Design selection criteria for discrimination between nested models for binomial data

T.H. Waterhouse [a], D.C. Woods [b,*], J.A. Eccleston [a], S.M. Lewis [b]

[a] *School of Physical Sciences, University of Queensland, Brisbane, QLD 4072, Australia*

[b] *Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK*

Abstract

The aim of an experiment is often to enable discrimination between competing forms for a response model. We consider this problem when there are two competing generalized linear models (GLMs) for a binomial response. These models are assumed to have a common link function with the linear predictor of one model nested within that of the other. We consider selection of a continuous design for use in a non-sequential strategy and investigate a new criterion, T_E-optimality, based on the difference in the deviances from the two models. A comparison is made with three existing design selection criteria, namely T-, D_s- and D-optimality. Issues are raised through the study of two examples in which designs are assessed using simulation studies of the power to reject the null hypothesis of the simpler model being correct, when the data are generated from the larger model. Parameter estimation for these designs is also discussed and a simple method is investigated of combining designs to form a *hybrid* design to achieve both model discrimination and estimation. Such a method may offer a computational advantage over the use of a compound criterion and the similar performance of the resulting designs is illustrated in an example.

MSC: primary 62K05, 62J12; secondary 62K20

Key words: Binary response, Continuous designs, Deviance, D-optimality, Hybrid designs, Likelihood ratio test, Logistic regression, T-optimality

* Corresponding author.
Email addresses: thw@maths.uq.edu.au (T.H. Waterhouse), D.C.Woods@maths.soton.ac.uk (D.C. Woods), jae@maths.uq.edu.au (J.A. Eccleston), S.M.Lewis@maths.soton.ac.uk (S.M. Lewis).

Preprint submitted to Elsevier Science *12 December 2006*

1 Introduction

When an experiment results in a binary outcome, the relationship between the k independent variables $x_1, \ldots, x_k$ and the response may be approximated by a generalized linear model (GLM) as described, for example, by McCullagh and Nelder (1989). In such models, the number, Y_j, of successes at the jth distinct design point follows a binomial distribution $\mathrm{Bin}(m_j, \pi_j)$, for $j = 1, \ldots, n$. Further, the success probability π_j is related to the jth treatment (combination of variable values) $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{kj})'$ through

$$g(\pi_j) = f(\boldsymbol{x}_j)'\boldsymbol{\beta}\,,$$

where $g(\cdot)$ is the *link function* and $\eta_j = f(\boldsymbol{x}_j)'\boldsymbol{\beta}$ is the *linear predictor*, with $f(\boldsymbol{x}_j)$ a $q \times 1$ vector of known functions and $\boldsymbol{\beta}$ a $q \times 1$ vector of unknown parameters. Thus the link function relates the success probability, π_j, to the linear predictor. Examples of link functions are the logit, $\eta_j = \log(\pi_j/(1-\pi_j))$; the probit, $\eta_j = \Phi^{-1}(\pi_j)$, where Φ is the normal cumulative distribution function; and the complementary log-log link, $\eta(\boldsymbol{x}_j) = \log\{-\log(1-\pi_j)\}$. It is assumed that observations from an experiment are independent and that a single observation is made on each of the $N = \sum_{j=1}^{n} m_j$ experimental units. Also, the units are assumed to be exchangeable in the sense that the distribution of the response to a treatment does not depend on the unit to which the treatment is applied.

Most work in the literature has focused on finding designs for GLMs that allow accurate estimation of the unknown model parameters; see, for example, Firth and Hinde (1997) and Woods, Lewis, Eccleston, and Russell (2005). However, earlier experimentation may aim to choose between two or more models, each of which offers a plausible description of the response. As an example, suppose that two models differ by one or more interaction terms. Then the ability to choose between these alternatives in an early experiment has an important impact on the effective design of subsequent investigations. For such model discrimination experiments, different designs may be required from those for parameter estimation. To find designs for linear or nonlinear models, Atkinson and Fedorov (1975a,b) proposed the T-optimality criterion. They showed that, for two competing models, T-optimal designs lead to the most powerful F-test for the lack of fit of one arbitrarily chosen model, under the assumption that the other model is "true". Recent work on designs for model discrimination includes the sequential approach for linear models of Dette and Kwiecien (2004) and methods for multi-response nonlinear models by Uciński and Bogacka (2005).

For discrimination designs for GLMs, Ponce de Leon and Atkinson (1992) formulated the T-optimality criterion in terms of the deviance, where the use of the deviance as a goodness-of-fit test statistic is equivalent to the use of an F-test for a linear model. A weighted sum of deviances was used by Müller and Ponce de Leon (1996a) to find sequential designs for model discrimination for binomial data. They used a simulation study to assess designs for discriminating between GLMs with logit and probit link functions and the same linear predictors. As the logit and probit link functions are almost linearly related over most of their common domain, the problem of discrimination between the corresponding GLMs is usually of less practical importance than that of discrimination between models with a common link function and differing linear predictors.

In this paper, we investigate a variety of optimality criteria for choosing designs to discriminate between two nested GLMs for binomial data; that is, between models M_1 and M_2 with the linear predictor for M_1 nested within that for M_2. We consider the set Ξ of all possible continuous designs, defined by probability measures on $\mathcal{X} = [-1, 1]^k$, the space of possible design points. Each design ξ has n distinct, or support, points and is represented as

$$\xi = \left\{ \begin{matrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \\ w_1 & w_2 & \cdots & w_n \end{matrix} \right\} ,$$

where the vector $\boldsymbol{x}_j$ holds the values of the k variables at the jth support point. The weight $w_j \in [0, 1]$ represents the proportion of the total experimental effort expended on the jth support point, so that $\sum_{j=1}^{n} w_j = 1$ (see, for example, Fedorov and Hackl, 1997).

In Section 2 we discuss four criteria for design selection: a T-optimality criterion based on deviance, D_S-optimality, D-optimality under the larger model, M_2, and a new criterion, T_E-optimality, for discriminating between nested models which is based on the difference between the deviances for the models. A critical comparison of the criteria is made and their effectiveness examined through simulation studies of power in Section 3. This is the first study for GLMs, to the best of our knowledge, that compares criteria in this way. In Section 4, we consider designs which have the dual aim of both estimating models M_1 and M_2 and discriminating between them. These hybrid designs, which are found by amalgamating a design for discrimination and a design for estimation, are compared with designs found using a compound objective criterion proposed by Atkinson (2005).

2 Criteria

The criteria discussed in this section are motivated by two closely related objectives: first, the testing of the assertion that model M_1 is correct (T- and T_E-optimality); secondly, the estimation of some or all of the model parameters in model M_2 as accurately as possible (D- and D_s-optimality).

2.1 Deviance-based T-optimality

The deviance is an often-used measure of the goodness-of-fit of a GLM to the observed responses $\boldsymbol{y}$ obtained from design ξ. For model M_i ($i = 1, 2$), it is defined by

$$D_i(\xi, \boldsymbol{y}) = 2l(\xi, \boldsymbol{y}, \boldsymbol{y}) - 2l(\xi, \hat{\boldsymbol{\pi}}_i, \boldsymbol{y}) \,, \tag{1}$$

where $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{1i}, \ldots, \hat{\pi}_{ni})'$ is the estimate of $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)'$ obtained by using maximum likelihood estimates of the q_i unknown parameters, $\boldsymbol{\beta}_i$, of model M_i, and

$$l(\xi, \boldsymbol{\gamma}, \boldsymbol{y}) = N \sum_{j=1}^{n} w_j \left[\tilde{\pi}_j \log \gamma_j + (1 - \tilde{\pi}_j) \log(1 - \gamma_j)\right] \tag{2}$$

is a log-likelihood function for $\boldsymbol{\gamma} = \tilde{\boldsymbol{\pi}}$ or $\hat{\boldsymbol{\pi}}_i$, with $\tilde{\pi}_j = y_j/m_j$ and $m_j = Nw_j$ ($j = 1, \ldots, n$). The maximum value of (2), corresponding to the saturated model, is $l(\xi, \tilde{\boldsymbol{\pi}}; \boldsymbol{y})$ and is achieved when the estimates and observations coincide.

A T-optimal design ξ_T^* for discriminating between models M_1 and M_2 maximizes the deviance of M_1 when M_2 is assumed to be the true model, that is,

$$D_1(\xi_T^*, \boldsymbol{\mu}_2) = \max_{\xi \in \Xi} D_1(\xi, \boldsymbol{\mu}_2) \,, \tag{3}$$

see Ponce de Leon and Atkinson (1992). In this equation, $\boldsymbol{\mu}_2$ is the vector of expected responses under M_2 having jth entry $m_j g^{-1}(\eta_j^{(2)})$, where $\eta_j^{(i)}$ is the value of the linear predictor at the jth support point for model M_i ($i = 1, 2$; $j = 1, \ldots, n$). As $\boldsymbol{\mu}_2$ depends on the value of $\boldsymbol{\beta}_2$, ξ_T^* is a *locally* optimal design. The observed response $\boldsymbol{y}$ is unknown prior to experimentation and hence the expected response under model M_2 is used in (3).

A potential disadvantage of T-optimality is that the deviance $D_1(\xi, \boldsymbol{Y})$ may not follow the $\chi^2_{n-q_1}$ distribution given by asymptotic theory, particularly for sparse data, see McCullagh and Nelder (1989, p.119-121). A particular concern is that a large deviance is not necessarily evidence against the null hypothesis that the data are adequately described by model M_1. Sparseness may be particularly acute for data from designed experiments (see Woods et al., 2005, for an example) and may also lead to other problems, such as non-convergence of the iterative procedures in maximum likelihood estimation (for example, Firth, 1993).

2.2 T_E-optimality

For nested GLMs, a more useful measure of the adequacy of M_1 is to test the hypothesis that M_1 is the correct model against the alternative that M_2 is correct, using the *reduction* in deviance achieved by fitting M_2 to the data compared with fitting M_1 (as discussed, for example, by Agresti, 2002, p.141). For a given design ξ, the reduction in deviance is the value of the likelihood ratio test statistic given by

$$\begin{aligned} R(\xi, \boldsymbol{y}) &= D_1(\xi, \boldsymbol{y}) - D_2(\xi, \boldsymbol{y}) \\ &= 2\left\{ l(\xi, \hat{\boldsymbol{\pi}}_2, \boldsymbol{y}) - l(\xi, \hat{\boldsymbol{\pi}}_1, \boldsymbol{y}) \right\} . \end{aligned} \tag{4}$$

Under the null hypothesis, $R(\xi, \boldsymbol{Y})$ follows asymptotically a χ^2 distribution on $(q_2 - q_1)$ degrees of freedom. For a finite number of support points, this approximation is regarded as being quite accurate for the difference in deviance, even though a χ^2 distribution may be an inadequate approximation for the deviance itself.

The T_E-optimality objective function is the expected value of $R(\xi, \boldsymbol{Y})$. Thus a design $\xi^*_{T_E}$ is T_E-optimal if

$$E\{R(\xi^*_{T_E}, \boldsymbol{Y})\} = \max_{\xi \in \Xi} E\{R(\xi, \boldsymbol{Y})\} . \tag{5}$$

The use of the random variable $\boldsymbol{Y}$ (as opposed to its expected value, as in T-optimality) recognizes that, even if M_2 provides a good approximation to the observed response, observations are likely to include noise. In the degenerate case when $\boldsymbol{Y}$ has mean $\boldsymbol{\mu}_2$ and variance-covariance matrix $\mathbf{0}$, the T_E-criterion reduces to the T-criterion.

2.3 D- and D_s-optimality

A D-optimal design for a GLM estimates the model parameters with minimum asymptotic generalized variance. Under model M_i ($i = 1, 2$), the objective function is given by

$$\phi_D(\xi, M_i) = |X_i' W_i X_i| \,, \tag{6}$$

where X_i is the $n \times q_i$ model matrix for M_i and W_i is a diagonal weight matrix with (j, j)th element

$$W_i(j, j) = m_j \left(\frac{d\pi_j}{d\eta_j}\right) / [\pi_j(1 - \pi_j)] \,, \qquad \text{for } j = 1, \ldots, n \,.$$

A design ξ_D^* is D-optimal if

$$\phi_D(\xi_D^*, M_i) = \max_{\xi \in \Xi} \phi_D(\xi, M_i) \,,$$

where ξ_D^* is dependent on the parameters of M_i, that is, ξ_D^* is only locally optimal. See Firth and Hinde (1997) and Woods et al. (2005) for methods of overcoming this limitation.

The D_s-optimality criterion seeks a design which estimates a particular subset of the parameters of a given model as accurately as possible and hence may be applied to the nested models M_1 and M_2. In order to find a D_s-optimal design $\xi_{D_s}^*$ for the additional parameters in M_2, the objective function is

$$\phi_{D_S}(\xi, M_1, M_2) = |X_2^{*\prime} W_2 X_2^* - X_2^{*\prime} W_2 X_1 (X_1' W_2 X_1)^{-1} X_1' W_2 X_2^*| \,,$$

where $X_2 = [X_1 | X_2^*]$ and X_2^* is the $n \times (q_2 - q_1)$ matrix with columns corresponding to the additional parameters in M_2. A D_s-optimal design $\xi_{D_s}^*$ maximizes this function, that is,

$$\phi_{D_S}(\xi_{D_s}^*, M_1, M_2) = \max_{\xi \in \Xi} \phi_{D_S}(\xi, M_1, M_2) \,.$$

The aim of accurate estimation of those terms that distinguish between the two models suggests that D_s-optimality might serve well as a criterion for selecting a design to discriminate between them, as proposed by Atkinson and Cox (1974) for linear models. See Müller and Ponce de Leon (1996b) for an application of D_s-optimality for probit models in the design of a study in economics.

2.4 Implementation

In order to compare the above criteria, each has been implemented in search algorithms to find designs. For the T-, D- and D_s-criterion, the BFGS Quasi-Newton method (Dennis and Schnabel, 1983) was used in the search. To implement the T_E-criterion, the expectation of $R(\xi, \boldsymbol{Y})$ is required, see (5), which is analytically intractable. Hence the objective function was approximated using Monte Carlo simulation (see, for example, Gentle, 2003, ch.7), with the binomial success probability at support point j obtained from independent draws from Beta(u_j,v_j) distributions, for $j = 1, \ldots, n$. The parameters u_j and v_j were chosen to make the mean of the jth Beta distribution equal to the probability of success, $\pi_j^{(2)}$, under M_2 and the variance proportional to the mean. That is,

$$\pi_j^{(2)} = \frac{u_j}{u_j + v_j} \qquad \text{and} \qquad c\pi_j^{(2)}(1 - \pi_j^{(2)}) = \frac{u_j v_j}{(u_j + v_j)^2(u_j + v_j + 1)}\,,$$

where $c > 0$ is a constant of proportionality. When the additional parameters have only small impact on the predicted response, then a small value of c should be used. This is because it is only possible to distinguish between closely similar models using a prediction-based criterion when negligible noise can be assumed. Note that the T-optimality criterion assumes zero noise. This issue is illustrated in Example 2 of Section 3.

The embedding of Monte Carlo function evaluation within design search algorithms has been successful in several areas of design; for example, Hamada, Martz, Reese, and Wilson (2001) and Woods (2005). Our implementation uses a simulated annealing algorithm (see, for example, Spall, 2003, ch.8) which has proved useful in optimizing noisy functions in a wide range of applications. Our algorithm used 1000 independent draws for the success probabilities for every function evaluation in order to achieve satisfactory control of the Monte Carlo error.

3 Comparison of criteria

A method of assessing the effectiveness of the four criteria is to evaluate the designs selected using an estimate of the power of a test of the null hypothesis H_0: M_1 is correct against the alternative hypothesis H_A: M_2 is correct, where the estimate is obtained using data from simulated experiments with a range of experiment sizes. In this section, we apply this method to investigate the criteria on two examples and use the difference in deviance between the two models as a test statistic. A hypothesis test against a specified alternative is preferred to a lack-of-fit test for M_1 based on the deviance, not only because of the previously discussed unreliable distributional properties of the deviance, but also because of the ambiguity in the number of degrees of freedom for the deviance according to whether the data are viewed as grouped or ungrouped (Davison, 2003, p.491-492).

For each criterion, a continuous design is obtained by search algorithm. An observation is generated for the jth support point as a random draw from a $\text{Bin}(m_j, \pi_j^{(2)})$ distribution, where m_j is approximated by integer rounding of Nw_j $(j = 1, \ldots, n)$. Both models M_1 and M_2 are fitted to the data set and the difference in deviance is compared with the $\chi^2_{(1-\alpha),(q_2-q_1)}$ percentile. This process is repeated n_0 times and the proportion of times that H_0 is rejected is recorded.

When the data from the simulated experiments give proportions near 0 or 1, the problem of infinite or unstable parameter estimates may arise. Although this clearly presents a problem for parameter estimation and prediction, it does not necessarily imply a poor fitting model as, usually, the fitted values, and therefore the deviance, converge to their limiting values (McCullagh and Nelder, 1989, p.117). Rather, this problem is indicative of the inability of the models to distinguish between "large" and "very large" values of the linear predictor, and similarly between "small" and "very small" values. These situations result in fitted probabilities of 1 or 0 respectively. This problem is particularly acute for the probit link. When this problem arises in the simulated experiments, we have not excluded fitted models with unstable parameters, as might be considered in an investigation concerning parameter estimation.

The following examples illustrate this method of evaluation and allow comparison of the criteria. As the T-, D- and D_s-optimality criteria find locally optimal designs, values of the parameters in M_2 need to be specified for each example. These parameters are also used to determine the Beta distribution used to evaluate the T_E-objective function. The parameters in M_1 are not needed, as the D-optimal design is only found for the larger model.

Example 1: Logistic regression with two factors: M_1 has an additive linear predictor with three parameters $(q_1 = 3)$; M_2 also includes the interaction term $(q_2 = 4)$. Thus, the linear predictors are

$$\eta^{(1)} = \beta_{01} + \beta_{11}x_1 + \beta_{21}x_2 \,,$$
$$\eta^{(2)} = \beta_{02} + \beta_{12}x_1 + \beta_{22}x_2 + \beta_{32}x_1x_2 \,.$$

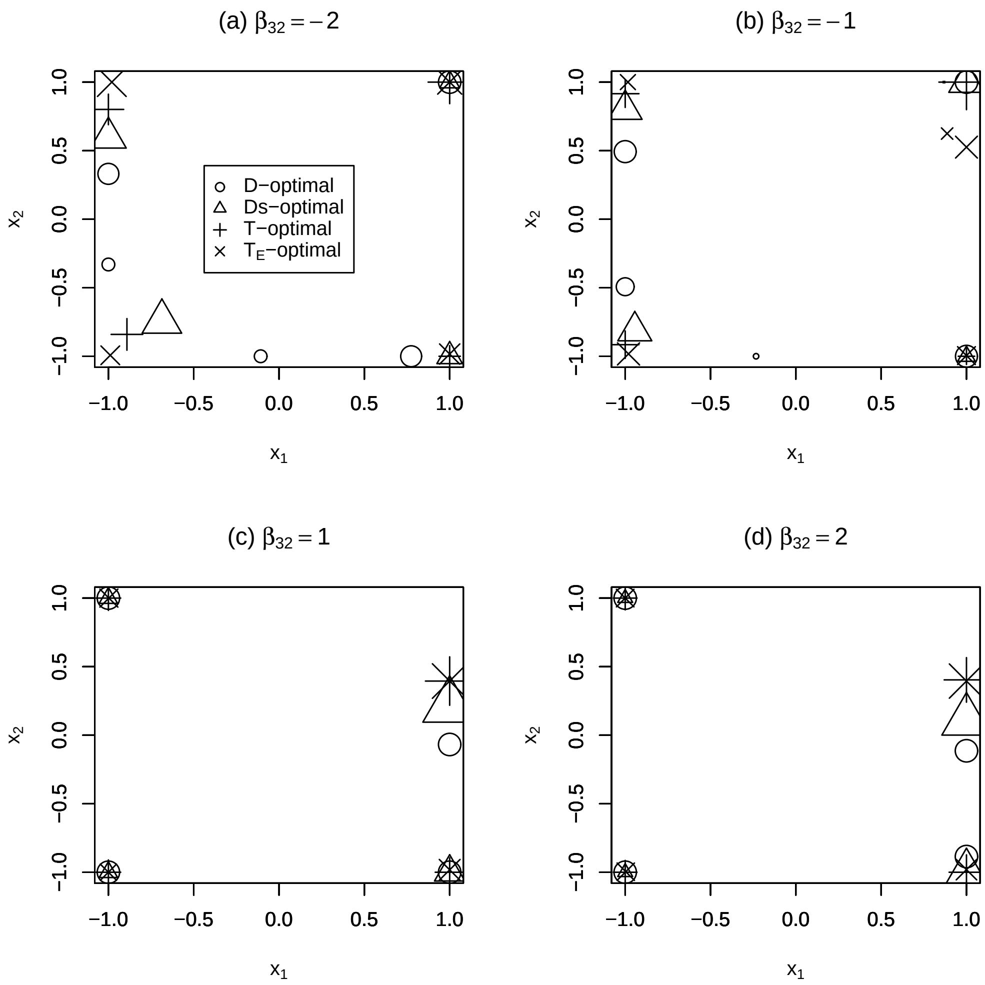

Fig. 1. Discrimination designs for four different values of additional parameter β_{32} in Example 1. The size of each plotted symbol is proportional to the weight assigned to the corresponding support point.

Four choices of parameter values for M_2 were investigated, namely $\beta_{02} = 1$, $\beta_{12} = 1$, $\beta_{22} = 2$ and $\beta_{32} \in \{-2, -1, 1, 2\}$.

Example 2: Probit regression with one factor: M_1 has a first-order linear predictor ($q_1 = 2$); M_2 is a second order model ($q_2 = 3$). Thus the linear predictors are

$$\eta^{(1)} = \beta_{01} + \beta_{11}x\,,$$
$$\eta^{(2)} = \beta_{02} + \beta_{12}x + \beta_{22}x^2\,.$$

Again, four choices of parameter values are considered: $\beta_{02} = 1$, $\beta_{12} = -2$ and $\beta_{22} \in \{-2, -1, 1, 2\}$.

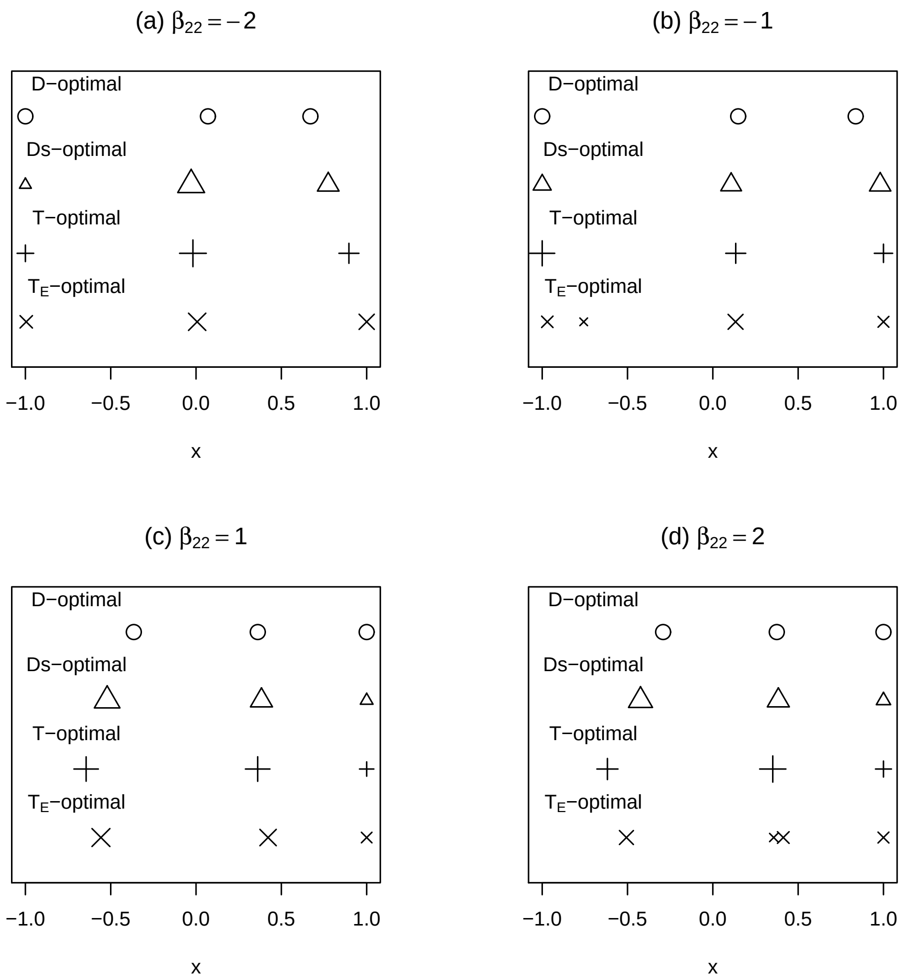

Fig. 2. Discrimination designs for four different values of additional parameter β_{22} in Example 2. The size of each plotted symbol is proportional to the weight assigned to the corresponding support point.

For each example, a design was found for each of the four criteria and for each of the four values of the additional parameter. The support points for these designs are shown in Figures 1 and 2, with the size of each plotted symbol proportional to the weight assigned to the support point. All the T_E-optimal designs were found using $c = 0.1$, except for the design in Figure 2(c) for which $c = 0.01$. For this set of parameter values, the random variation in the simulated response with $c = 0.1$ dominates the differences in the predictions from the two similar models. Hence, the T_E-optimality criterion with $c = 0.1$ produced a degenerate design with all the support points at, or very near to, $x = 1$, the point at which the variation in the response from model 2 was at its greatest. The choice of $c = 0.01$ allowed the difference between the expected responses from the two models to be distinguished from the noise and resulted in the design shown in Figure 2(c). For each example, and choice of model parameters, the locations of the support points are broadly similar for the

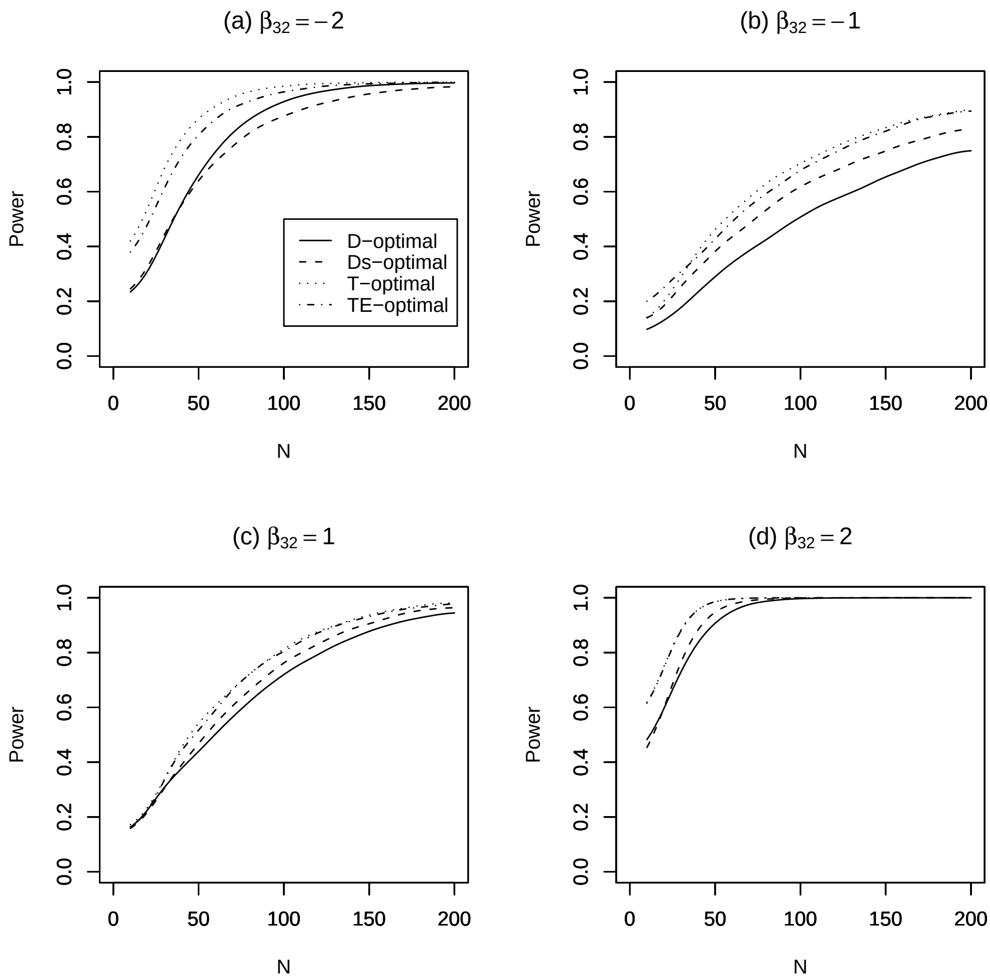

Fig. 3. Power for testing H_0 against H_A, against experiment size N, for four values of the additional parameter β_{32} in Example 1.

designs found from the four criteria. However, the locations differ according to the sign of the additional model parameter in each example.

The performances of the 16 designs were compared for each example using the estimated power of the test of H_0 against H_A for experiment sizes ranging from $N = 10$ to $N = 200$ runs. For each experiment, the power was approximated by simulating $n_0 = 1000$ sets of experimental data and performing the test with a significance level of $\alpha = 0.05$. Figures 3 and 4 show how the power for each of the designs found from the four criteria is related to the number of runs in the design for Examples 1 and 2, respectively. The curves have been smoothed using a normal kernel smoother (Eubank, 1999, ch.4), with bandwidth equal to 20 and chosen by eye to reduce the impact of Monte Carlo error.

All the designs led to higher power for a given N when there is a greater difference between M_1 and M_2 and hence an easier discrimination problem. This

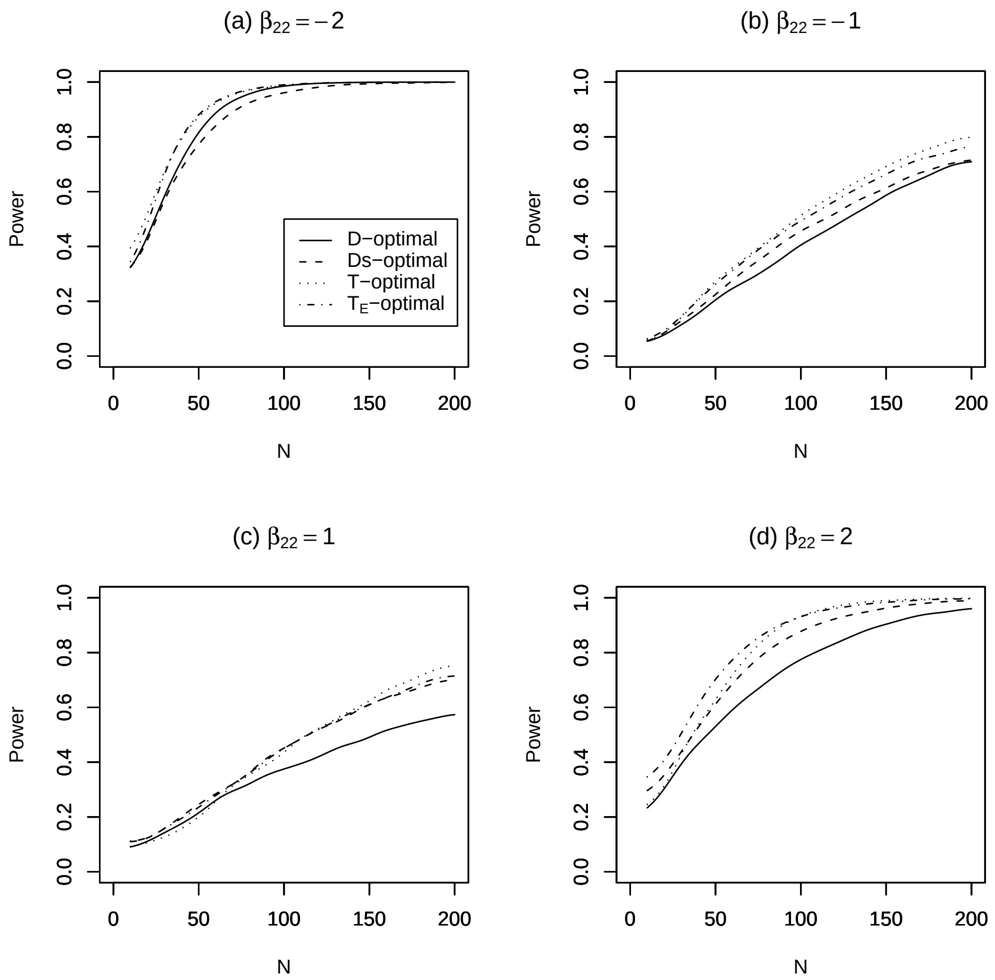

Fig. 4. Power for testing H_0 against H_A, against experiment size N, for four values of the additional parameter β_{22} in Example 2.

occurs for the larger absolute values of β_{32} (Example 1) and β_{22} (Example 2). For the more challenging problems, there are greater differences between the performances of the designs, see Figures 3(b) and 4(c), with the D-optimal designs generally having the poorest performance. For every comparison, the T-optimal design consistently gives the greatest increase in power for increasing N. The T_E-optimal designs generally have very similar performance to the T-optimal designs, with the power curves virtually indistinguishable in Figures 3(c) and 3(d).

4 Parameter estimation

Designs tailored to model discrimination may be inefficient for parameter estimation for either, or both of, M_1 and M_2. Indeed, T- and D_s-optimal designs may not even allow estimation of the parameters in M_2, particularly if there is a considerable difference between the two models, such as more than one additional term. In this section, we describe a selection criterion for designs for estimating parameters in two models. We then investigate a method of combining designs for discrimination and estimation to provide designs capable of both efficient estimation and discrimination.

4.1 Compound estimation criterion

In order to produce designs for efficient estimation of the parameters in two models, M_1 and M_2, Atkinson and Cox (1974) suggested maximization of the objective function

$$\phi_c(\xi, M_1, M_2) = (\phi_D(\xi, M_1))^{1/q_1}(\phi_D(\xi, M_2))^{1/q_2}\,, \tag{7}$$

where ϕ_D is defined in (6). This is a special case of the criterion in the seminal work of Läuter (1974) and has recently been used to find designs for GLMs by Woods et al. (2005).

4.2 Hybrid designs

We consider *hybrid* designs which are capable of both discrimination and parameter estimation. Such a design is formed as a "weighted sum" of a discrimination design ξ^d, such as a T- or T_E-optimal design with n_d support points, and a design ξ^e for estimation, such as a design maximizing (7) with n_e support points. A hybrid design ξ^h is defined as

$$\begin{aligned}\xi^h &= (1-a)\xi^d + a\xi^e \\ &= \left\{ \begin{matrix} \boldsymbol{x}_1^d & \ldots & \boldsymbol{x}_{n_d}^d & \boldsymbol{x}_1^e & \ldots & \boldsymbol{x}_{n_e}^e \\ (1-a)w_1^d & \ldots & (1-a)w_{n_d}^d & aw_1^e & \ldots & aw_{n_e}^e \end{matrix} \right\},\end{aligned}$$

where $\boldsymbol{x}_j^d$, $\boldsymbol{x}_l^e$ and w_j^d, w_l^e ($j = 1, \ldots, n_d$; $l = 1, \ldots, n_e$) are the support points and corresponding weights of the discrimination and estimation designs respectively, and $0 \leq a \leq 1$. The choice of a allows adjustment of the relative importance of discrimination and estimation: $a = 0$ gives the discrimination design; $a = 1$ gives the estimation design.

Previous work on designing experiments for the dual purpose of discrimination and estimation, ranging from Läuter (1974) to Atkinson (2005), has focused on compound and constrained criteria which combine the two objectives. The use of hybrid designs offers a computational advantage over these alternative approaches, as it is not required to carry out a separate design search for several values of a in order to investigate designs offering different trade-offs between discrimination and estimation. This is a considerable benefit if a simulation-based criterion, such as T_E-optimality, is used for discrimination.

We investigate the performance of hybrid designs formed from a T-optimal design and an estimation design for the models in Example 1. The T-optimal designs are those discussed in Section 3 and the estimation designs were obtained by design search through maximization of (7). For 101 equally spaced values of a in $[0, 1]$, hybrid designs were assessed through (i) power, as described in Section 3, with N chosen to be 100, and (ii) through their D-efficiency under model i, defined as

$$\text{Eff}_{D_i} = \left[\frac{\phi_D(\xi^h, M_i)}{\phi_D(\xi_i^*, M_i)} \right]^{1/q_i} ,$$

where ϕ_D is defined in (6) and ξ_i^* is the D-optimal design under model M_i ($i = 1, 2$). Parameter values for M_1 were obtained by fitting M_1 to data generated from model M_2 using the T-optimal design. Alternatively, any available prior knowledge of the parameters in M_1 could be used.

The results of the investigation are shown in Figure 5, where the curves for power have been smoothed using a kernel smoother with bandwidth equal to 0.1. The trade-off between discrimination and estimation ability as a varies depends upon the difference between the models. When a large difference exists, such as when $\beta_{32} = 2$, high D-efficiency, under both M_1 and M_2, and high power can be achieved simultaneously through setting $a = 0.8$, see Figure 5(d). When there is less difference between the models, maximum performance for both estimation and discrimination cannot be achieved and a substantial trade-off between the two requirements is necessary, see Figure 5(b).

4.3 DT-optimality

Atkinson (2005) described the DT-optimality criterion for estimation of, and discrimination between, two or more models with normally distributed errors. We extend this criterion to the comparison of two GLMs, M_1 and M_2, through use of the objective function

$$\phi_{DT}(\xi, \boldsymbol{\mu}_2) = (1 - a) \log D_1(\xi, \boldsymbol{\mu}_2) + a\phi_c(\xi, M_1, M_2)/2 ,$$

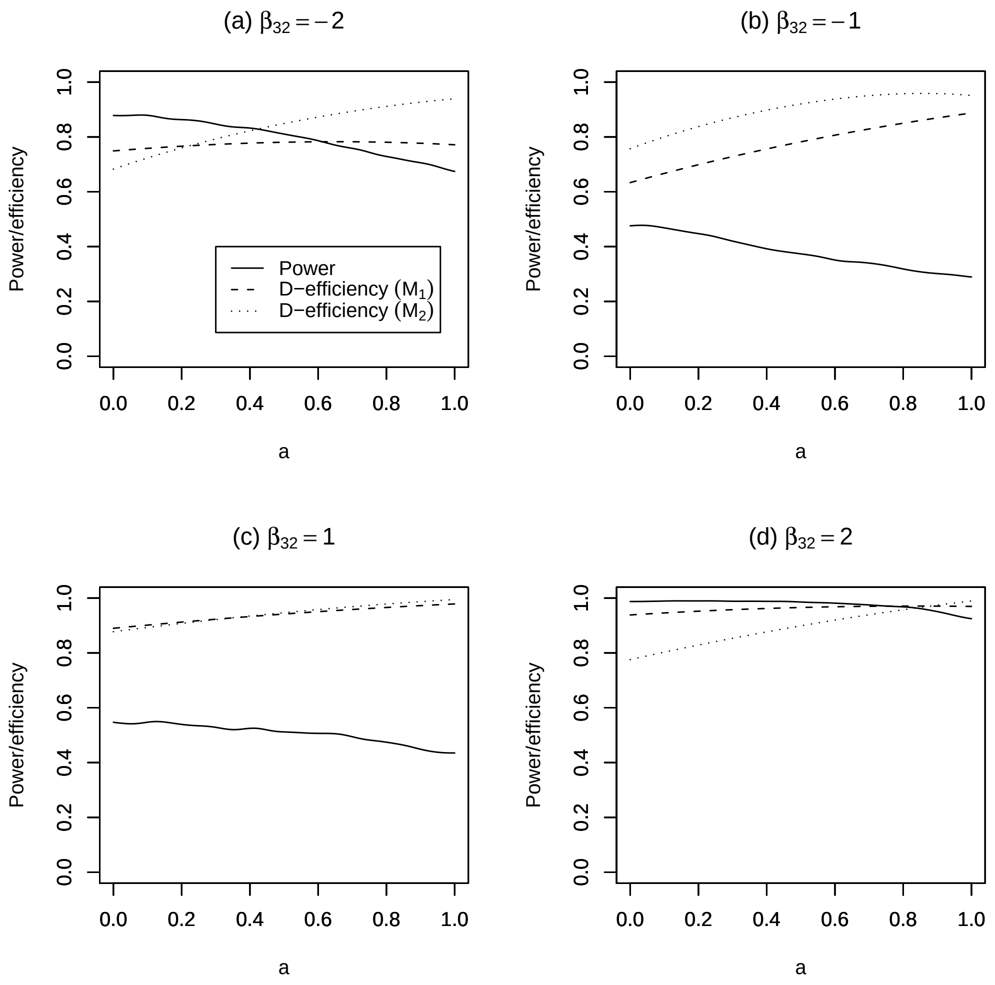

Fig. 5. The power and D-efficiency of the hybrid designs of Example 1 for $0 \leq a \leq 1$.

where $0 \leq a \leq 1$ again represents the relative importance of discrimination and estimation. We use the criterion of maximizing this function to find DT-optimal designs for the models in Example 1, for the same values of a used to calculate the hybrid designs in Section 4.2. As for the hybrid designs, these designs were assessed in terms of their power and D-efficiency and the results are shown in Figure 6. These results indicate that DT-optimal designs tend to achieve slightly lower D-efficiencies under both M_1 and M_2 than the hybrid designs and slightly higher power to discriminate between the models; see also Waterhouse and Eccleston (2005) for other nonlinear models. A direct comparison for $\beta_{32} = -1$ is shown in Figure 7. The differences in performance between the hybrid and DT-optimal designs are, however, minimal and are unlikely to influence the choice of one design over another.

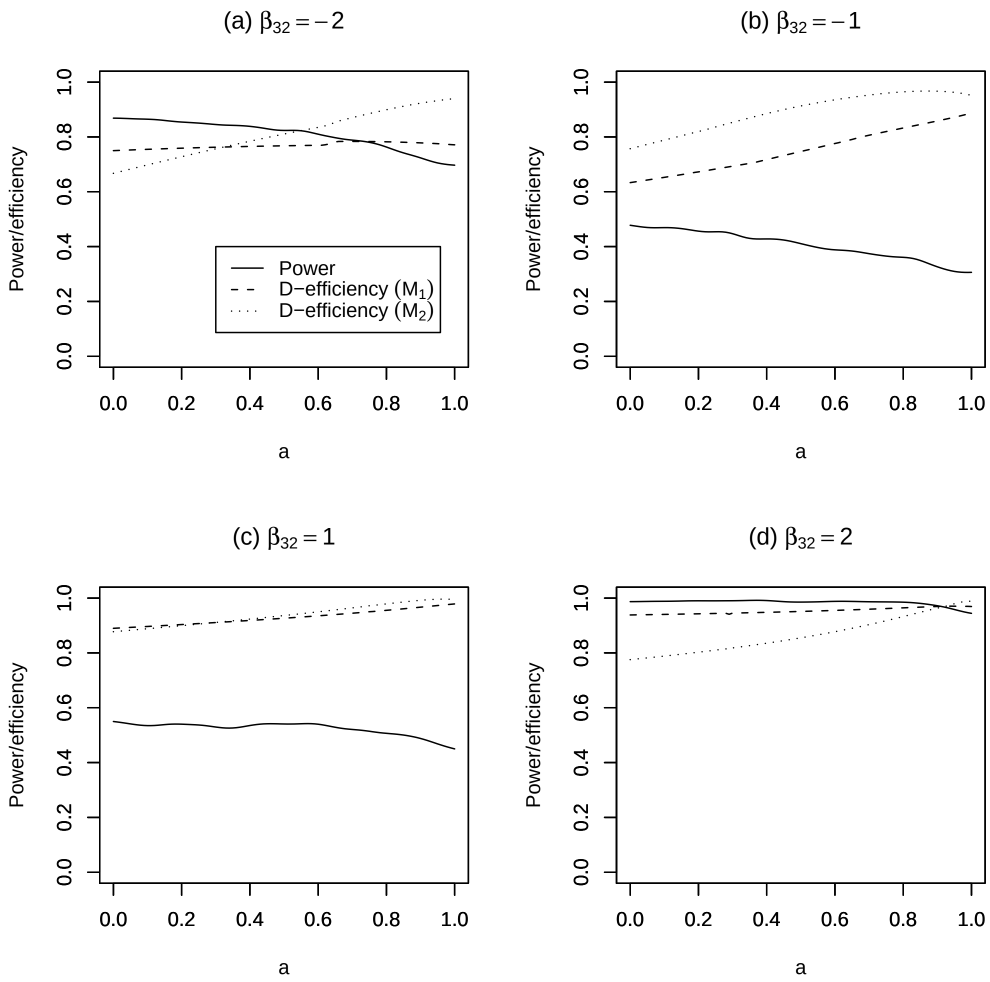

Fig. 6. The power and D-efficiency of the DT-optimal designs of Example 1 for $0 \leq a \leq 1$.

5 Discussion

When the aim of an experiment is to maximize the probability of rejecting the smaller of two nested models, given data generated from the larger model, there can be advantages in using a design which is optimal for discrimination, rather than one which is optimal under an estimation criterion. In particular, use of a discrimination design may allow a smaller experiment to be used to achieve some specified power.

Alternative methods for this problem apply a sequential approach, such as that suggested by Müller and Ponce de Leon (1996a). Although such an approach is intuitively appealing, it is not always feasible in practice and, for linear models, is often inferior in power to a non-sequential method, see Dette and Kwiecien (2004).

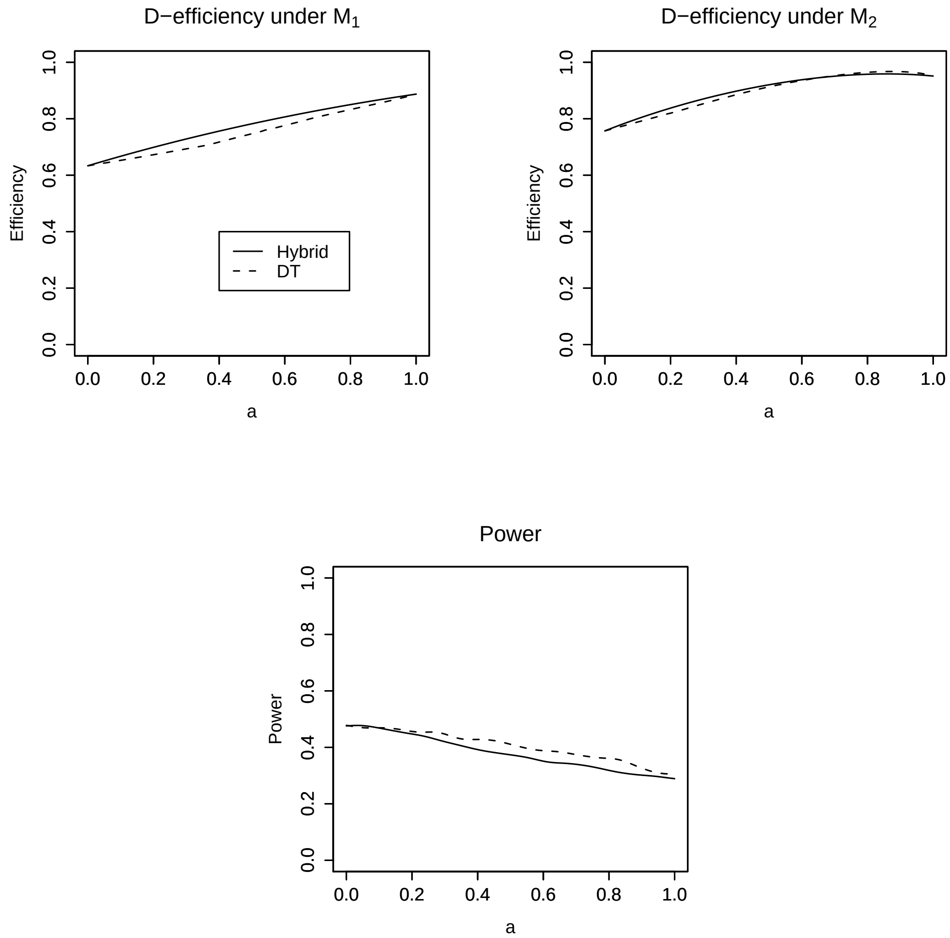

Fig. 7. The performance of hybrid and DT-optimal designs for Example 1 with $\beta_{32} = -1$, measured by their D-efficiencies under M_1 and M_2, and the power to discriminate between the two models.

An important issue in discrimination problems concerns the size of the difference between competing models. If the models are very close, in the sense of producing very similar predictions, then the experimenter needs to consider if there is a scientific need to discriminate between them. If the models are substantially different, then it is unlikely that a tailored design will be necessary to distinguish between them with reasonable power.

The four criteria discussed in this paper may be ranked according to the amount of prior information required for their implementation. For two nested GLMs, to find a D-optimal design requires knowledge of which model is correct and the corresponding parameter values; for both T- and D_s-optimal designs, the parameter values must be known for the larger model; for a T_E-optimal design, only knowledge of suitable prior distributions for the response is needed.

An advantage of a T_E-optimal design is that it is guaranteed to allow a test of model M_1 against M_2 using the difference in deviance. Such a test is not guaranteed to be possible for a T-optimal design, particularly when the models differ by more than one term, when the T-optimal design may have insufficient support points for the estimation of M_2.

Further research needed in this area includes investigations of methods to overcome the dependence of T-optimal designs on the parameter values for M_2, such as the approach of Ponce de Leon and Atkinson (1991). Another interesting extension is to find designs capable of discriminating between three or more models, as was investigated, for example, by Dette and Kwiecien (2004) for linear models.

Acknowledgements

This research was supported by grants GR/S85139 and EP/C008863 from the Engineering and Physical Sciences Research Council. The first author was supported by an Australian Postgraduate Award. The work benefited from visits by J.A. Eccleston and T.H. Waterhouse to the Southampton Statistical Sciences Research Institute and by S.M. Lewis and D.C. Woods to the University of Queensland, with D.C. Woods supported by a grant from the Royal Society.

References

Agresti, A., 2002. Categorical Data Analysis, 2nd Edition. Wiley, New York.

Atkinson, A. C., 2005. DT-optimum designs for model discrimination and parameter estimation. Submitted for publication.

Atkinson, A. C., Cox, D. R., 1974. Planning experiments for discriminating between models (with discussion). Journal of the Royal Statistical Society, Ser. B 36, 321–48.

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

Atkinson, A. C., Fedorov, V. V., 1975b. Experiments for discriminating between several models. Biometrika 62, 289–303.

Davison, A. C., 2003. Statistical Models. Cambridge University Press, Cambridge.

Dennis, Jr., J. E., Schnabel, R. B., 1983. Numerical methods for unconstrained optimization and nonlinear equations. Prentice Hall Series in Computational Mathematics. Prentice Hall Inc., Englewood Cliffs, NJ.

Dette, H., Kwiecien, R., 2004. A comparison of sequential and non-sequential designs for discrimination between nested regression models. Biometrika 91, 165–176.

Eubank, R. L., 1999. Nonparametric Regression and Spline Smoothing. Marcel Dekker, New York.

Fedorov, V. V., Hackl, P., 1997. Model-Oriented Design of Experiments. Springer, New York.

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

Firth, D., Hinde, J. P., 1997. Parameter neutral optimum design for non-linear models. Journal of the Royal Statistical Society Ser. B 59, 799–811.

Gentle, J. E., 2003. Random Number Generation and Monte Carlo Methods, 2nd Edition. Springer, New York.

Hamada, M., Martz, H. F., Reese, C. S., Wilson, A. G., 2001. Finding near-optimal Bayesian experimental designs via genetic algorithms. Amer. Statistican 62, 175–181.

Läuter, E., 1974. Experimental design in a class of models. Mathematische Operationsforschung und Statistik 5, 379–396.

McCullagh, P., Nelder, J. A., 1989. Generalized Linear Models, 2nd Edition. Chapman and Hall, New York.

Müller, W. G., Ponce de Leon, A. C. M., 1996a. Discrimination between two binary data models: Sequentially designed experiments. Journal of Statistical Computation and Simulation 55, 87–100.

Müller, W. G., Ponce de Leon, A. C. M., 1996b. Optimal design of an experiment in economics. The Economics Journal 106, 122–127.

Ponce de Leon, A. C. M., Atkinson, A. C., 1991. Optimum experimental design for discriminating between two rival models in the presence of prior information. Biometrika 78, 601–608.

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

Spall, J. C., 2003. Introduction to Stochastic Search and Optimization. Wiley, Hoboken, NJ.

Uciński, D., Bogacka, B., 2005. T-optimum designs for discrimination between two multiresponse dynamic models. J. Roy. Statist. Soc. Ser. B 67, 3–18.

Waterhouse, T. H., Eccleston, J. A., 2005. Optimal design criteria for discrimination and estimation in nonlinear models. Tech. rep., University of Queensland, School of Physical Sciences.

Woods, D. C., 2005. Designing experiments under random contamination with application to polynomial spline regression. Statist. Sinica 15, 619–635.

Woods, D. C., Lewis, S. M., Eccleston, J. A., Russell, K. G., 2005. Designs for generalized linear models with several variables and model uncertainty. Technometrics, in press.