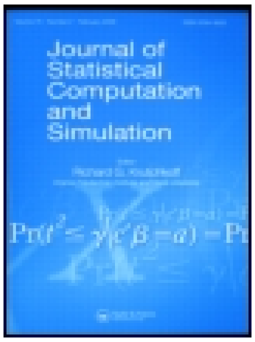

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: http://www.tandfonline.com/loi/gscs20

An algorithmic approach to construct D-optimal saturated designs for logistic model

Shwetank Lall, Seema Jaggi, Eldho Varghese, Cini Varghese & Arpan Bhowmik

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2018
https://doi.org/10.1080/00949655.2018.1425689

An algorithmic approach to construct D-optimal saturated designs for logistic model

Shwetank Lall, Seema Jaggi, Eldho Varghese*, Cini Varghese and Arpan Bhowmik

ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

ABSTRACT
In this paper, locally D-optimal saturated designs for a logistic model with one and two continuous input variables have been constructed by modifying the famous Fedorov exchange algorithm. A saturated design not only ensures the minimum number of runs in the design but also simplifies the row exchange computation. The basic idea is to exchange a design point with a point from the design space. The algorithm performs the best row exchange between design points and points form a candidate set representing the design space. Naturally, the resultant designs depend on the candidate set. For gain in precision, intuitively a candidate set with a larger number of points and the low discrepancy is desirable, but it increases the computational cost. Apart from the modification in row exchange computation, we propose implementing the algorithm in two stages. Initially, construct a design with a candidate set of affordable size and then later generate a new candidate set around the points of design searched in the former stage. In order to validate the optimality of constructed designs, we have used the general equivalence theorem. Algorithms for the construction of optimal designs have been implemented by developing suitable codes in R.

ARTICLE HISTORY
Received 4 August 2017
Accepted 5 January 2018

KEYWORDS
Fisher information matrix; standardized variance function; D-optimality; Fedorov exchange algorithm; general equivalence theorem

1. Introduction

The logistic model with the binary response can be defined as follows:

$$P(Y=1)=\frac{e^{\eta}}{1+e^{\eta}}=\frac{1}{1+e^{-\eta}},\qquad Y=0,1 \tag{1}$$

where $\eta = f(\mathbf{x}), \mathbf{x} \in \chi$ is called the linear predictor. It gives the functional form and number of input ($\mathbf{x}$) to be included in the model. Let ξ be an exact design with m support points and $\boldsymbol{\theta}$ is the vector of unknown parameters. Following Atkinson et al. [1], Fisher information matrix (FIM) for parameter vector $\boldsymbol{\theta}$ at design ξ is defined as follows:

$$\mathbf{M}(\xi,\boldsymbol{\theta})=\sum_{i=1}^{m} w(x_i,\boldsymbol{\theta})f(\mathbf{x}_i)f(\mathbf{x}_i)'w_i \tag{2}$$

CONTACT Eldho Varghese eldhoiasri@gmail.com
*Present address: ICAR-Central Marine Fisheries Research Institute, Kochi, India.

or

$$\mathbf{M}(\xi,\boldsymbol{\theta}) = \sum_{i=1}^{m} f_1(\mathbf{x}_i) f_1(\mathbf{x}_i)' w_i, \tag{3}$$

where $w(\mathbf{x},\boldsymbol{\theta}) = e^{-\eta}/(1+e^{-\eta})^2, f_1(\mathbf{x}_i) = \sqrt{w(\mathbf{x}_i,\boldsymbol{\theta})} f(\mathbf{x}_i)$ and w_i is the weight given to the support point $\mathbf{x}_i$, $w_i \in [0,1]$ and $\sum_{i=1}^{n} w_i = 1$.

Furthermore, the standardized variance function at a given input setting $\mathbf{x}$ is given by

$$d(\mathbf{x},\xi) = w(\mathbf{x},\boldsymbol{\theta}) f(\mathbf{x})' \mathbf{M}^{-1}(\xi,\boldsymbol{\theta}) f(\mathbf{x}) \tag{4}$$

or

$$d(\mathbf{x},\xi) = f_1(\mathbf{x})' \mathbf{M}^{-1}(\xi,\boldsymbol{\theta}) f_1(\mathbf{x}) \tag{5}$$

The logistic model is one of the most extensively studied nonlinear models in biological and social sciences. It establishes the nonlinear relationship between the input variables and binary/ordinal response. Chernoff [2] addressed the issue of developing designs for nonlinear models and discussed the use of initial parameter guesses. Such designs depending on initial parameter guesses are called local. Nelder and Wedderburn [3] introduced generalized linear models (GLM) and their analysis. Ford et al. [4], Sebastiani and Settini [5], Mathew and Sinha [6], and Li and Majumdar [7] provided sound theoretical results to construct D-optimal designs for the logistic model in one variable. But as the number of variables increases, the problem becomes mathematically intractable. Following Woods et al. [8], Dror and Steinberg [9] proposed modifications in existing available algorithms of linear models. Dror and Steinberg [9] basically focussed on generating robust designs for logistic models using a clustering approach assuming that the generation of locally optimal designs for the logistic model was easily possible. Khuri et al. [10] discussed the design issues for GLM and provided the review till then.

D-optimal designs for the logistic model depend on initial parameter guess and are difficult to compute analytically. The computational complexity increases with an increase in the number of design variables. Also, there are no straightforward computer-based solutions available for constructing designs in this situation. Toolkits like GOSSET [11], OPTEX procedure in SAS, statistical toolbox in MATLAB and AlgDesign package in R are not applicable here. These software tools are well suited for linear models with discrete factors. With continuous input variables, the computational cost increases heavily. In the present study, we loosely followed Woods et al. [8] and Dror and Steinberg [9] to construct locally D-optimal saturated designs for logistic models by searching in a continuous design space. We modified and simplified the row exchange computation in the Fedorov [12] algorithm to generate D-optimal saturated designs for the following cases of linear predictors:

$$\eta_i = \alpha + \beta x_{1i}, \tag{6}$$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \tag{7}$$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}, \tag{8}$$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i}, \tag{9}$$

2. Methodology

Technically, developing designs for nonlinear models is an optimization problem. The objective function being some optimality criterion (in this study, D-) defined on FIM and the candidate set representing the design space is defined for the input variables. Based on the constraints, a set of support points/design points/experimental runs are obtained from the candidate set. The choice of the candidate set is important. As we are interested in continuous input variables, no candidate set can actually represent the true design space. Intuitively, we generated a uniform grid with low discrepancy as a candidate set using the R [13] function 'expand.grid'. The idea is to search a design for this candidate set and then generate a new candidate set in the vicinity of the generated design assuming that the optimal design exists nearby. Dror and Steinberg [9] proposed using multivariate normal random numbers to generate a new candidate set. Using previously searched design (say intermediate design) and the new candidate set, the final design is obtained. In our study, we found that this new candidate set does not yield any betterment to the intermediate design. So, we devised a new way to generate the candidate set for the next stage. We simply extended the use of 'expand.grid' function of R to generate a grid of design points around the support points of the intermediate design.

In order to search the optimal design over a candidate set, we used the famous Fedorov exchange algorithm [12]. Apart from the candidate set, an initial design is required to implement this algorithm. We have taken a suitable subset of a factorial structure with non-singular information matrix as the initial design. The Fedorov exchange algorithm is basically a row exchange algorithm which replaces a row of a design at a given iteration with a row from the candidate set so as to optimize the objective function.

Algorithm:

I. Initial design: Select a design with a non-singular information matrix.
II. Candidate set: Generate a grid of points to represent the true continuous design space.
III. Exchange: Exchange points between the design and the candidate set to attain maximum gain in the determinant of information matrix of the resulting design.
IV. Repeat exchange step till no further improvement in determinant of information matrix for resultant design is observed,
V. New candidate set: Generate a grid of points in the vicinity of design found in step IV and make it the new candidate set.
VI. Repeat steps III and IV with the new candidate set.

The optimality of designs generated by the algorithmic approach above is not guaranteed. So, to check the optimality of the generated designs, we used the general equivalence theorem given by Kiefer and Wolfowitz [14] and extended to nonlinear models by White [15] and Whittle [16].

2.1. Row exchange computation

Saturated designs have runs or support points equal to the number of unknown parameters (say k) to be estimated in the model. Imposing the restriction of saturated designs has mainly three advantages. Firstly, the generated designs have the minimum possible number of support points. Secondly, calculation of weights is not needed anymore, as discussed by Silvey [17] all support points have equal weight for the case of saturated designs. And lastly, the row exchange computation can be further simplified.

Let ξ^j be the design at iteration j and '**b**' be the row to be replaced by a row '**a**' from the candidate set. Here, $\mathbf{a} = f_1(\mathbf{x}_{\text{in}})$, $\mathbf{x}_{\text{in}}$ is a point from the candidate set and $\mathbf{b} = f_1(\mathbf{x}_{\text{out}})$, x_{out} is a point from design ξ^j. Since saturated D-optimal designs have equal weights, $w_i = 1/m$, we define $\mathbf{N}(\xi,\boldsymbol{\theta}) = \mathbf{M}(\xi,\boldsymbol{\theta}) \times m$. From the appendix given at the end, we have

$$|\mathbf{N}_{j+1}| = |\mathbf{N}_j + \mathbf{aa}' - \mathbf{bb}'| = |\mathbf{N}_j|\{1 + \mathbf{a}'\mathbf{N}_j^{-1}\mathbf{a} - \mathbf{b}'\mathbf{N}_j^{-1}\mathbf{b} - (\mathbf{a}'\mathbf{N}_j^{-1}\mathbf{a})(\mathbf{b}'\mathbf{N}_j^{-1}\mathbf{b}) + (\mathbf{a}'\mathbf{N}_j^{-1}\mathbf{b})^2\} \tag{10}$$

or

$$|\mathbf{N}_{j+1}| = |\mathbf{N}_j|\Delta(\mathbf{x}_{\text{in}}, \mathbf{x}_{\text{out}}) \tag{11}$$

In other words, we have to select $\mathbf{x}_{\text{in}}$ and $\mathbf{x}_{\text{out}}$ such that $\Delta(\mathbf{x}_{\text{in}}, \mathbf{x}_{\text{out}})$ is maximized.

Lemma 1: *Let H be the class of designs with exactly k support points with equal weights, k is the number of unknown parameters in the model. Then, for each $\xi \in H$, $d(\xi, \boldsymbol{x}) = k, \forall \boldsymbol{x} \in \xi$.*

Proof: Let ξ be any design with k support points with equal weights $1/k$. So,

$$\mathbf{M}(\xi, \boldsymbol{\theta}) = \sum_{i=2}^{k} f(\mathbf{x}_i)f(\mathbf{x}_i)'/k = \mathbf{F}'\mathbf{F}/k. \tag{12}$$

Let $\mathbf{N} = \mathbf{F}'\mathbf{F}$, it can be easily shown that $\mathbf{F}\mathbf{N}^{-1}\mathbf{F}' = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix of order k. Since $f(\mathbf{x}_i)'$ is the i-th row of matrix $\mathbf{F}$, the following result follows from matrix algebra.

$$f(\mathbf{x}_i)'\mathbf{N}^{-1}f(\mathbf{x}_i) = 1 \quad \forall \mathbf{x}_i \in \xi, i = 1, 2, \ldots, k. \tag{13}$$

Subsequently,

$$f(\mathbf{x}_i)'\mathbf{M}^{-1}f(\mathbf{x}_i) = k \quad \forall \mathbf{x}_i \in \xi, i = 1, 2, \ldots, k.$$

Hence, for saturated designs in Equation (10), we get, $\mathbf{b}'\mathbf{N}_j{}^{-1}\mathbf{b} = 1$.

Thus,

$$\Delta(\mathbf{x}_{\text{in}}, \mathbf{x}_{\text{out}}) = (\mathbf{b}'\mathbf{N}_j{}^{-1}\mathbf{a})^2. \tag{14}$$

The above results can be extended to the logistic model case by simply replacing $f(\mathbf{x}_i)$ with $f_1(\mathbf{x}_i)$, where $f_1(\mathbf{x}_i) = \sqrt{w(\mathbf{x}_i, \boldsymbol{\theta})}f(\mathbf{x}_i)$ and $w(\mathbf{x}_i, \boldsymbol{\theta}) = e^{-\eta_i}/(1 + e^{-\eta_i})^2$. ■

3. Designs for logistic models

3.1. One variable and two parameters

The linear predictor defined in Equation (6) gives the logistic model with one input variable. For this case, a standard result on two-point D-optimal designs can be easily found algebraically. Here, for a given initial parameter guess say, a and b for α and β, respectively, the D-optimal design is obtained by solving $\pm 1.5434 \approx a + bx_{1i}$ [4]. But this result

Table 1. D-optimal designs for the logistic model in one variable.

	Initial parameter guess		Support points			
Design	a	b	x_{11}	x_{12}	$\|\mathbf{M}\|$	max $d(x,\xi)$
D_1	0.1	0.5	−1	1	0.054968	2
D_2	1	1	−1	1	0.026248	2
D_3	1	4	−0.636	0.136	0.003132	2

holds for an unbounded design space. The theoretical framework for bounded design space was discussed by Sebastiani and Settini [5] in Theorem 2 of their paper. The present study strictly considers the design space [−1,1] and D-optimal designs were found using an algorithmic approach. One can find the related literature for the logistic model in one variable in [4–7].

We have used a sequence from −1 to 1 with the increment of 0.01 as the candidate set and applying the algorithm explained in the previous section to obtain D-optimal design. The D-optimal designs were found in few iterations only and are reported in Table 1.

3.2. First-order model with two variables

Haines et al. [18] reported the D-optimal designs for the logistic model with a linear predictor of the first order in two variables without interaction. For the transformed design space with the adjusted initial parameter guesses as $\tilde{\beta}_0 = 9$, $\tilde{\beta}_1 = 5$ and $\tilde{\beta}_2 = 5$, Table 2 shows the designs generated from our approach. Both designs are very close and this illustration tells the success of the proposed procedure for the construction of D-optimal exact designs.

3.3. First-order model with two variables with interaction

For model (8), the linear predictor gives a first-order model with two variables and their interaction. There are 4 unknown parameters in this model and let for example take the initial parameter guesses as $\tilde{\beta}_0 = -1$, $\tilde{\beta}_1 = \tilde{\beta}_2 = 2$ and $\tilde{\beta}_3 = 0.01$. Using initial design D_7 and applying modified Fedorov algorithm design, D_8 is obtained. A new candidate set is constructed around design D_8 and ultimately design D_9 is found. The designs are reported in Table 3.

3.4. Second-order full model with two variables and six parameters

With six unknown parameters the linear predictor defined for the logistic model in Equation (9) is considered here with initial parameter guesses $\tilde{\beta}_0 = -1$, $\tilde{\beta}_1 = 2$, $\tilde{\beta}_2 = 0.5$,

Table 2. Designs for the first-order logistic model in two variables.

Initial design (D_4)		Final design (D_5)		Haenis design (D_6)	
x_1	x_2	x_1	x_2	x_1	x_2
−1	1	−1	−1	−1	−1
1	−1	−1	−0.44	−0.4408	−1
−1	−1	−0.44	−1	−1	−0.4408
$\|\mathbf{M}\|$	max $d(x,\xi)$	$\|\mathbf{M}\|$	max $d(x,\xi)$	$\|\mathbf{M}\|$	max $d(x,\xi)$
1.77E−09	233.0174	1.06E−05	3	1.06E−05	3

Table 3. Designs for the first-order logistic model in two variables with interaction.

Initial design (D_7)		Intermediate design (D_8)		Final design (D_9)	
x_1	x_2	x_1	x_2	x_1	x_2
−1	−1	−1	1	−1	1
−1	1	1	−1	1	−1
1	−1	0.64	0.64	0.64	0.64
1	1	−0.28	−0.32	−0.3024	−0.3008
$\|\mathbf{M}\|$	max $d(x, \xi)$	$\|\mathbf{M}\|$	max $d(x, \xi)$	$\|\mathbf{M}\|$	max $d(x, \xi)$
1.15E−05	9.978745	3.86E−05	4.001469	3.86E−05	4

Table 4. Designs for the second-order logistic model with 51 × 51 points in the candidate set.

Initial design (D_{10})		Intermediate design (D_{11})		Final design (D_{12})	
x_1	x_2	x_1	x_2	x_1	x_2
−1	−1	−1	1	−1	1
−1	1	1	−1	1	−1
1	−1	1	0	1	−0.0256
1	1	0.12	1	0.1424	1
0	1	0.04	0.08	−1	−0.7008
1	0	−1	−0.72	0.056	0.0672
$\|\mathbf{M}\|$	max $d(x, \xi)$	$\|\mathbf{M}\|$	max $d(x, \xi)$	$\|\mathbf{M}\|$	max $d(x, \xi)$
2.74E−09	47.53295	1.24E−08	6.574738	1.24E−08	6.643901

Table 5. Designs for the second-order logistic model with 101 × 101 points in the candidate set.

Initial design (D_{13})		Intermediate design (D_{14})		Final design (D_{15})	
x_1	x_2	x_1	x_2	x_1	x_2
−1	−1	−1	1	−1	1
−1	1	1	−1	1	−1
1	−1	0.06	0.06	−1	−0.7
1	1	1	−0.02	0.0568	0.0664
0	1	0.14	1	1	−0.0264
1	0	−1	−0.7	0.1432	1
$\|\mathbf{M}\|$	max $d(x, \xi)$	$\|\mathbf{M}\|$	max $d(x, \xi)$	$\|\mathbf{M}\|$	max $d(x, \xi)$
2.74E−09	47.55261	1.24E−08	6.622323	1.24E−08	6.646048

$\tilde{\beta}_3 = 2$, $\tilde{\beta}_4 = 0.1$ and $\tilde{\beta}_5 = 0.01$. We also intend to study the effect of density of the candidate set in our approach. For this purpose, we used two candidate sets, first with 51 × 51 points and later with 101 × 101 points.

Though the final designs were almost the same, the intermediate designs differ with respect to support points. The respective designs are reported in Tables 4 and 5.

4. Discussion

The present study mainly focuses on the construction of D-optimal designs by modifying the Fedorov algorithm in a continuous design space for more than one variable in the logistic models. The general equivalence theorem plays a key role in validating the optimality of designs constructed using an algorithmic approach. In simple words, the D-optimal design should minimize the maximum of standardized variance function value over entire design space. This minimum value is found to be the number of unknown parameters, attained at the design points only. In this study, we have used candidate sets, a grid of discrete points to represent a continuous design space. We used the candidate set to exchange and produce new designs and also checked the conditions of general equivalence theorem. For

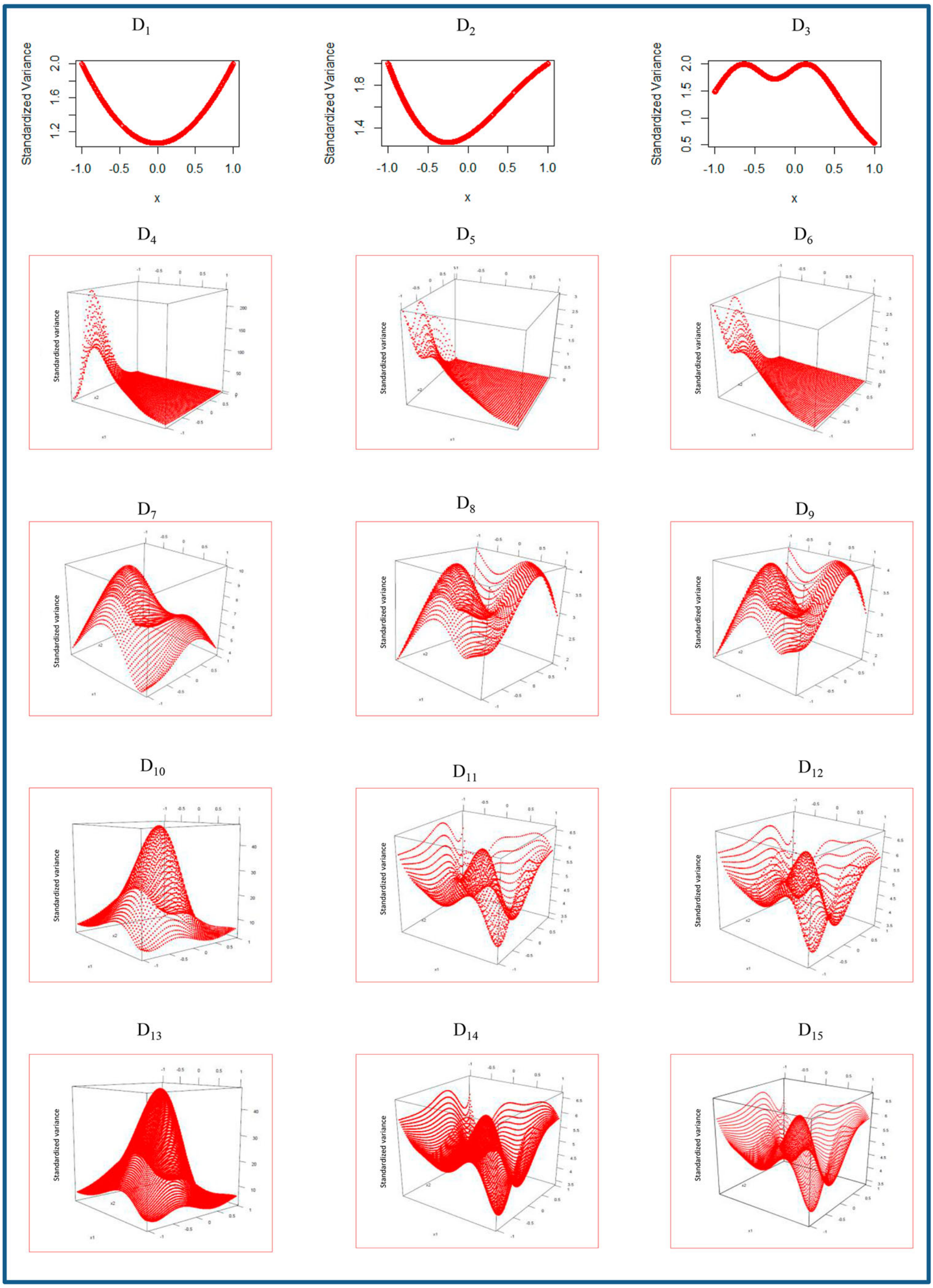

Figure 1. Standardized variance function value vs. design space of constructed designs.

Note: In this study, we have considered rescaled design variables in all examples. So the reader is advised to adjust the initial parameter guesses accordingly.

most cases, namely designs for Equations (6)–(8), we found D-optimal designs in accordance with the general equivalence theorem. But for Equation (9), we cannot say for sure that design D_{15} is D-optimal. For any given design, a plot of the standardized variance function $d(x, \xi)$ over the candidate set helps in visualizing the conditions of the general equivalence theorem. It can be seen in Figure 1 that for optimal designs, namely D_1, D_2, D_3, D_6, D_9, D_{12} and D_{15}, the maximum of $d(x, \xi)$ is k and this maximum is occurring at the design points itself. Furthermore, we found that the construction of the new candidate set around intermediate design as in step V of the proposed algorithm helps in searching optimal design even with smaller initial candidate set, thus it reduces computation as well as storage cost in implementing computer-based programs.

Acknowledgement

Authors are grateful to the Managing Editor and the Referee for the constructive comments which helped in improving the quality of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Atkinson AC, Donev AN, Tobias R. Optimum experimental designs with SAS. Oxford: Oxford University Press; 2007.
[2] Chernoff H. Locally optimal designs for estimating parameters. Ann Math Stat. 1953;24:586–602.
[3] Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc Ser A. 1972;135:370–384.
[4] Ford I, Torsney B, Wu CFJ. The use of a canonical form in the construction of locally optimal designs for non-linear problems. J R Stat Soc Ser B. 1992;54:569–583.
[5] Sebastiani P, Settimi R. A note on D-optimal designs for a logistic regression model. J Stat Plan Inference. 1997;59:359–368.
[6] Mathew T, Sinha BK. Optimal designs for binary data under logistic regression. J Stat Plan Inference. 2001;93:295–307.
[7] Li G, Majumdar D. D-optimal designs for logistic models with three and four parameters. J Stat Plan Inference. 2008;138:1950–1959.
[8] Woods DC, Lewis SM, Eccleston JA, et al. Designs for generalized linear models with several variables and model uncertainty. Technometrics. 2006;48:284–292.
[9] Dror HA, Steinberg DM. Robust experimental design for multivariate generalized linear models. Technometrics. 2006;48:520–529.
[10] Khuri AI, Mukherjee B, Sinha BK, et al. Design issues for generalized linear models: a review. Stat Sci. 2006;21:376–399.
[11] Hardin RH, Sloane NJA. A new approach to the construction of optimal designs. J Stat Plan Inference. 1993;37:339–369.
[12] Fedorov VV. Theory of optimal experiments. New York: Academic Press; 1972.
[13] R Core Team R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing; 2016. https://www.R-project.org/.
[14] Kiefer J, Wolfowitz J. The equivalence of two extremum problems. Canad J Math. 1960;12:363–366.
[15] White LV. An extension of the general equivalence theorem to nonlinear models. Biometrika. 1973;60(2):345–348.
[16] Whittle P. Some general points in the theory of optimal experimental designs. J R Stat Soc Ser B. 1973;35:123–130.
[17] Silvey SD. Optimal design. London: Chapman and Hall; 1980.

[18] Haines LM, Kabera G, Ndlovu P, et al. D-optimal designs for logistic regression in two variables. In: López-Fidalgo J, Rodríguez-Díaz JM, Torsney B, editors. mODa 8 – Advances in model-oriented design and analysis. Contributions to statistics. Heidelberg: Physica-Verlag HD; 2007.

Appendix

Fedorov [12,p.100]: Let **W** be a non-singular matrix of order $p \times p$ and **V** be a $p \times q$ order matrix. Then for any real λ

$$|\mathbf{W} + \boldsymbol{\lambda} \boldsymbol{V}\boldsymbol{V'}| = |\mathbf{W}||I + \boldsymbol{\lambda}\mathbf{V}'\mathbf{W}^{-1}\mathbf{V}|. \tag{A1}$$

Let **a** be a vector of order $p \times 1$ and **W** as defined earlier, from matrix algebra, we have

$$(\mathbf{W} + \mathbf{a}a')^{-1} = \mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{a}(1 + \mathbf{a}'\mathbf{W}^{-1}\mathbf{a})^{-1}\mathbf{a}'\mathbf{W}^{-1}. \tag{A2}$$

Now let **b** be a vector of order $p \times 1$, then using Equations (A1) and (A2):

$$\begin{aligned} |\mathbf{W} + \mathbf{aa}' - \mathbf{bb}'| &= |\mathbf{W} + \mathbf{aa}'|\{1 - \mathbf{b}'(\mathbf{W} + \mathbf{aa}')^{-1}\mathbf{b}\} \\ &= |\mathbf{W}|(1 + \dot{\mathbf{a}}'\mathbf{W}^{-1}a)[1 - \mathbf{b}'\{\mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{a}(1 + \mathbf{a}'\mathbf{W}^{-1}\mathbf{a})^{-1}\mathbf{a}'\mathbf{w}^{-1}\}\mathbf{b}] \qquad \text{(A3)} \\ &= |\mathbf{W}|\{1 + \mathbf{a}'\mathbf{W}^{-1}\mathbf{a} - \mathbf{b}'\mathbf{W}^{-1}\mathbf{b} - (\mathbf{a}'\mathbf{W}^{-1}\mathbf{a})(\mathbf{b}'\mathbf{W}^{-1}\mathbf{b}) + (\mathbf{a}'\mathbf{W}^{-1}\mathbf{b})^2\}. \end{aligned}$$