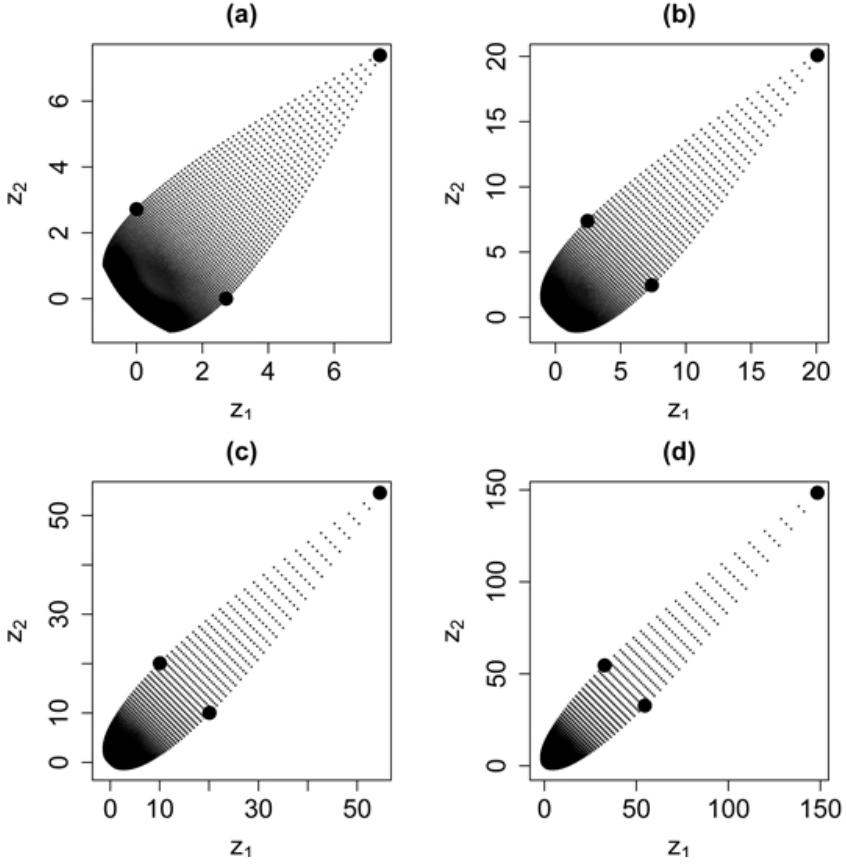


Figure 13: Support points in the induced design space \mathcal{Z} for D-optimal designs for Poisson regression with (41) and $\boldsymbol{\theta} = (0, r, r)$. (a) $r = 2$; (b) $r = 3$; (c) $r = 4$; (d) $r = 5$.

$$\Psi(\xi) = - \int_{\Theta} \log |\mathbf{M}(\xi; \boldsymbol{\theta})| p(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad (43)$$

$$\begin{aligned} &= - \int_{\Theta} 2 \log |\mathbf{X}| p(\boldsymbol{\theta}) d\boldsymbol{\theta} - \int_{\Theta} \log \prod_{i=1}^t w_i \exp(\eta_i) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= -2 \log |\mathbf{X}| - \sum_{i=1}^t \left[\log w_i + \int_{\Theta} \eta_i p(\boldsymbol{\theta}) d\boldsymbol{\theta} \right] \\ &= -2 \log |\mathbf{X}| - \sum_{i=1}^t [\log w_i - \eta_i^*] \\ &= -\log |\mathbf{M}(\xi; \boldsymbol{\theta}^*)|, \end{aligned} \quad (44)$$

where $\boldsymbol{\theta}^* = E(\boldsymbol{\theta})$ and $\eta_i^* = \theta_0^* + \sum_{j=1}^k \theta_j^* x_j$. The fourth line above follows as η_i is a linear function of $\boldsymbol{\theta}$. The equality of (43) and (44) establishes that, provided $|E(\theta_i)(u_i - l_i)| \geq 2$, design (42) is Bayesian D-optimal among the class of minimally supported designs.

To illustrate Theorem 4, we find a Bayesian minimally supported D-optimal design for (41) and $k = 5$ factors with $x_i \in [-1, 1]$, $\theta_0 = 0$ and each $\theta_i \sim U(a, b)$ ($i = 1, \dots, 5$). The values of a and b are given in Table 9 (a) in terms of a common parameter α . Increasing α leads to more diffuse prior densities. However, for any $\alpha \geq 2$, the Bayesian minimally-supported D-optimal design is given by the locally D-optimal design for $\theta_0 = 0$ and $\theta_i = (a + b)/2 = (-1)^{(i+1)}(1 + \alpha/2)$; see Table 9 (b).

Table 9: Bayesian minimally supported D-optimal design: (a) Ranges for the uniform, $U(a, b)$, prior densities for $\theta_1, \dots, \theta_5$; (b) Equally weighted support points; $\beta = [(\alpha - 2)/(\alpha + 2)]$ for $\alpha = 2, 5, 10, 20$.

(a) Parameter ranges

(b) Support points

| Parameter | Limits | | x_1 | x_2 | x_3 | x_4 | x_5 |
|------------|---------------|--------------|-------|-------|----------|---------|----------|
| | a | b | | | | | |
| θ_1 | 1 | $1 + \alpha$ | 2 | 1 | $-\beta$ | 1 | -1 |
| θ_2 | $-1 - \alpha$ | -1 | 3 | 1 | -1 | β | -1 |
| θ_3 | 1 | $1 + \alpha$ | 4 | 1 | -1 | 1 | $-\beta$ |
| θ_4 | $-1 - \alpha$ | -1 | 5 | 1 | -1 | 1 | $-\beta$ |
| θ_5 | 1 | $1 + \alpha$ | 6 | 1 | -1 | 1 | -1 |

We assess the performance of these designs through simulation of 10,000 parameter vectors from the uniform distributions defined by Table 9 (a) for $\alpha = 2, 5, 10$ and 20. For each parameter vector, we derive the locally D-optimal design from Theorem 3, and then calculate the D-efficiency (27) for the design in Table 9 (b). The induced empirical cumulative distribution functions are given in Figure 14.

For relatively precise prior information ($\alpha = 2$), high efficiency is maintained for all the samples from the prior distribution, with minimum efficiency of 79% and median of 93%. As the prior distribution becomes more diffuse (with increasing α), the induced efficiency distribution also becomes more diffuse. For $\alpha = 5$, the minimum and median efficiencies are 53% and 85% respectively; the corresponding summaries for $\alpha = 10$ are 34% and 80% and for $\alpha = 20$ are 21% and 75%. The high median efficiencies maintained

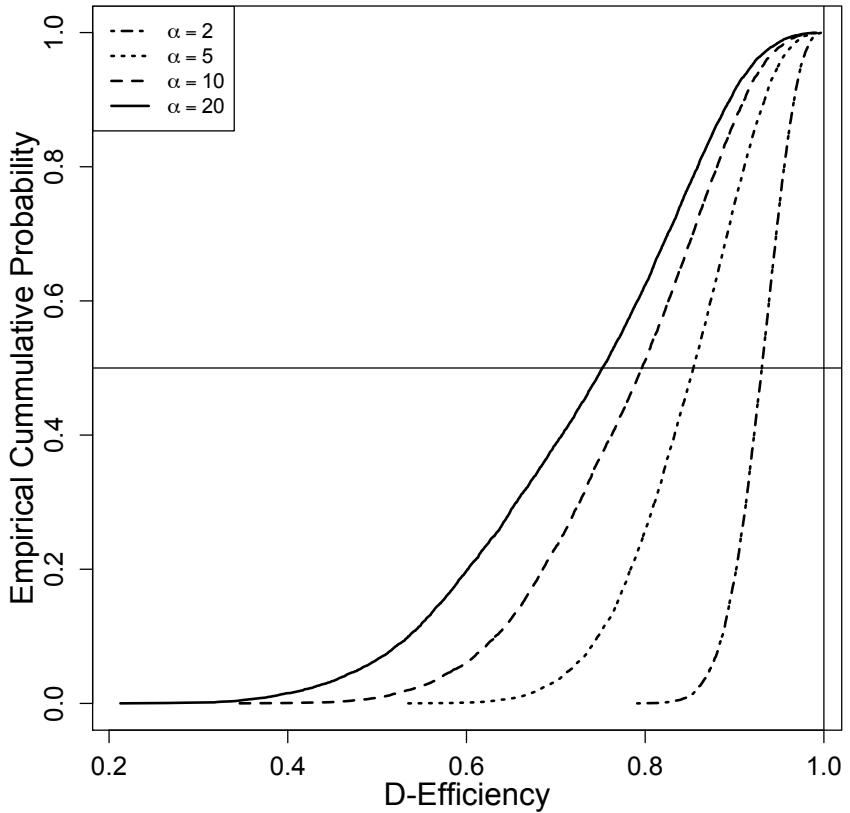


Figure 14: Empirical cumulative distributions for the D-efficiency of the Bayesian minimally supported D-optimal design for four different prior distributions; see Table 9.

for more diffuse distributions are not typical of minimally supported designs in general (for example, see van de Ven and Woods, 2014 for binary data designs). However, it seems the structure of the Poisson designs, with all the support points on the boundary, makes Bayesian minimally supported designs an efficient choice (see also McGree and Eccleston, 2012)

6 Designs with dependent data

There is a variety of practical experiments in which the observed responses may be dependent; see also Chapter 5. Most obviously, and perhaps most importantly, are experiments where there is a natural grouping in the ex-

perimental units such that, given the same treatment, two observations from within the same group are expected to be more similar than two observations in different groups. Common examples for non-normal data include longitudinal studies in clinical and pharmaceutical trials (Diggle et al., 2002) and blocked experiments in industry (Robinson et al., 2004). Our exposition in this section focuses on block designs but is equally appropriate for other grouped experiments.

For experiments involving the observation of a continuous response that is appropriately described by a normal distribution, there is, of course, a wealth of literature on the design and analysis of blocked experiments for linear models (see Part II of this book). There is also a considerable literature on design for nonlinear models and dependent data, much of it stemming from the world of pharmacokinetics/pharmacodynamics (PK/PD) and the seminal work of Mentré et al. (1997). For experiments with discrete data, with a binary or count response, there are rather fewer results available, although the results we describe below naturally share some similarities with the PK/PD literature, particularly the so-called first-order approximation (see Retout and Mentré, 2003 and Bazzoli et al., 2010).

As with linear models, the first decision is whether to model the block effects as fixed or random. We will choose the latter option and take a mixed model approach as: (i) in industrial experiments, the blocks are usually a nuisance factor and not of interest in themselves; (ii) random block effects allow prediction of the response for unobserved blocks; and (iii) pragmatically, when finding optimal designs for nonlinear models, the use of random block effects reduces the number of unknown parameters for which prior information is required. For example, with b blocks, a fixed-effects design would require the specification of $b - 1$ block effects. See Stufken and Yang (2012) for locally optimal designs with fixed group effects and a single variable.

6.1 Random intercept model

To model the responses from a blocked experiment, we adopt the framework of Generalized Linear Mixed Models (GLMMs; Breslow and Clayton, 1993) and, in particular, apply random intercept models. We develop our framework for b blocks each of equal size m . For the j th unit in the i th block,

$$y_{ij} | \gamma_i \sim \pi(\mu_{ij}), \quad \text{for } i = 1, \dots, b; j = 1, \dots, m,$$

where $\pi(\cdot)$ is a distribution from the exponential family with mean μ_{ij} and

$$g(\mu_{ij}) = \mathbf{f}'(\mathbf{x}_{ij})\boldsymbol{\theta} + \gamma_i.$$

Here $g(\cdot)$ is the link function and \mathbf{x}_{ij} is the i, j th combination of variable values. As before, the vector $\mathbf{f}(\mathbf{x}_{ij})$ holds known functions of the k variables and $\boldsymbol{\theta}$ holds the p unknown regression parameters. The unobservable random block effects γ_i are assumed to follow independent $N(0, \sigma_\gamma^2)$ distributions, with σ_γ^2 known. Under this model, observations in different blocks are independent. More general models including additional random effects, such as random slopes, may also be defined.

6.2 Continuous block designs

We choose to generalise (17) to include blocks of fixed size m through

$$\xi = \begin{Bmatrix} \zeta_1 & \dots & \zeta_t \\ w_1 & \dots & w_t \end{Bmatrix},$$

where $\zeta_l \in \mathcal{X}^m$ is the set of design points that form the l th block (or ‘‘support point’’), $0 < w_l \leq 1$ is the corresponding weight; $\sum_{i=1}^t w_i = 1$. See Cheng (1995) and Woods and van de Ven (2011). For example, if $k = m = 2$, a possible continuous design is

$$\xi = \begin{Bmatrix} \zeta_1 = \{(-1, -1), (1, 1)\} & \zeta_2 = \{(1, -1), (-1, 1)\} \\ 0.5 & 0.5 \end{Bmatrix},$$

that is, one-half of the b blocks in a realised exact design would contain design points $x_1 = x_2 = -1$ and $x_1 = x_2 = 1$, and the other half would contain design points $x_1 = 1, x_2 = -1$ and $x_1 = -1, x_2 = 1$.

6.3 Information matrix for a random intercept model

To apply the usual design selection criteria, for example, D-optimality, we need to derive and evaluate the information matrix for $\boldsymbol{\theta}$. As observations in different blocks are independent,

$$\mathbf{M}(\xi; \boldsymbol{\tau}) = \sum_{l=1}^t w_l \mathbf{M}(\zeta_l; \boldsymbol{\tau}),$$

where $\boldsymbol{\tau} = (\boldsymbol{\theta}', \sigma_\gamma^2)$, and the information matrix for the l th block is, by definition,

$$\mathbf{M}(\zeta_l, \boldsymbol{\tau}) = E_{\mathbf{y}_l} \left\{ -\frac{\partial^2 \log p(\mathbf{y}_l | \boldsymbol{\tau}, \zeta_l)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right\}, \quad (45)$$

where the m -vector $\mathbf{y}_l = (y_{l1}, \dots, y_{lm})'$ holds the responses from block ζ_l .

Direct calculation of the expectation in (45) is possible for small experiments with binary data through

$$\mathbf{M}(\zeta_l, \boldsymbol{\tau}) = \sum_{\mathbf{y}_l \in \{0,1\}^m} -\frac{\partial^2 \log p(\mathbf{y}_l | \boldsymbol{\tau}, \zeta_l)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} p(\mathbf{y}_l | \boldsymbol{\tau}, \zeta_l),$$

although both the marginal likelihood and its derivative will require numerical approximation (see Waite, 2013). For more practically-sized experiments, this direct calculation will be computationally infeasible.

6.4 Approximating the information matrix using estimating equations

For model estimation, the marginal likelihood can be approximated using methods such as quadrature, Markov Chain Monte Carlo or the EM algorithm (McCulloch et al., 2008). However, for the purposes of finding an optimal design using numerical search, repeated evaluation of the information matrix, or some alternative, is required. Hence, a fast approximation to (45) is needed.

An approximate variance-covariance matrix for $\boldsymbol{\theta}$ is available from the theory of estimating equations (see, for example, Godambe, 1991). For a GLMM, standard unbiased estimating equations are an extension of the score equations for a GLM and have the form

$$\sum_{l=1}^t w_l \mathbf{X}_l' \boldsymbol{\Delta}_l \mathbf{V}_l^{-1} (\mathbf{y}_l - \boldsymbol{\mu}_l) = \mathbf{0},$$

where $\boldsymbol{\Delta}_l = \text{diag}[d\mu_{lj}/d\eta_{lj}]$, $\eta_{lj} = \mathbf{f}'(\mathbf{x}_{lj})\boldsymbol{\theta}$, and \mathbf{X}_l and $\boldsymbol{\mu}_l$ are the $m \times p$ model matrix and $m \times 1$ mean vector defined for the l th block; \mathbf{V}_l is a $m \times m$ weight matrix for the observations from the l th block. Depending on the approximation, $\boldsymbol{\mu}_l$ may be either the *marginal* mean response or the *conditional* mean response given $\gamma_l = 0$.

The approximate variance-covariance of the resulting estimators is given by

$$\text{Var}(\hat{\boldsymbol{\theta}}) \approx \left(\sum_{l=1}^t w_l \mathbf{X}_l' \boldsymbol{\Delta}_l \mathbf{V}_l^{-1} \boldsymbol{\Delta}_l \mathbf{X}_l \right)^{-1} \quad (46)$$

The inverse of this variance-covariance matrix can be used as an approximation to the information matrix $\mathbf{M}(\xi, \boldsymbol{\tau})$.

Various different choices of \mathbf{V}_l and $\boldsymbol{\mu}_l$ have been proposed, of which we will discuss the following three:

1. Quasi-Likelihood (QL): $\mathbf{V}_l = \text{Var}(\mathbf{Y}_l)$ and $\boldsymbol{\mu}_l = E(\mathbf{Y}_l)$, the marginal variance and mean; see Wedderburn (1974). The marginal variance-covariance matrix for \mathbf{Y}_l is generally not available in closed-form for non-normal data. One notable exception is for the Poisson distribution, for which Niaparast (2009) and Niaparast and Schwabe (2013) used quasi-likelihood to find D-optimal designs.
2. Marginal quasi-likelihood (MQL): $\mathbf{V}_l = \text{diag}[\text{Var}(Y_{lj})] + \boldsymbol{\Delta}_l \mathbf{J} \boldsymbol{\Delta}_l \sigma_\gamma^2$, with $\mathbf{J} = \mathbf{1}\mathbf{1}'$, and $\boldsymbol{\mu}_l = E(\mathbf{Y}_l | \gamma_l = 0)$; see Breslow and Clayton (1993). Here, a linear mixed model approximation is used for the marginal variance-covariance of \mathbf{Y}_l . This approximation has been used to find designs for binary data by authors such as Moerbeek and Maas (2005), Tekle et al. (2008) and Ogungbenro and Aarons (2011).
3. Generalized estimating equations (GEEs):

$$\mathbf{V}_l = \{\text{diag}[\text{Var}(Y_{lj})]\}^{1/2} \mathbf{R} \{\text{diag}[\text{Var}(Y_{lj})]\}^{1/2},$$

with \mathbf{R} an intra-block marginal “working correlation” matrix, and $\boldsymbol{\mu}_l = E(\mathbf{Y}_l | \gamma_l = 0)$; see Liang and Zeger (1986). The matrix \mathbf{R} is assumed to be independent of \mathbf{x} and usually chosen to have a standard form, for example, exchangeable or known up to a small number of correlation parameters. For discrete data, it is in fact often impossible for either of these two assumptions to hold but the resulting estimators can be relatively efficient compared to a full maximum likelihood approach (Chaganty and Joe, 2004). Methodology for D-optimal designs using this approach was developed by Woods and van de Ven (2011).

Note that for each of these approximations, if $\sigma_\gamma^2 \rightarrow 0$ (or equivalently, $\mathbf{R} \rightarrow \mathbf{I}$), the variance-covariance matrix for \mathbf{Y}_l reverts to that for a simple GLM. Waite and Woods (2013) developed, assessed and compared a variety of methods of approximating \mathbf{M} to find D-optimal designs for a GLMM.

6.5 Comparison of approximations

We use a small example to perform a simple comparison of designs from the three approximations in the previous section. Consider an experiment in blocks of size $m = 2$ with a single variable x to collect count data. Conditional on the random block effect, we assume $Y_{ij} | \gamma_i \sim \text{Poisson}(\mu_{ij})$ and we choose a second-order predictor and the log link

$$\log(\mu_{ij}) = \eta = \gamma_i + \theta_0 + \theta_1 x + \theta_2 x^2,$$

for $x \in [-1, 1]$. For the purposes of finding designs, we assume point prior information and set $\theta_0 = 0$, $\theta_1 = 5$ and $\theta_2 = 1$. The random block effect has distribution $\gamma_i \sim N(0, \sigma_\gamma^2)$ for $i = 1, \dots, b$ and $\sigma_\gamma^2 = 0.5$.

For the log link, $\Delta_l = \text{diag}\{\mu_{lj}\}$. We consider each of the three approximations, with μ_{lj} representing either the marginal or conditional mean response for the j th point in the l th block of support ($l = 1, \dots, t; j = 1, \dots, m$).

Quasi-Likelihood: Here, $\mu_{lj} = E(Y_{lj}) = \exp(\eta_{lj} + \sigma_\gamma^2/2)$, the marginal mean response, and

$$\mathbf{V}_l = \text{diag}\{\exp(\eta_{lj} + \sigma_\gamma^2/2)\} + \{\exp(\sigma_\gamma^2) - 1\} \bar{\boldsymbol{\mu}}_l \bar{\boldsymbol{\mu}}_l'$$

where $\bar{\boldsymbol{\mu}}_l' = \{\exp(\eta_{lj} + \sigma_\gamma^2/2)\}_{j=1}^m$.

Marginal Quasi-Likelihood: For this approximation, $\mu_{lj} = \exp(\eta_{lj})$, the conditional mean response given $\gamma_l = 0$, and

$$\mathbf{V}_l = \text{diag}\{\exp(\eta_{lj})\} + \sigma_\gamma^2 \boldsymbol{\mu}_l \boldsymbol{\mu}_l'$$

where $\boldsymbol{\mu}_l' = \{\exp(\eta_{lj})\}_{j=1}^m$.

Generalized Estimating Equations: Now, $\mu_{lj} = \exp(\eta_{lj})$, the conditional mean response given $\gamma_l = 0$, and

$$\mathbf{V}_l = \text{diag}\{\exp(\eta_{lj})\}^{\frac{1}{2}} \mathbf{R} \text{diag}\{\exp(\eta_{lj})\}^{\frac{1}{2}},$$

where the working correlation matrix for this example is

$$\mathbf{R} = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}.$$

Table 10: D-optimal continuous block designs with blocks of size $m = 2$ for a Poisson example and Quasi-Likelihood, Marginal Quasi-Likelihood and Generalized Estimating Equations approaches (to 2 decimal places).

| | Support blocks | | |
|---------|----------------|-----------|-----------|
| | Block 1 | Block 2 | Block 3 |
| QL/MQL | (0.10, 0.88) | (0.75, 1) | — |
| Weights | 0.5 | 0.5 | — |
| GEE | (0.02, 0.84) | (0.72, 1) | (0.26, 1) |
| Weights | 0.38 | 0.35 | 0.27 |

Table 11: Efficiencies of three optimal designs under three approximations to the information matrix.

| Design | Approximation | |
|--------|---------------|------|
| | QL/MQL | GEE |
| QL/MQL | 1 | 0.90 |
| GEE | 0.87 | 1 |

For the GEE design, we redefine $\boldsymbol{\tau} = (\boldsymbol{\theta}', \alpha)$.

Locally D-optimal designs under the three approximations are given in Table 10. Note that the same design was found under the QL and MQL approximations and that, for the GEE design, $\alpha = 0.5$ was chosen so that the working correlation closely matched the average intra-block correlation (≈ 0.49) from the other design. Table 11 gives the D-efficiencies (27) of each design under each approximation; the GEE design is 87% efficient under the QL and MQL approximations, whilst the QL/MQL design is 90% efficient under the GEE approximation.

Optimality of these designs can be confirmed, as in §3.1, via the application of a multivariate equivalence theorem; see Atkinson (2008b) and Woods and van de Ven (2011). A necessary and sufficient condition for a design ξ^* to be locally D-optimal is

$$\psi(\mathbf{x}, \xi^*; \boldsymbol{\tau}) = p - \text{trace} \{ \mathbf{M}(\zeta; \boldsymbol{\tau}) \mathbf{M}^{-1}(\xi^*; \boldsymbol{\tau}) \} \geq 0, \quad (47)$$

for all $\zeta \in \mathcal{X}^m$. This condition can be verified numerically; Figure 15 plots the derivative function for each of the three approximations, with the support points of the optimal designs marked. Notice that: (i) the support points occur at minima of the derivative surface, with $\psi(\mathbf{x}, \xi^*; \boldsymbol{\tau}) = 0$; (ii) with blocks of size $m = 2$, the derivative function must be symmetric in the line $x_{l1} = x_{l2}$; and (iii) that the derivative surfaces for QL and MQL are very similar.

7 Extensions and further reading

Although we have described designs for three types of response and several link functions, interest in designs for generalized linear models continues to focus on binary responses and the logistic link. Much of this reflects the rapid growth of applications of the discrete choice models described in detail in Chapter 29.

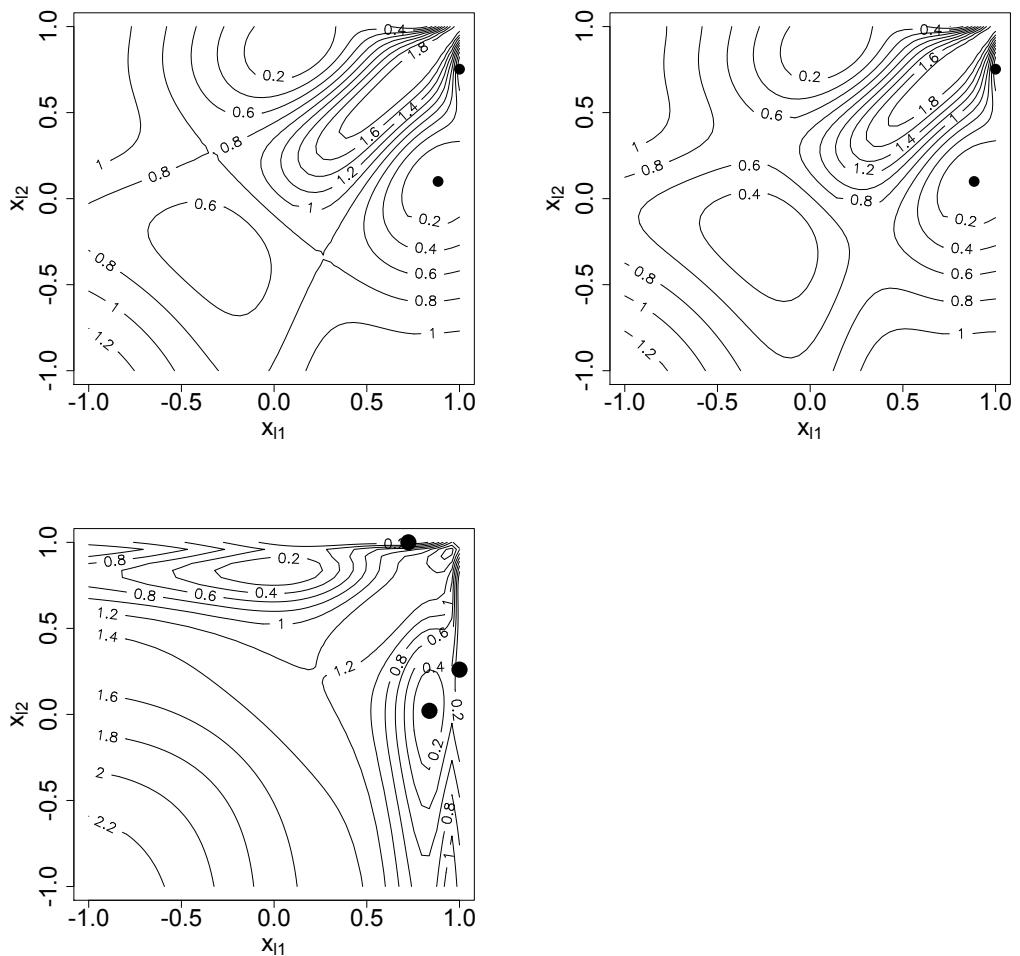


Figure 15: Derivative surfaces, $\psi(\mathbf{x}, \xi^*; \boldsymbol{\tau})$, and D-optimal designs for QL (top left), MQL (top right) and GEE (bottom left) approximations.

There is also appreciable interest in design for logistic regression in medical statistics, particularly drug development. A typical problem is to find the dose giving a specified probability of toxicity. The natural way to proceed is the use of a sequential design as described in §3.4.2. The appropriate design criterion is c -optimality, which designs to estimate this dose with minimum variance. Such designs for nonlinear regression models are exemplified by Atkinson et al. (2007, §17.5) and by Ford et al. (1992) for GLMs. Often, however, the locally optimal design has a single support point at the estimated dose. The sequential design may then fail to provide sufficient information to guarantee identification of the required dose (Pronzato, 2000; O’Quigley and Conaway, 2010). The designs need to provide sufficient perturbation in the experimental conditions to ensure convergence.

The designs we have exemplified, particularly for first-order models, often have the number of support points equal to the number of parameters. They therefore fail to provide any information for model checking and choice. Designs for discriminating between two regression models were introduced by Fedorov and Malyutov (1972) and by Atkinson and Fedorov (1975) who called them T-optimal. Ponce de Leon and Atkinson (1992) and Waterhouse et al. (2008) extend T-optimality to GLMs. A general design criterion for discrimination between models using Kullback-Leibler distances is that of López-Fidalgo et al. (2007). A potential disadvantage of these designs is that they focus exclusively on model discrimination. Compound designs for the joint problem of parameter estimation and model discrimination, called DT-optimal, are given, for regression models, by Atkinson (2008a). Waterhouse et al. (2008) also attend to the quality of parameter estimates, but not through use of a compound criterion. D-optimal designs robust to the form of the linear predictor were developed by Woods et al. (2006) and Woods and Lewis (2011).

In some applications it is not necessary to establish a model that holds globally. In the context of dose finding, O’Quigley and Conaway (2010) recommends the use of a model that has adequate local behaviour. A more formal approach to model uncertainty is in Chapter 24. In particular, Li and Wiens (2011), considers approximate models in dose response experiments, whereas Wiens (2009) provides robust designs for discrimination between regression models.

In §2.5 we mentioned that response transformation is sometimes an alternative to the use of a gamma model. Atkinson and Cook (1997) find optimal designs for estimation of the transformation in linear models whereas Atkinson (2005) studies designs for transforming both sides of a nonlinear model.

Finally, we note that some interesting models for non-normal responses have a structure as in §2.1 but with a predictor which is nonlinear in the

parameters. Optimal designs for one particular generalized nonlinear model are given by Biedermann and Woods (2011).

Acknowledgments

The authors thank Dr T.W. Waite (University of Southampton, UK) for assistance with the computation in §6. D.C. Woods was supported by a Fellowship from the UK Engineering and Physical Sciences Research Council (EP/J018317/1), and the work was partly undertaken while the authors were Visiting Fellows at the Isaac Newton Institute for the Mathematical Sciences (Cambridge, UK).

References

Agresti, A. (2002). *Categorical Data Analysis, 2nd Ed.* New York: Wiley.

Atkinson, A. C. (2005). Robust optimum designs for transformation of the responses in a multivariate kinetic model. *Technometrics* 47, 478–487.

Atkinson, A. C. (2006). Generalized linear models and response transformation. In A. I. Khuri (Ed.), *Response Surface Methodology and Related Topics*, Chapter 8. Singapore: World Scientific.

Atkinson, A. C. (2008a). DT-optimum designs for model discrimination and parameter estimation. *Journal of Statistical Planning and Inference* 138, 56–64.

Atkinson, A. C. (2008b). Examples of the use of an equivalence theorem in constructing optimum experimental designs for random-effects nonlinear regression models. *Journal of Statistical Planning and Inference* 138, 2595–2606.

Atkinson, A. C. (2010). The non-uniqueness of some designs for discriminating between two polynomial models in one variable. In A. Giovagnoli, A. C. Atkinson, and B. Torsney (Eds.), *mODa9 - Advances in Model-Oriented Design and Analysis*, pp. 9–16. Heidelberg: Springer-Verlag.

Atkinson, A. C. and R. D. Cook (1995). D-optimum designs for heteroscedastic linear models. *Journal of the American Statistical Association* 90, 204–212.

Atkinson, A. C. and R. D. Cook (1997). Designing for a response transformation parameter. *Journal of the Royal Statistical Society B* 59, 111–124.

Atkinson, A. C., A. N. Donev, and R. D. Tobias (2007). *Optimum Experimental Designs, with SAS*. Oxford: Oxford University Press.

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

Atkinson, A. C., V. V. Fedorov, A. M. Herzberg, and R. Zhang (2014). Elemental information matrices and optimal design for generalized regression models. *Journal of Statistical Planning and Inference* 144, 81–91.

Atkinson, A. C. and M. Riani (2000). *Robust Diagnostic Regression Analysis*. New York: Springer–Verlag.

Bazzoli, C., S. Retout, and F. Mentré (2010). Design evaluation and optimisation in multiple response nonlinear mixed effect models: Pfim 3.0. *Computer methods and programs in biomedicine* 98, 55–65.

Berger, M. and W. K. Wong (2009). *An Introduction to Optimal Designs for Social and Biomedical Research*. New York: Wiley.

Biedermann, S. and D. C. Woods (2011). Optimal designs for generalized non-linear models with application to second-harmonic generation experiments. *Journal of the Royal Statistical Society C* 60, 281–299.

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology* 22, 134–167.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B* 26, 211–246.

Box, G. E. P. and N. R. Draper (1963). The choice of a second order rotatable design. *Biometrika* 50, 335–352.

Box, G. E. P. and N. R. Draper (2007). *Response Surfaces, Mixtures, and Ridge Analysis, 2nd Edition*. New York: Wiley.

Box, G. E. P. and W. G. Hunter (1965). Sequential design of experiments for nonlinear models. In *Proceedings IBM Scientific Computing Symposium: Statistics*, pp. 113–137. New York: IBM.

Box, G. E. P. and H. L. Lucas (1959). Design of experiments in nonlinear situations. *Biometrika* 46, 77–90.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88, 9–25.

Burridge, J. and P. Sebastiani (1994). D-optimal designs for generalised linear models with variance proportional to the square of the mean. *Biometrika* 81, 295–304.

Chaganty, N. R. and H. Joe (2004). Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society B* 66, 851–860.

Chaloner, K. and K. Larntz (1989). Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference* 21, 191–208.

Chaloner, K. and I. Verdinelli (1995). Bayesian experimental design: a review. *Statistical Science* 10, 273–304.

Cheng, C.-S. (1995). Optimal regression designs under random block-effects models. *Statistica Sinica* 5(2), 485–497.

Collett, D. (2002). *Modelling Binary Data*, 2nd Ed. London: Chapman and Hall/CRC.

Cox, D. R. (1988). A note on design when response has an exponential family distribution. *Biometrika* 75, 161–164.

Dette, H., V. B. Melas, and P. Shpilev (2012). T-optimal designs for discrimination between two polynomial models. *Annals of Statistics* 40, 188–205.

Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. Zeger (2002). *Analysis of Longitudinal Data* (2nd ed.). Oxford: Oxford University Press.

Dobson, A. (2001). *An Introduction to Generalized Linear Models*, 2nd edition. London: Chapman and Hall.

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

Dror, H. A. and D. M. Steinberg (2008). Sequential experimental designs for generalized linear models. *Journal of the American Statistical Association* 103, 288–298.

Fedorov, V. and M. Malyutov (1972). Optimal designs in regression problems. *Math. Operationsforsch. Statist.* 3, 281–308.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. New York: Academic Press.

Fedorov, V. V. and P. Hackl (1997). *Model-Oriented Design of Experiments*. Lecture Notes in Statistics 125. New York: Springer Verlag.

Fedorov, V. V. and S. L. Leonov (2014). *Optimal Design for Nonlinear Response Models*. Boca Raton: Chapman and Hall/ CRC Press.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80, 27–38.

Ford, I., B. Torsney, and C. F. J. Wu (1992). The use of a canonical form in the construction of locally optimal designs for non-linear problems. *Journal of the Royal Statistical Society B* 54, 569–583.

Godambe, V. P. (Ed.) (1991). *Estimating Functions*. Oxford: Oxford Science Publications.

Goos, P. and B. Jones (2011). *Optimal Design of Experiments: A Case Study Approach*. New York: Wiley.

Gotwalt, C. M., B. A. Jones, and D. M. Steinberg (2009). Fast computation of designs robust to parameter uncertainty for nonlinear settings. *Technometrics* 51, 88–95.

Khuri, A. I., B. Mukherjee, B. K. Sinha, and M. Ghosh (2006). Design issues for generalized linear models: A review. *Statistical Science* 21, 376–399.

Kiefer, J. (1959). Optimum experimental designs (with discussion). *Journal of the Royal Statistical Society B* 21, 272–319.

Kiefer, J. and J. Wolfowitz (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics* 12, 363–366.

King, J. and W. K. Wong (2000). Minimax D-optimal designs for the logistic model. *Biometrics* 56, 1263–1267.

Li, P. and D. P. Wiens (2011). Robustness of design in dose–response studies. *Journal of the Royal Statistical Society B* 73, 215–238.

Liang, K. and S. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.

López-Fidalgo, J., C. Trandafir, and C. Tommasi (2007). An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society B* 69, 231–242.

Magnus, J. R. and H. Neudecker (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester: Wiley.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models (2nd edition)*. London: Chapman and Hall.

McCulloch, C. E., S. R. Searle, and J. M. Neuhaus (2008). *Generalized, Linear and Mixed Models* (2nd ed.). Hoboken, NJ: Wiley.

McGree, J. M. and J. A. Eccleston (2012). Robust designs for Poisson regression models. *Technometrics* 54, 64–72.

Mentré, F., A. Mallet, and D. Baccar (1997). Optimal design in random-effects regression models. *Biometrika* 84, 429–442.

Minkin, S. (1993). Experimental design for clonogenic assays in chemotherapy. *Journal of the American Statistical Association* 88, 410–420.

Moerbeek, M. and C. J. M. Maas (2005). Optimal experimental designs for multilevel logistic models with two binary predictors. *Communications in Statistics - Theory and Methods* 34, 1151–1167.

Morgan, B. J. T. (1992). *Analysis of Quantal Response Data*. London: Chapman and Hall.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley.

Myers, R. H., D. C. Montgomery, G. G. Vining, and T. J. Robinson (2010). *Generalized Linear Models: with Applications in Engineering and the Sciences* (2nd ed.). New York: Wiley.

Nelson, W. (1981). The analysis of performance-degradation data. *IEEE Transactions on Reliability* R-30, 149–155.

Niaparast, M. (2009). On optimal design for a Poisson regression model with random intercept. *Statistics and Probability Letters* 79, 741–747.

Niaparast, M. and R. Schwabe (2013). Optimal design for quasi-likelihood estimation in Poisson regression with random coefficients. *Journal of Statistical Planning and Inference* 143, 296–306.

Nyquist, H. (2013). Convergence of an algorithm for constructing minimax designs. In D. Uciński, A. C. Atkinson, and M. Patan (Eds.), *mODa 10 – Advances in Model-Oriented Design and Analysis*, pp. 187–194. Heidelberg: Springer.

Ogungbenro, K. and L. Aarons (2011). Population Fisher information matrix and optimal design of discrete data responses in population pharmacodynamic experiments. *Journal of Pharmacokinetics and Pharmacodynamics* 38, 449–469.

O’Quigley, J. and M. Conaway (2010). Continual reassessment and related dose-finding designs. *Statistical Science* 25, 202–216.

Pázman, A. (1986). *Foundations of Optimum Experimental Design*. Dordrecht: Reidel.

Ponce de Leon, A. M. and A. C. Atkinson (1992). The design of experiments to discriminate between two rival generalized linear models. In L. Fahrmeir, B. Francis, R. Gilchrist, and G. Tutz (Eds.), *Advances in GLIM and Statistical Modelling: Proceedings of the GLIM92 Conference, Munich*, pp. 159–164. New York: Springer.

Pronzato, L. (2000). Adaptive optimization and D-optimum experimental design. *Annals of Statistics* 28, 1743–1761.

Pronzato, L. and A. Pázman (2013). *Design of Experiments in Nonlinear Models*. New York: Springer.

Pukelsheim, F. (1993). *Optimal Design of Experiments*. New York: Wiley.

Retout, S. and F. Mentré (2003). Further developments of the Fisher information matrix in nonlinear mixed effects models with evaluation in population pharmacokinetics. *Biopharmaceutical Statistics* 13, 209–227.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *Annals of Mathematical Statistics* 22, 400–407.

Robinson, T., R. Myers, and D. Montgomery (2004). Analysis considerations in industrial split-plot experiments with non-normal responses. *Journal of Quality Technology* 36(2), 180–191.

Russell, K. G., D. C. Woods, S. M. Lewis, and J. A. Eccleston (2009). D-Optimal designs for Poisson regression models. *Statistica Sinica* 19, 721–730.

Santner, T. J., B. Williams, and W. Notz (2003). *The Design and Analysis of Computer Experiments*. New York: Springer-Verlag.

Sibson, R. (1974). D_A -optimality and duality. In J. Gani, K. Sarkadi, and I. Vincze (Eds.), *Progress in Statistics, Vol.2 – Proc. 9th European Meeting of Statisticians, Budapest*. Amsterdam: North-Holland.

Silvey, S. D. (1980). *Optimum Design*. London: Chapman and Hall.

Silvey, S. D. and D. M. Titterington (1973). A geometric approach to optimal design theory. *Biometrika* 60, 15–19.

Sitter, R. R. (1992). Robust designs for binary data. *Biometrics* 48, 1145–1155.

Sitter, R. R. and B. Torsney (1995a). Optimal designs for binary response experiments with two variables. *Statistica Sinica* 5, 405–419.

Sitter, R. R. and B. Torsney (1995b). D-optimal designs for generalized linear models. In C. P. Kitsos and W. G. Müller (Eds.), *MODA 4 – Advances in Model-Oriented Data Analysis*, pp. 87–102. Heidelberg: Physica-Verlag.

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika* 12, 1–85.

Stufken, J. and M. Yang (2012). Optimal designs for generalized linear models. In K. Hinkelmann (Ed.), *Design and Analysis of Experiments, Special Designs and Applications*, Chapter 4, pp. 137–165. New York: Wiley.

Stufken, J. and M. Yang (2012). On locally optimal designs for generalized linear models with group effects. *Statistica Sinica* 22, 1765–1786.

Tekle, F. B., F. E. S. Tan, and M. P. F. Berger (2008). Maximin D-optimal designs for binary longitudinal responses. *Computational Statistics and Data Analysis* 52, 5253–5262.

Torsney, B. and N. Gunduz (2001). On optimal designs for high dimensional binary regression models. In A. C. Atkinson, B. Bogacka, and A. Zhigljavsky (Eds.), *Optimal Design 2000*, pp. 275–285. Dordrecht: Kluwer.

van de Ven, P. M. and D. C. Woods (2014). Optimal blocked minimum-support designs for non-linear models. *Journal of Statistical Planning and Inference* 144, 152–159.

von Eye, A., E. Mun, and P. Mair (2011). Log-linear modeling. *Wiley Interdisciplinary Reviews: Computational Statistics* 4, 218–223.

Waite, T. W. (2013). *Designs for Experiments with Mixed Effects and Discrete Responses plus Related Topics*. Ph. D. thesis, University of Southampton.

Waite, T. W. and D. C. Woods (2013). Designs for generalized linear models with random block effects. *Under revision*.

Wang, Y., R. H. Myers, E. P. Smith, and K. Ye (2006). D-optimal designs for Poisson regression models. *Journal of Statistical Planning and Inference* 136, 2831–2845.

Wang, Y., E. P. Smith, and K. Ye (2006). Sequential designs for a Poisson regression model. *Journal of Statistical Planning and Inference* 136, 3187–3202.

Waterhouse, T. H., D. C. Woods, J. A. Eccleston, and S. M. Lewis (2008). Design selection criteria for discrimination/estimation for nested models and a binomial response. *Journal of Statistical Planning and Inference* 138, 132 – 144.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61, 439–447.

Whittle, P. (1973). Some general points in the theory of optimal experimental design. *Journal of the Royal Statistical Society, Series B* 35, 123–130.

Wiens, D. P. (2009). Robust discrimination designs. *Journal of the Royal Statistical Society, Series B* 71, 805–829.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. New York: Springer.

Woods, D. C. (2010). Robust designs for binary data: applications of simulated annealing. *Journal of Statistical Computation and Simulation* 80, 29–41.

Woods, D. C. and S. M. Lewis (2011). Continuous optimal designs for generalized linear models under model uncertainty. *Journal of Statistical Theory and Practice* 5, 137–145.

Woods, D. C., S. M. Lewis, J. A. Eccleston, and K. G. Russell (2006). Designs for generalized linear models with several variables and model uncertainty. *Technometrics* 48, 284–292.

Woods, D. C. and P. van de Ven (2011). Blocked designs for experiments with non-normal response. *Technometrics* 53, 173–182.

Yang, M., B. Zhang, and S. Huang (2011). Optimal designs for generalized linear models with multiple design variables. *Statistica Sinica* 21, 1415–1430.