Singular prior distributions in Bayesian D-optimal design for nonlinear models

Timothy W. Waite
School of Mathematics, University of Manchester, UK

June 9, 2015

Abstract: For Bayesian D-optimal design, we define a *singular prior distribution* to be a prior distribution such that the determinant of the Fisher information matrix has a prior geometric mean of zero for all designs. For such a prior distribution, the Bayesian D-optimality criterion fails to select a design. For the exponential decay model, we characterize singularity of the prior distribution in terms of the expectations of a few elementary transformations of the parameter. For a compartmental model and multi-parameter logistic regression we establish sufficient conditions for singularity of a prior distribution. For logistic regression we also obtain sufficient conditions for non-singularity. The results are applied to show that the weakly informative prior distribution proposed as a default for inference by Gelman, Jakulin, Pittau and Su (2008) should not be used for Bayesian D-optimal design. Additionally, we develop methods to derive and assess Bayesian D-efficient designs for logistic regression when numerical evaluation of the objective function fails due to ill-conditioning.

Key words and phrases: Compartmental model, exponential decay model, ill-conditioning, logistic regression.

1 Introduction

In recent years, much effort has been devoted to developing D-optimal design methods for nonlinear problems; for example, nonlinear models (e.g. Yang (2010)), generalized linear models (Khuri, Mukherjee, Sinha and Ghosh (2006); Woods, Lewis, Eccleston and Russell (2006); Yang, Zhang and Huang (2011)), and linear models with mixed effects (Jones and Goos (2009)). In all of these areas, the set of D-optimal designs depends on the unknown values of the model parameters, $\theta \in \Theta \subseteq \mathbb{R}^p$.

One approach is to assume a particular best guess for the parameter values, and calculate a corresponding locally D-optimal design, $\xi_\theta^* \in \arg\max_{\xi\in\Xi} |M(\xi;\theta)|$, where $M(\xi;\theta)$ is the Fisher information matrix for design $\xi \in \Xi$. However, the performance of a locally optimal design may be highly sensitive to misspecification of the value of θ. Then a Bayesian approach is often used to derive designs that are efficient for a variety of plausible values for θ. This approach requires the adoption of a prior distribution, $\mathcal{P}$, on the parameters, and maximization of the value of an objective function that quantifies the expected information contained in the experiment. Throughout, we assume that $\mathcal{P}$ is a probability measure on the measure space (Θ, Σ), with Σ the Borel σ-algebra over Θ. A widely used objective function is the logarithm of the geometric mean of $|M(\xi;\theta)|$,

$$\phi(\xi;\mathcal{P}) = \int_\Theta \log|M(\xi;\theta)| d\mathcal{P}(\theta)\,, \tag{1}$$

for example, see Chaloner and Larntz (1989) and Gotwalt, Jones and Steinberg (2009). We adopt the measure-theoretic formulation of integration, under which the notation $\int_\Theta g(\theta)d\mathcal{P}(\theta) =$

$-\infty$ is standard and has a well-defined meaning for a Σ-measurable function, g, mapping Θ to the extended real line.

A design that maximizes (1) is said to be *(pseudo-)Bayesian D-optimal*, and may be used whether or not a Bayesian analysis will be performed (e.g. Woods, Lewis, Eccleston and Russell (2006)). Maximization of (1) is equivalent to maximization of an asymptotic approximation to the Shannon information gain from prior to posterior (Chaloner and Verdinelli (1995)).

In nonlinear problems, for certain *singular parameter vectors*, θ, the Fisher information matrix, $M(\xi;\theta)$, has determinant zero for *any* design ξ. For these θ, it is difficult to estimate the parameters no matter which design is used, often because of a lack of model identifiability (see Section 2.3). In this situation, the local D-optimality criterion fails to select a design. We now define the analogue of a singular parameter vector for Bayesian D-optimality.

Definition 1. (a) Given $\xi \in \Xi$, and a prior distribution, $\mathcal{P}$, we say that ξ is a *Bayesian singular design with respect to* $\mathcal{P}$ if $\phi(\xi;\mathcal{P}) = -\infty$.

(b) Given a prior distribution, $\mathcal{P}$, we say that $\mathcal{P}$ is a *singular prior distribution* if *all* $\xi \in \Xi$ are Bayesian singular with respect to $\mathcal{P}$.

Equivalently, $\mathcal{P}$ is a singular prior distribution if the geometric mean of $|M(\xi;\theta)|$ under $\mathcal{P}$ is zero for all $\xi \in \Xi$. In many models, such as the exponential decay model and logistic regression, it is straightforward to detect singular parameter vectors, θ, by inspection of the information matrix. However, as we will show, it is more difficult to detect whether $\mathcal{P}$ is a singular prior distribution, except in the case of point priors.

A different, but related, problem is the presence of ill-conditioned information matrices in a quadrature approximation to (1). This causes failure of numerical selection of Bayesian D-optimal designs, and can occur even for theoretically non-singular prior distributions.

In this paper, we clarify and extend the set of priors for which Bayesian D-optimal design is feasible for three important classes of models. In Sections 2.1, 2.2, and 2.3, respectively, we give examples of singular prior distributions for the one-factor exponential decay model, a three-parameter compartmental model, and the multi-factor logistic regression model. In Section 2.3 the default weakly informative prior proposed for logistic regression by Gelman, Jakulin, Pittau and Su (2008) is shown to be singular. For the exponential and logistic models, sufficient conditions for a prior distribution to be non-singular are established. These conditions are easily checked to ensure that the Bayesian D-optimality criterion can be used to select designs under $\mathcal{P}$. In Section 3, we develop novel methods that enable the selection of highly Bayesian D-efficient designs for logistic regression when the quadrature approximation to (1) is ill-conditioned. Finally, in Section 4 we discuss possible alternative approaches to finding efficient designs when $\mathcal{P}$ is a singular prior distribution.

2 Singularity of priors for three standard models

2.1 Exponential decay model

In this section, we derive necessary and sufficient conditions for a prior distribution to be singular for the exponential decay model. We consider two parameterizations: by rate, $\beta > 0$, and by 'lifetime', $\theta = 1/\beta > 0$. The response y is the concentration of a compound, and the explanatory variable is time, $x \in \mathcal{X} = [0, \infty)$. The model in terms of β is

$$y_i = e^{-\beta x_i} + \epsilon_i\,, \quad \epsilon_i \sim N(0, \sigma^2)\,, \tag{2}$$

where $i = 1, \ldots, n$, $x_i \geq 0$, and $\sigma > 0$.

We assume the set of competing designs is $\Xi = \mathcal{X}^n$. For a design $\xi = (x_1, \ldots, x_n) \in \Xi$, the information matrix is

$$M_\beta(\xi; \beta) = \sum_{i=1}^{n} x_i^2 e^{-2\beta x_i} \,.$$

Suppose that at least one $x_i > 0$ and let $S_{xx} = \sum_{i=1}^n x_i^2$. Then

$$-2\beta \max_{i=1,\ldots,n} \{x_i\} \leq \log |M_\beta(\xi;\beta)| - \log S_{xx} \leq -2\beta \min_{i\,:\,x_i>0} \{x_i\} \,. \tag{3}$$

By taking expectations, the following result is obtained.

Proposition 1. *Assume that at least one $x_i > 0$. For the β-parameterization, $\phi(\xi; \mathcal{P}) > -\infty$ if and only if $E_{\mathcal{P}}(\beta) < \infty$.*

Thus, here the prior, $\mathcal{P}$, is non-singular provided the rate parameter has finite expectation but, for example, is singular if β distributed a priori as the absolute value of a Cauchy random variable.

For the θ-parameterization, we have by a standard argument that

$$\log |M_\theta(\xi;\theta)| = \log |M_\beta(\xi;\beta)| - 4 \log \theta \,. \tag{4}$$

This enables derivation of the following result; for proof see the appendix.

Proposition 2. *For the θ-parameterization, the prior distribution $\mathcal{P}$ is singular if and only if either $E_{\mathcal{P}}(1/\theta) = \infty$ or $E_{\mathcal{P}}(\log \theta) = \infty$.*

In the context of designs maximizing $\phi(\xi; \mathcal{P})$ for nonlinear models, Chaloner and Verdinelli (1995) refer to potential 'technical problems using prior distributions with unbounded support where [...] $M(\xi;\theta)$ may be arbitrarily close to being singular'. Corollary 1 below shows that, even with bounded support, seemingly innocuous prior distributions can cause Bayesian D-optimality to fail as a design selection criterion.

Corollary 1. *For the θ-parameterization, the prior distribution $\mathcal{P} = U(0, a)$, $a > 0$, is singular.*

2.2 Compartmental model

In this section, we derive sufficient conditions for a prior distribution to be singular for a three-parameter compartmental model. The model is:

$$y_i = \theta_3 \{e^{-\theta_1 x_i} - e^{-\theta_2 x_i}\} + \epsilon_i \,, \qquad \epsilon_i \sim N(0, \sigma^2) \,, \tag{5}$$

where $x_i \geq 0$, $i = 1, \ldots, n$, $\theta_2 > \theta_1 > 0$, $\theta_3 > 0$ and $\sigma > 0$. The set of competing designs is $\Xi = [0, \infty)^n$. In applications, often the response y_i is a concentration of a compound in a system, and the x_i are the observation times.

The information matrix for the ith time point is

$$M(x_i;\theta) = \begin{pmatrix} x_i^2\theta_3^2 e^{-2\theta_1 x_i} & -x_i^2\theta_3^2 e^{-(\theta_1+\theta_2)x_i} & -f_i x_i e^{-\theta_1 x_i} \\ -x_i^2\theta_3^2 e^{-(\theta_1+\theta_2)x_i} & x_i^2\theta_3^2 e^{-2\theta_2 x_i} & f_i x_i e^{-\theta_2 x_i} \\ -f_i x_i e^{-\theta_1 x_i} & f_i x_i e^{-\theta_2 x_i} & f_i^2/\theta_3^2 \end{pmatrix} ,$$

where $f_i = \theta_3\{e^{-\theta_1 x_i} - e^{-\theta_2 x_i}\}$. We have $|M(\xi;\theta)| = 0$ when $\theta_1 = \theta_2$ or $\theta_3 = 0$, and $|M(\xi;\theta)| \to 0$ when $\theta_1 \to \infty$. Thus it is clear that for $\mathcal{P}$ to be a non-singular prior distribution it must not be too likely that θ_2 and θ_1 are very close, θ_3 is small, or θ_1 is large. This is formalized by Proposition 3. Let $\delta = \theta_2 - \theta_1 > 0$.

Lemma 1. *We have the following bounds on* $\log|M(\xi;\theta)|$,

$$-6\theta_1 x_{\max} \le \log|M(\xi;\theta)| - 4\log\theta_3 - \log|\tilde{M}_{\delta,1}| \le -6\theta_1 x_{\min}\,, \tag{6}$$

where above:

$$\tilde{M}^{(i)}_{\delta,\theta_3} = \begin{pmatrix} x_i^2\theta_3^2 & -x_i^2\theta_3^2 e^{-\delta x_i} & -x_i\theta_3(1-e^{-\delta x_i}) \\ -x_i^2\theta_3^2 e^{-\delta x_i} & x_i^2\theta_3^2 e^{-2\delta x_i} & x_i\theta_3 e^{-\delta x_i}(1-e^{-\delta x_i}) \\ -x_i\theta_3(1-e^{-\delta x_i}) & x_i\theta_3 e^{-\delta x_i}(1-e^{-\delta x_i}) & (1-e^{-\delta x_i})^2 \end{pmatrix}\,,$$

$$x_{\min} = \min_{i\,:\,x_i>0}\{x_i\}\,,\quad x_{\max} = \max_{i=1,\dots,n} x_i\,,\quad \tilde{M}_{\delta,\theta_3} = \sum_{i=1}^{n} \tilde{M}^{(i)}_{\delta,\theta_3}\,.$$

Lemma 2. *If* $\int_{\delta<1}\log\delta\, d\mathcal{P}(\theta) = -\infty$*, then* $E_{\mathcal{P}}(\log|\tilde{M}_{\delta,1}|) = -\infty$.

Proposition 3. *Suppose we have* $\int_{\theta_3>1}\log\theta_3\, d\mathcal{P}(\theta) < \infty$*. For the compartmental model* (5)*, the prior* $\mathcal{P}$ *is singular if* $E_{\mathcal{P}}(\theta_1) = \infty$*,* $\int_{\theta_3<1}\log\theta_3\, d\mathcal{P}(\theta) = -\infty$*, or* $\int_{\delta<1}\log\delta\, d\mathcal{P}(\theta) = -\infty$.

Heavy-tailed priors such as the half-Cauchy are increasingly recommended as weakly informative priors in various models (Gelman et al. (2008); Polson and Scott (2012)). Here, $\mathcal{P}$ is singular if θ_1 is half-Cauchy distributed, though for physiological compartmental models often more specific prior information is used (Gelman et al. (1996)). Establishment of sufficient conditions for non-singularity of $\mathcal{P}$ for this model is highly involved and beyond the scope of this paper.

2.3 Logistic regression

Suppose there are n experimental units, with associated design points $x_i = (x_{i1},\dots,x_{in})^{\mathrm{T}} \in \mathcal{X}$, and responses $Y_i \sim \text{Bernoulli}(\pi_i)$, $0 \le \pi_i \le 1$, $i = 1,\dots,n$. We assume a generalized linear model formulation (McCullagh and Nelder (1989)), with linear predictor

$$\eta_i = f^{\mathrm{T}}(x)\beta\,, \tag{7}$$

and $h(\pi_i) = \eta_i$, where $h(\pi) = \log\{\pi/(1-\pi)\}$. Above, $f(x) = (f_0(x),\dots,f_{p-1}(x))^{\mathrm{T}}$ is a vector of regression functions $f_j(x) : \mathcal{X} \to \mathbb{R}$, $j = 0,\dots,p-1$, and $\beta = (\beta_0,\beta_1,\dots,\beta_{p-1})^{\mathrm{T}}$ is a vector of regression parameters. We assume that $\mathcal{X} = [-1,1]^q$, and $\Xi = \mathcal{X}^n$.

For design $\xi = (x_1,\dots,x_n)$ and model (7) we have

$$M(\xi;\beta) = \sum_{i=1}^{n} w_i\, f(x_i) f^{\mathrm{T}}(x_i) \tag{8}$$

$$w(\eta_i) = \exp(-|\eta_i|)\,\mathrm{expit}(|\eta_i|)^2\,, \tag{9}$$

with $w_i = w(\eta_i)$, $i = 1,\dots,n$, and $\mathrm{expit}(\eta) = 1/\{1+e^{-\eta}\}$.

The following lemma enables results on singular prior distributions to be derived, and facilitates the development of numerical methods to overcome ill-conditioning in Section 3. Let F be the model matrix with rows $f^{\mathrm{T}}(x_i)$, noting that $\sum_{i=1}^{n} f(x_i)f^{\mathrm{T}}(x_i) = F^{\mathrm{T}}F$ is the information matrix of ξ under a linear model with regressors specified by f.

Lemma 3. *For logistic regression, the information matrix satisfies*

$$\min_{i=1,\dots,n}\{w_i\}F^{\mathrm{T}}F \preceq M(\xi;\beta) \preceq \max_{i=1,\dots,n}\{w_i\}F^{\mathrm{T}}F\,.$$

The inequality above is with respect to the Loewner partial ordering on real symmetric matrices, in which $M_1 \preceq M_2$ if and only if $M_2 - M_1$ is non-negative definite. Lemma 3 can be used to establish sufficient conditions for the prior distribution to be non-singular for logistic regression.

Theorem 1. *Suppose that $\mathcal{P}$ is such that $E_{\mathcal{P}}(|\beta_j|) < \infty$, for $j = 0, \ldots, p-1$. If ξ is non-singular for the linear model with regressors given by f, that is $\left|F^{\mathrm{T}}F\right| > 0$, then $\phi(\xi; \mathcal{P}) > -\infty$, i.e. ξ is also Bayesian non-singular with respect to $\mathcal{P}$ for the logistic model.*

Note that there is no requirement for $\mathcal{P}$ to have bounded support. In particular, this result provides theoretical reassurance that Bayesian D-optimality can be used to select a design with a normal or log-normal prior on the parameters.

Other important prior distributions do not satisfy the conditions of Theorem 1; for example that proposed by Gelman, Jakulin, Pittau and Su (2008), which we refer to as $\mathcal{P}_G$. Those authors recommend applying a scaling before fitting the model. For observational studies, each explanatory variable is transformed to have mean zero and a standard deviation of $1/2$. This ensures that the method reflects the widely-held default prior belief that higher order interactions are likely to have a smaller contribution to the linear predictor. The combination of $\mathcal{P}_G$ with this scaling was shown to have improved predictive performance relative to both maximum likelihood and penalized logistic regression. A reasonable analogue of the above method for designed experiments would be to combine $\mathcal{P}_G$ with a standardization of the design variables to have range $[-1/2, 1/2]$. This achieves a similar penalization on higher order interactions.

It is possible to obtain a partial converse to Theorem 1.

Proposition 4. *Given $j \in \{0, \ldots, p-1\}$, suppose that:*

(i) $\mathcal{P}$ *is such that* $\Pr(\beta_j > 1) > 0$
(ii) $\mathcal{P}$ *is such that, for any* $\delta > 0$, *and any* $k = 0, \ldots, p-1$ *with* $k \neq j$, *we have that* $\Pr(|\beta_k| < \delta) > 0$
(iii) $\mathcal{P}$ *is such that* $\beta_0, \ldots, \beta_{p-1}$ *are independent*
(iv) $\mathcal{P}$ *is such that* $E_{\mathcal{P}}[\beta_j \mid \beta_j > 1] = \infty$
(v) ξ *is such that* $\min_{i=1,\ldots,n} |f_j(x_i)| > 0$.

Then $\phi(\xi; \mathcal{P}) = -\infty$, i.e. ξ is Bayesian singular with respect to $\mathcal{P}$.

The Gelman prior distribution, $\mathcal{P}_G$, is such that

$$\beta_0 = 10C_0\,, \qquad \beta_j = (5/2)C_j\,, \quad j = 1, \ldots, p-1\,,$$

where $C_0, \ldots C_{p-1}$ are independent standard Cauchy random variables, which have undefined mean. For a model with an intercept term, $f_0(x) = 1$, and we may apply Proposition 4 with $j = 0$ to find the following:

Corollary 2. *For a logistic model with an intercept term, the prior distribution $\mathcal{P}_G$ is singular.*

For logistic models with a single controllable variable, scalar x, Bayesian D-optimal design has also been studied for a different parameterization (for example, Chaloner and Larntz (1989)):

$$h(\pi_i) = \beta_1(x - \mu)\,, \tag{10}$$

which can be obtained from (7) via $\beta_0 = -\beta_1\mu$. The following result, which is straightforward to prove, provides sufficient conditions for a prior distribution to be non-singular for this form of the model.

Proposition 5. *For the (μ, β_1)-parameterization in (10), if (i) $E_{\mathcal{P}}(|\mu\beta_1|) < \infty$, (ii) $E_{\mathcal{P}}(|\beta_1|) < \infty$ and (iii) $E_{\mathcal{P}}(\log|\beta_1|) > -\infty$, then any design with two or more support points is Bayesian non-singular with respect to $\mathcal{P}$. Hence (i)–(iii) are sufficient for $\mathcal{P}$ to be non-singular. In this case, ξ is Bayesian D-optimal for (β_0, β_1) if and only if it is Bayesian D-optimal for (μ, β_1).*

3 Numerical methods to overcome ill-conditioning

3.1 Objective function bounds for logistic regression

When performing a numerical search for Bayesian D-optimal designs it is necessary to approximate the objective function, usually via a weighted sum,

$$\phi(\xi;\mathcal{P}) \approx \phi(\xi;\mathcal{Q}) = \sum_{l=1}^{N_\mathcal{Q}} v_l \log|M(\xi;\beta^{(l)})|\,, \tag{11}$$

over a weighted sample,

$$\mathcal{Q} = \left\{ \begin{array}{ccc} \beta^{(1)} & \ldots & \beta^{(N_\mathcal{Q})} \\ v_1 & \ldots & v_{N_\mathcal{Q}} \end{array} \right\},$$

of parameter vectors, $\beta^{(l)}$, $l = 1,\ldots,N_\mathcal{Q}$, with corresponding integration weights v_l, satisfying $\sum_{l=1}^{N_\mathcal{Q}} v_l = 1$.

The sample $\mathcal{Q}$ may be obtained, for example, by space-filling criteria, as used by Woods, Lewis, Eccleston and Russell (2006), Latin hypercube sampling, or a quadrature scheme, such as that applied by Gotwalt, Jones and Steinberg (2009). Quadrature methods, and in particular the Gotwalt method, can often yield highly accurate approximations.

A problem with approximation (11), that occurs even for non-singular $\mathcal{P}$, is that for multi-parameter models numerical evaluation of $\phi(\xi;\mathcal{Q})$ can fail due to ill-conditioning in one or more of the matrices $M(\xi;\beta^{(l)})$. When this happens for all $\xi \in \Xi$, we say that $\mathcal{Q}$ is an *ill-conditioned quadrature scheme.* For logistic regression, ill-conditioning of $M(\xi;\beta)$ often occurs when some of the parameters are large. Thus, for prior distributions with unbounded support, ill-conditioning of $\mathcal{Q}$ is made more likely by: (i) choice of a large $N_\mathcal{Q}$, needed for $\phi(\xi;\mathcal{Q})$ to be an accurate approximation to $\phi(\xi;\mathcal{P})$; and (ii) choice of a quadrature method, e.g. the Gotwalt method, that oversamples the tails of the distribution for β.

For some important models, it is possible to obtain bounds that allow approximation of $\phi(\xi;\mathcal{Q})$ when $\mathcal{Q}$ is ill-conditioned. We focus on the case of logistic regression, but the results of Lemma 1 can be used in a similar way for the compartmental model. Using Lemma 3 and (9), we see that $\phi(\xi;\beta) = \log|M(\xi;\beta)|$ lies in the interval $[\phi_L(\xi;\beta), \phi_U(\xi;\beta)]$, where

$$\begin{aligned}\phi_L(\xi;\beta) &= \log|F^\mathrm{T}F| + p \min_{i=1,\ldots,n}\{-|\eta_i| + 2\log\operatorname{expit}|\eta_i|\} \\ \phi_U(\xi;\beta) &= \log|F^\mathrm{T}F| + p \max_{i=1,\ldots,n}\{-|\eta_i| + 2\log\operatorname{expit}|\eta_i|\}\,.\end{aligned}$$

Let $\mathcal{S}$ be the set of l in $\{1,\ldots,N_\mathcal{Q}\}$ for which $M(\xi;\beta^{(l)})$ is ill-conditioned, then:

$$\phi_L(\xi;\mathcal{Q}) \le \phi(\xi;\mathcal{Q}) \le \phi_U(\xi;\mathcal{Q})\,, \tag{12}$$

where

$$\begin{aligned}\phi_L(\xi;\mathcal{Q}) = &\sum_{l\in\{1,\ldots,N_\mathcal{Q}\}\setminus\mathcal{S}} v_l \log|M(\xi;\beta^{(l)})| + \sum_{l\in\mathcal{S}} v_l \log|F^\mathrm{T}F| \\ &+ \sum_{l\in\mathcal{S}} v_l\, p \min_{i=1,\ldots,n}\{-|f^\mathrm{T}(x_i)\beta^{(l)}| + 2\log\operatorname{expit}|f^\mathrm{T}(x_i)\beta^{(l)}|\} \\ \phi_U(\xi;\mathcal{Q}) = &\sum_{l\in\{1,\ldots,N_\mathcal{Q}\}\setminus\mathcal{S}} v_l \log|M(\xi;\beta^{(l)})| + \sum_{l\in\mathcal{S}} v_l \log|F^\mathrm{T}F| \\ &+ \sum_{l\in\mathcal{S}} v_l\, p \max_{i=1,\ldots,n}\{-|f^\mathrm{T}(x_i)\beta^{(l)}| + 2\log\operatorname{expit}|f^\mathrm{T}(x_i)\beta^{(l)}|\}\,.\end{aligned}$$

The bounds $\phi_L(\xi;\mathcal{Q})$, $\phi_U(\xi;\mathcal{Q})$ are much better conditioned than $\phi(\xi;\mathcal{Q})$. The bounds for $\log|M(\xi;\beta^{(l)})|$, $l\in\mathcal{S}$, are often wide. However, as the corresponding v_l is often very small, we may nonetheless obtain from (12) a relatively narrow interval for $\phi(\xi;\mathcal{Q})$. Note that (12) specifies an interval that contains the approximation $\phi(\xi;\mathcal{Q})$, and not necessarily the value of $\phi(\xi;\mathcal{P})$.

In the remainder of Section 3, we use the following example to show how the bounds enable an extension of the set of prior distributions for which Bayesian D-efficient designs can be obtained. We begin by illustrating the use of bounds for the objective function.

Example 1. Potato-packing experiment (Woods, Lewis, Eccleston and Russell (2006)). We use one of the models, defined by

$$\begin{aligned} f(x) &= (1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3)^{\mathrm{T}} \\ \beta &= (\beta_0, \beta_1, \beta_2, \beta_3, \beta_{12}, \beta_{13}, \beta_{23})^{\mathrm{T}}\,, \end{aligned}$$

where $q=3$, $x=(x_1,x_2,x_3)^{\mathrm{T}}$. We adopt a different prior distribution, namely $\log\beta_0\sim N(-1,2)$, $\beta_1\sim N(2,2)$, $\beta_2\sim N(1,2)$, $\beta_3\sim N(-1,2)$, and $\beta_{12},\beta_{13},\beta_{23}\sim N(0.5,2)$ independently. From Theorem 1, this prior distribution is non-singular.

For a double-replicate of the 2^3 full factorial design, the value of $\phi(\xi;\mathcal{P})$ was approximated using the Gotwalt quadrature scheme, with 5 radial points and 4 random rotations. Direct numerical evaluation of $\phi(\xi;\mathcal{Q})$ failed, since $|\mathcal{S}|=39$. However, we have that $\phi(\xi;\mathcal{Q})\in[-6.85,-6.78]$ using (12).

3.2 Use of bounds in design optimization and assessment

We can also use the bounds from (12) within an optimization algorithm to help find Bayesian D-efficient designs. The *Bayesian D-efficiency* is

$$\text{Bayes-eff}(\xi;\mathcal{P}) = \exp\{[\phi(\xi;\mathcal{P}) - \phi(\xi^*_{\mathcal{P}};\mathcal{P})]/p\}\times 100\%\,,$$

where $\xi^*_{\mathcal{P}}\in\arg\max_\xi\phi(\xi;\mathcal{P})$ is a Bayesian D-optimal design. Bayesian D-efficiencies near 100% indicate that ξ achieves a near-optimal trade-off in performance for different β.

When $\mathcal{Q}$ is well-conditioned, the Bayesian D-efficiency may be approximated by numerically searching for $\xi^*_{\mathcal{Q}}\in\arg\max_\xi\phi(\xi;\mathcal{Q})$ maximizing the quadrature approximated objective function, and substituting the design found into

$$\text{Bayes-eff}(\xi;\mathcal{Q}) = \exp\{[\phi(\xi;\mathcal{Q}) - \phi(\xi^*_{\mathcal{Q}};\mathcal{Q})]/p\}\times 100\%\,.$$

However, if $\mathcal{Q}$ is ill-conditioned then this method fails since (i) $\phi(\xi;\mathcal{Q})$ cannot be evaluated directly, and (ii) $\xi^*_{\mathcal{Q}}$ cannot be found using a numerical search. We may nonetheless use numerical methods to find designs $\xi^*_{\mathcal{Q},L}$ and $\xi^*_{\mathcal{Q},U}$ maximizing the lower and upper bounds respectively, i.e. $\xi^*_{\mathcal{Q},L}\in\arg\max_\xi\phi_L(\xi;\mathcal{Q})$ and $\xi^*_{\mathcal{Q},U}\in\arg\max_\xi\phi_U(\xi;\mathcal{Q})$. Then a lower bound for the Bayesian efficiency of $\xi^*_{\mathcal{Q},L}$ can be approximated, via substitution of the designs found into the inequality,

$$\text{Bayes-eff}(\xi^*_{\mathcal{Q},L};\mathcal{Q}) \geq \exp\{[\phi_L(\xi^*_{\mathcal{Q},L};\mathcal{Q}) - \phi_U(\xi^*_{\mathcal{Q},U};\mathcal{Q})]/p\}\times 100\%\,. \tag{13}$$

To find exact designs maximizing the bounds we use a continuous co-ordinate exchange algorithm similar to that of Gotwalt, Jones and Steinberg (2009).

Example 1 (continued)**.** A co-ordinate exchange algorithm was used, with 100 random starts, to search for $\xi^*_{\mathcal{Q},L}$, $\xi^*_{\mathcal{Q},U}$ among exact designs with $n=16$ runs. The quadrature scheme $\mathcal{Q}$ was generated using the Gotwalt method, with 3 radial points, and one random rotation, yielding a total of 217 support points for $\mathcal{Q}$. The design $\xi^*_{\mathcal{Q},L}$, given in Table 1, was found to have $\text{Bayes-eff}(\xi^*_{\mathcal{Q},L};\mathcal{Q})\gtrsim 99.4\%$.

Run	x_1	x_2	x_3	Run	x_1	x_2	x_3
1	0.456	1.000	1.000	9	-1.000	-1.000	1.000
2	-1.000	-1.000	-1.000	10	-0.269	1.000	1.000
3	-1.000	0.512	-1.000	11	1.000	-1.000	-1.000
4	-0.137	-1.000	-1.000	12	1.000	-1.000	0.045
5	1.000	-1.000	1.000	13	-1.000	-1.000	-0.124
6	1.000	1.000	-1.000	14	0.085	-1.000	1.000
7	1.000	-0.038	1.000	15	-1.000	1.000	-0.213
8	-1.000	1.000	1.000	16	-0.149	1.000	-1.000

Table 1: Example 1, Bayesian design, $\xi^*_{\mathcal{Q},L}$, maximizing the lower bound $\phi_L(\xi;\mathcal{Q})$.

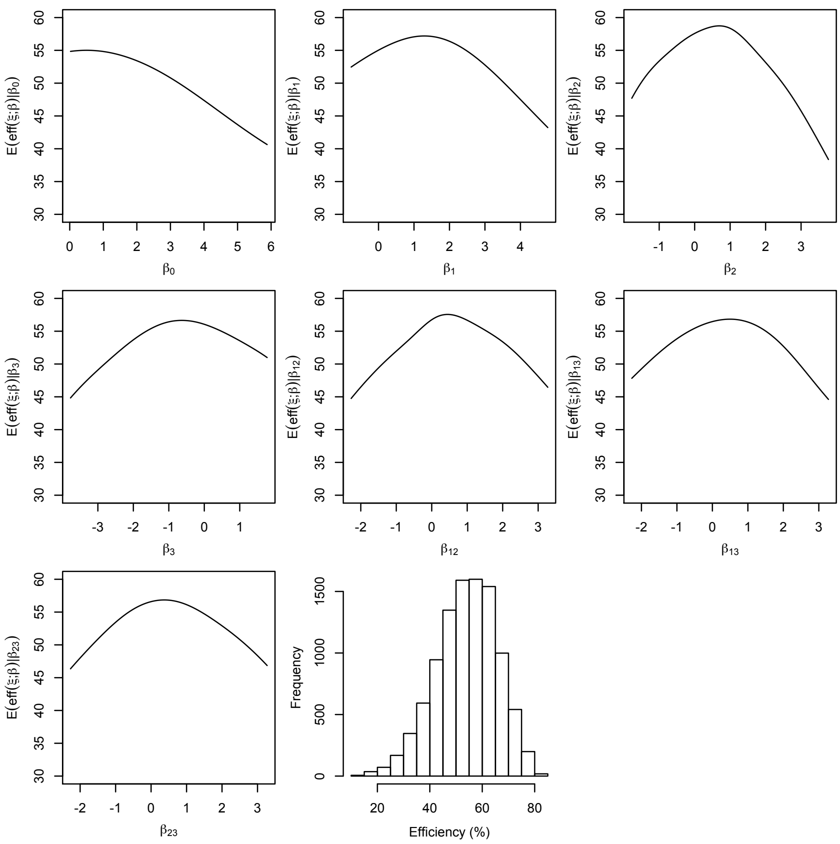

Figure 1: Kriging-approximated conditional mean efficiencies, $E_{\mathcal{P}}\{\mathrm{eff}(\xi^*_{\mathcal{Q},L};\beta)\,|\,\beta_j\}$, for all parameters, of the Bayesian design, $\xi^*_{\mathcal{Q},L}$, maximizing the lower bound (13). The lower and upper limits of the β-axes correspond to the 2.5% and 97.5% prior quantiles, respectively. The histogram shows an approximate sample from the local efficiency distribution for $\xi^*_{\mathcal{Q},L}$ induced by the prior distribution on β.

Note that the computation of the lower bound is approximate since we cannot be certain to have found the global optimum $\xi^*_{\mathcal{Q},U}$, although in the above example an assessment of the objective function values from the different random initializations of the algorithm suggests that the number of starts is adequate.

To assess the performance of a given design, ξ, for different β, we use the local D-efficiency,

$$\mathrm{eff}(\xi;\beta) = \{|M(\xi;\beta)|/|M(\xi^*_\beta;\beta)|\}^{1/p}\,. \tag{14}$$

For some β, $M(\xi;\beta)$ is well-conditioned for most $\xi \in \Xi$. In this case, the local D-efficiency can be approximated by searching numerically for the locally D-optimal design, ξ^*_β, and substituting the design found into (14). For other β, $M(\xi;\beta)$ is ill-conditioned for all $\xi \in \Xi$. Then, approximate bounds on the efficiency can be derived by numerically searching for the designs $\xi^*_{U,\beta} \in \arg\max_\xi \phi_U(\xi;\beta)$ and $\xi^*_{L,\beta} \in \arg\max_\xi \phi_L(\xi;\beta)$, and using the fact that

$$\exp\frac{1}{p}[\phi_L(\xi;\beta) - \phi_U(\xi^*_{U,\beta};\beta)] \le \mathrm{eff}(\xi;\beta) \le \exp\frac{1}{p}[\phi_U(\xi;\beta) - \phi_L(\xi^*_{L,\beta};\beta)]\,. \tag{15}$$

To visualize the dependence of the local efficiency on the individual parameters, we plot approximations to the conditional means, $E_{\mathcal{P}}\{\mathrm{eff}(\xi;\beta)\,|\,\beta_j\}$, as univariate functions of each of the regression coefficients, β_j. Owing to the need to search for a locally D-optimal design, evaluation of $\mathrm{eff}(\xi;\beta)$ is computationally intensive. Thus, before computing conditional means it is advantageous to first build a statistical emulator of $\mathrm{eff}(\xi;\beta)$ as a function of β, using Gaussian process interpolation. This is analogous to the approach followed in the computer experiments literature when visualizing the main effects of a computationally expensive simulator (e.g. Santner, Williams and Notz (2003, Ch.7)). A similar method was used by Waite and Woods (2015) to visualize the efficiency profile of Bayesian designs for logistic models with random effects.

Example 1 (continued)**.** We consider further the performance of the design, $\xi^*_{\mathcal{Q},L}$, maximizing the lower bound for $\phi(\xi;\mathcal{Q})$. We use the support points of the quadrature scheme to train our emulator of $\mathrm{eff}(\xi^*_{\mathcal{Q},L};\beta)$. In our example, only three out of the 217 β vectors in $\mathcal{Q}$ led to $M(\xi;\beta)$ being ill-conditioned for all $\xi \in \Xi$. For these vectors, the efficiency bounds in (15) gave no additional information beyond $\mathrm{eff}(\xi^*_{\mathcal{Q},L};\beta) \in [0\%, 100\%]$. Thus we decided to omit these β vectors from the training set, as including the bounds $[0\%, 100\%]$ would not substantively reduce our uncertainty about the efficiency at these β. Figure 1 shows the approximations to the conditional means, $E_{\mathcal{P}}\{\mathrm{eff}(\xi^*_{\mathcal{Q},L};\beta)\,|\,\beta_j\}$, resulting from application of the Gotwalt integration method in $p-1$ dimensions (with 5 radial abscissae and one random rotation) to integrate the mean of the Kriging emulator with respect to all parameters except β_j. Also shown is a histogram giving an approximation to the distribution of local efficiencies of $\xi^*_{\mathcal{Q},L}$ induced by the prior distribution on β. This is derived by computing the Kriging-based estimates of $\mathrm{eff}(\xi^*_{\mathcal{Q},L};\beta)$ for a Monte Carlo sample of 10,000 β vectors from the prior distribution. From Figure 1 it appears that the modal efficiency is in the range 55-60%. The lower and upper quartiles of the efficiency distribution are approximately 46% and 62%. Overall, the design appears moderately robust to likely β, despite the possibility of very large β_0.

4 Discussion

One of the best possible situations for (pseudo-)Bayesian design is when $\mathcal{P}$ is non-singular. In this case we may proceed to find Bayesian D-optimal designs using standard methods, or if the quadrature scheme is ill-conditioned, using bounds such as those developed for logistic regression in Section 3. We may also apply these methods in the case where $\mathcal{P}$ is singular, but can be replaced with an alternative non-singular $\mathcal{P}'$ that plausibly represents our prior uncertainty. However, we should in general avoid selecting prior distributions for analytical convenience if they do not accurately represent our prior beliefs, and so if no such $\mathcal{P}'$ exists, neither $\phi(\xi;\mathcal{P})$ nor $\phi(\xi;\mathcal{P}')$ can be used to help guide the choice of design. In this case, we must consider different criteria for design selection.

One alternative approach is to select ξ to maximize the *mean local efficiency*,

$$\Psi(\xi;\mathcal{P}) = E_{\mathcal{P}}\{\mathrm{eff}(\xi;\theta)\}\,,$$

which is much less sensitive to the presence of θ with $|M(\xi;\theta)| \approx 0$. This is a special case (Φ_1 in their notation) of the optimality criterion discussed by Dette and Wong (1996). The above criterion does not have the interpretation of being approximately equivalent to maximizing Shannon information gain. As an example of the use of this criterion, consider the exponential decay model from Section 2.1. From Corollary 1, when $\mathcal{P} = U(0,a)$, $a > 0$, all designs are Bayesian singular with respect to $\phi(\xi;\mathcal{P})$ for θ-parameterization. By contrast, it is shown trivially that the design with a single support point $x = a/2$ is Ψ-optimal with a mean efficiency of approximately 67%. This design is locally D-optimal when θ is equal to its prior mean, but highly inefficient when θ is very small. Thus, Ψ-optimal designs are much less strongly driven by their worst-case behaviour.

Acknowledgement

I am grateful to Professors David Woods and Susan Lewis (University of Southampton) for several invaluable discussions and comments. This work was supported by funding from the UK Engineering and Physical Sciences Research Council, via a PhD studentship, Doctoral Prize, and a project grant. The work made use of the Iridis computational cluster at the University of Southampton.

A Appendix. Proofs of analytical results

Proof of Proposition 2. Assume that at least one $x_i > 0$. For the θ parameterization, we demonstrate two implications. Namely (i) if $E_{\mathcal{P}}(1/\theta) < \infty$ and $E_{\mathcal{P}}(\log\theta) < \infty$, then $\phi(\xi;\mathcal{P}) > -\infty$; and (ii) if $E_{\mathcal{P}}(\log\theta) = \infty$ or $E_{\mathcal{P}}(1/\theta) = \infty$, then $\phi(\xi;\mathcal{P}) = -\infty$. Here, $\phi(\xi;\mathcal{P}) = E\{\log|M_\theta(\xi;\theta)|\}$, where $\log|M_\theta(\xi;\theta)|$ is given by (4).

For (i), observe that $-\infty \leq E_{\mathcal{P}}\left\{(2/\theta)\max_{i=1,\ldots,n}\{x_i\} + 4\log\theta\right\} < \infty$, and so considering the left hand side of (3), and the reparameterization (4), we have

$$-\infty < \log\sum_{i=1}^n x_i^2 - E_{\mathcal{P}}\left\{(2/\theta)\max_{i=1,\ldots,n}\{x_i\} + 4\log\theta\right\} \leq \phi(\xi;\mathcal{P})\,,$$

as required.

For (ii), note that in addition to (3), the following weaker inequality holds:

$$\phi(\xi;\mathcal{P}) \leq \log\sum_{i=1}^n x_i^2 - 4\log\theta\,.$$

Taking expectations of both sides, if $E_{\mathcal{P}}(\log\theta) = \infty$ then $\phi(\xi;\mathcal{P}) = -\infty$.

For the other case, first let

$$b(\theta) = \frac{1}{\theta}\left\{2\min_{i=1,\ldots,n}\{x_i : x_i > 0\} + 4\theta\log\theta\right\}\,.$$

Note that $\theta \log \theta \to 0$ as $\theta \to 0$, and so there is some $\delta > 0$ such that, for $\theta < \delta$,

$$b(\theta) \geq (1/\theta) \min_{i=1,\dots,n} \{x_i : x_i > 0\} .$$

The expectation of $b(\theta)$ therefore satisfies

$$\begin{aligned} E_{\mathcal{P}}\{b(\theta)\} &\geq E_{\mathcal{P}}\{b(\theta)\mathbb{I}(\theta < \delta) + \inf_{\theta > \delta} b(\theta)\, \mathbb{I}(\theta > \delta)\} \\ &\geq \min_{i=1,\dots,n} \{x_i : x_i > 0\} E_{\mathcal{P}}\{(1/\theta)\mathbb{I}(\theta < \delta)\} + (4 \log \delta) \Pr(\theta > \delta) \end{aligned} \tag{16}$$

If $E_{\mathcal{P}}(1/\theta) = \infty$, then $E_{\mathcal{P}}\{(1/\theta)\mathbb{I}(\theta < \delta)\} = \infty$, and so by (16), $E_{\mathcal{P}}\{b(\theta)\} = \infty$, regardless of whether $E_{\mathcal{P}}(\log \theta) = -\infty$. Recall from (3) that

$$\phi(\xi; \mathcal{P}) \leq \log \sum_{i=1}^{n} x_i^2 - E_{\mathcal{P}}\{b(\theta)\} .$$

Hence if $E_{\mathcal{P}}(1/\theta) = \infty$, we have $\phi(\xi; \mathcal{P}) = -\infty$. This is sufficient to establish the proposition. $\square$

Proof of Lemma 1. Observe that $M(x_i; \theta) = e^{-2\theta_1 x_i} M^{(i)}_{\delta,\theta_3}$. Moreover, for $i = 1, \dots, n$, either (i) $x_i = 0$ or (ii) $x_i \geq x_{\min}$. In case (ii), we have

$$e^{-2\theta_1 x_{\max}} \tilde{M}^{(i)}_{\delta,\theta_3} \preceq M(x_i; \theta) \preceq e^{-2\theta_1 x_{\min}} \tilde{M}^{(i)}_{\delta,\theta_3} . \tag{17}$$

Moreover, the above holds also in case (i) since then $M(x_i; \theta)$ and $M^{(i)}_{\delta,\theta_3}$ are matrices of zeroes. Summing (17) over $i = 1, \dots, n$, we obtain:

$$e^{-2\theta_1 x_{\max}} \tilde{M}_{\delta,\theta_3} \preceq M(\xi; \theta) \preceq e^{-2\theta_1 x_{\min}} \tilde{M}_{\delta,\theta_3} . \tag{18}$$

Taking log-determinants of all sides of (18) yields the result, when combined with the fact that $|\tilde{M}_{\delta,\theta_3}| = \theta_3^4 |\tilde{M}_{\delta,1}|$. $\square$

Define $g_\xi(\delta) = |\tilde{M}_{\delta,1}|$. The following is needed to establish Lemma 2.

Lemma 4. *Suppose that ξ contains at least three distinct $x_i > 0$. Then the derivatives of $g_\xi(\delta)$ satisfy: (i) $g_\xi^{(k)}(0) = 0$, $k = 1, \dots, 7$, (ii) $g_\xi^{(8)}(0) > 0$.*

Proof of Lemma 4. Part (i) can be verified using symbolic computation, e.g. using Mathematica. It can also be shown that

$$g_\xi^{(8)}(0) = 280\{S_2 S_4 S_6 - S_2 S_5^2 - S_3^2 S_6 + S_3 S_4 S_5 + S_3 S_4 S_5 - S_4^3\} ,$$

where $S_l = \sum_{i=1}^{n} x_i^l$. Define the following,

$$K = \begin{pmatrix} S_2 & S_3 & S_4 \\ S_3 & S_4 & S_5 \\ S_4 & S_5 & S_6 \end{pmatrix} , \qquad K' = \sum_{i : x_i > 0} \begin{pmatrix} 1 & x_i & x_i^2 \\ x_i & x_i^2 & x_i^3 \\ x_i^2 & x_i^3 & x_i^4 \end{pmatrix} ,$$

and $x_{\min} = \min\{x_i : x_i > 0\}$. Note

$$K \succeq x_{\min}^2 K' .$$

We have

$$g_\xi^{(8)}(0) = 280|K| \geq 280 x_{\min}^6 |K'| .$$

Observe also that K' is the information matrix of the design $\xi' = (x_i : x_i > 0)$ under the linear model with regressors $1, x, x^2$. By the assumption that there are at least three distinct $x_i > 0$, the above linear model is estimable and so $|K'| > 0$. This establishes part (ii). $\square$

Proof of Lemma 2. In the case that ξ contains fewer than three distinct $x_i > 0$, then $M_{\delta,1}$ has rank at most two, and so $E_{\mathcal{P}}(\log|M_{\delta,1}|) = -\infty$ for any prior $\mathcal{P}$. Thus we may assume that ξ has at least three distinct $x_i > 0$. From Lemma 4 is is clear that, for small δ, we have $g_{\xi}(\delta) \approx (\kappa/2)\delta^8$, where $\kappa > 0$. We show that the approximation is sufficiently close that $E_{\mathcal{P}}(\log|\tilde{M}_{\delta,1}|) = -\infty$ if $\int_{\delta<1} \log\delta\, d\mathcal{P}(\theta) = -\infty$. By Taylor's theorem, there is $\epsilon_1 > 0$ and $\lambda > 0$ such that, for $\delta \in (0, \epsilon_1)$,

$$|g(\delta) - (\kappa/2)\delta^8| \leq \lambda\delta^9\,.$$

Hence, for $\delta \in (0, \epsilon_1)$,

$$|2g(\delta)/(\delta^8\kappa) - 1| \leq (2\lambda/\kappa)\delta^9\,.$$

As the logarithm function has derivative 1 at argument 1, there exists $0 < \epsilon_2 \leq \epsilon_1$ such that for $\delta \in (0, \epsilon_2)$,

$$\left|\log\frac{2g(\delta)}{\delta^8\kappa} - \log 1\right| \leq 2|2g(\delta)/(\delta^8\kappa) - 1| \leq (4\lambda/\kappa)\delta^9\,.$$

Thus, for $\delta \in (0, \epsilon_2)$, we have the following approximation of $\log g(\delta)$,

$$|\log g(\delta) - \log(\kappa\delta^8/2)| \leq (4\lambda/\kappa)\delta^9\,,$$

so that

$$\left|\int_{\delta<\epsilon_2} \log g(\delta) d\mathcal{P}(\theta) - \int_{\delta<\epsilon_2} \{8\log\delta + \log(\kappa/2)\} d\mathcal{P}(\theta)\right| \leq (4\lambda/\kappa)\epsilon_2^9\,.$$

Hence it is clear that $\int_{\delta<\epsilon_2} \log g(\delta) d\mathcal{P}(\theta) = -\infty$ if and only if $\int_{\delta<\epsilon_2} \log\delta\, d\mathcal{P}(\theta) = -\infty$. Moreover note that $g_{\xi}(\delta)$ is bounded above, and

$$\int \log g(\delta) d\mathcal{P} = \int_{\delta<\epsilon_2} \log g(\delta) d\mathcal{P} + \int_{\delta>\epsilon_2} \log g(\delta) d\mathcal{P}\,.$$

Thus, $\int \log g(\delta) d\mathcal{P} = -\infty$ if $\int_{\delta<\epsilon_2} \log\delta\, d\mathcal{P}(\theta) = -\infty$. The result is finally established by observing that $\int_{\delta<\epsilon_2} \log\delta\, d\mathcal{P}(\theta) = -\infty$ if $\int_{\delta<1} \log\delta\, d\mathcal{P}(\theta) = -\infty$.

□

Proof of Proposition 3. Suppose that ξ has at least three distinct $x_i > 0$. From Lemmas 1 and 2 it is also clear that if $\int_{\theta_3>1} \log\theta_3\, d\mathcal{P}(\theta) < \infty$, and in addition $\int_{\theta_3<1} \log\theta_3\, d\mathcal{P}(\theta) = -\infty$, $\int_{\delta<1} \log\delta\, d\mathcal{P}(\theta) = -\infty$, or $E_{\mathcal{P}}(\theta_1) = \infty$, then also $E_{\mathcal{P}}\{\log|M(\xi;\theta)|\} = -\infty$. This establishes the result.

□

Proof of Theorem 1. Using the Loewner bounds on the information matrix, and the fact that taking determinants respects the Loewner partial ordering (i.e. if $M_1 \preceq M_2$ then $|M_1| \leq |M_2|$), we have from Lemma 3 that

$$\log|M(\xi;\beta)| \geq \log|F^{\mathrm{T}}F| + p\min_i \log w_i\,.$$

From (9) it is clear that $w(\eta) \geq (1/4)e^{-|\eta|}$. Thus,

$$\begin{aligned}\log|M(\xi;\beta)| &\geq \log|F^{\mathrm{T}}F| + p\log\left[(1/4)e^{-\max_i|\eta_i|}\right]\\ &\geq \log|F^{\mathrm{T}}F| - p\max_i|\eta_i| - p\log 4\,.\end{aligned}$$

Moreover, by the triangle inequality, $\max_i|\eta_i| \leq \sum_j \max_i|f_j(x_i)||\beta_j|$, and so

$$\log|M(\xi;\beta)| \geq \log|F^{\mathrm{T}}F| - p\log 4 - p\sum_j \max_i|f_j(x_i)||\beta_j|\,. \tag{19}$$

The right hand side of (19) has expectation greater than $-\infty$ due to the assumptions that $E_{\mathcal{P}}(|\beta_j|) < \infty$, $|F^{\mathrm{T}}F| > 0$, therefore $E_{\mathcal{P}}\{\log|M(\xi;\beta)|\} > -\infty$. □

Proof of Proposition 4. From Lemma 3 we know that

$$\log|M(\xi;\beta)| \leq \log|F^{\mathrm{T}}F| + p\max_i \log w_i$$

It can also be shown that $w(\eta)$ is a decreasing function of $|\eta|$, and from (9) it is clear that $w(|\eta|) \leq \exp(-|\eta|)$, so

$$\begin{aligned}\log|M(\xi;\beta)| &\leq \log|F^{\mathrm{T}}F| + p\log w(\min_i|\eta_i|)\\ &\leq \log|F^{\mathrm{T}}F| - p\min_i|\eta_i|\end{aligned}$$

Now we need only prove $E_{\mathcal{P}}(\min_i|\eta_i|) = \infty$ to establish that $E_{\mathcal{P}}\{\log|M(\xi;\beta)|\} = -\infty$. To show this, we condition on an event where the parameter β_j dominates.

Given $j \in \{0,\ldots,p-1\}$, let $E \in \Sigma$ be an event such that (a) $\beta_j > 1$, and (b) $\sum_{k\neq j}|f_k(x_i)||\beta_k| < \epsilon$ for all i, where $\epsilon > 0$ is such that

$$||f_j(x_i)| - |f_j(x_{i'})|| > 2\epsilon \qquad \text{for any } i, i' \text{ with } |f_j(x_i)| \neq |f_j(x_{i'})|\,.$$

We can guarantee (a) and (b), for example by taking

$$E = \{\beta : \beta_j > 1,\, |\beta_k| < \epsilon/[(p-1)\max_i|f_k(x_i)|]\,,\ \text{for } k \neq j\} \in \Sigma\,,$$

which satisfies $\Pr(E) > 0$, by assumptions (i)–(iii) of the proposition.

By standard properties of the modulus under addition, on event E,

$$||\eta_i| - |f_j(x_i)|\beta_j| \leq \sum_{k\neq j}|f_k(x_i)||\beta_k| \leq \epsilon\,, \qquad \text{by (b)} \tag{20}$$

Since on E the term from β_j dominates, to find the minimum of $|\eta_i|$ we just need to minimize the β_j term. To see this formally, observe that if $|f_j(x_i)|\beta_j > |f_j(x_{i'})|\beta_j$ then by (b)

$$|f_j(x_i)|\beta_j - |f_j(x_{i'})|\beta_j > 2\epsilon\beta_j > 2\epsilon\,,$$

and so, by (20),

$$|\eta_{i'}| < |f_j(x_{i'})|\beta_j + \epsilon < |f_j(x_i)|\beta_j - \epsilon < |\eta_i|\,.$$

Thus $|f_j(x_i)|\beta_j > |f_j(x_{i'})|\beta_j$ implies $|\eta_i| > |\eta_{i'}|$. Hence

$$\begin{aligned}\min_i|\eta_i| &= |\eta_{i^*}|\,, \quad i^* \in \arg\min_i|f_j(x_i)|\\ &\geq |f_j(x_{i^*})|\beta_j - \epsilon\,.\end{aligned}$$

Consequently,

$$\begin{aligned}E_{\mathcal{P}}(\min_i|\eta_i| \mid E) &\geq |f_j(x_{i^*})|\,E_{\mathcal{P}}(\beta_j \mid E) - \epsilon\\ &\geq |f_j(x_{i^*})|\,E_{\mathcal{P}}(\beta_j \mid \beta_j > 1) - \epsilon\,,\ \text{by assumption (iii) of the proposition}\\ &= \infty \quad \text{by assumptions (iv) and (v) of the proposition.}\end{aligned}$$

Considering the marginal expectation, note that $\Pr(E) > 0$, and so,

$$E_{\mathcal{P}}(\min_i|\eta_i|) \geq \Pr(E)E_{\mathcal{P}}(\min_i|\eta_i| \mid E) = \infty\,.$$

$\square$

References

Chaloner, K. and Larntz, K. (1989), 'Optimal Bayesian design applied to logistic regression experiments', *J. Statist. Plann. Infer.* **21**, 191–208.

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

Dette, H. and Wong, W. K. (1996), 'Optimal Bayesian designs for models with partially specified heteroscedastic structure', *Ann. Statist.* **24**, 2108–2127.

Gelman, A., Bois, F. and Jiang, J. (1996), 'Physiological pharmacokinetic analysis using population modeling and informative prior distributions', *J. Amer. Statist. Assoc.* **91**, 1400–1412.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008), 'A weakly informative default prior distribution for logistic and other regression models', *Ann. Appl. Statist.* **2**, 1360–1383.

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

Jones, B. and Goos, P. (2009), 'D-optimal design of split-split-plot experiments', *Biometrika* **96**, 67–82.

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

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edn, Chapman and Hall, Boca Raton, FL.

Polson, N. G. and Scott, J. G. (2012), 'On the half-Cauchy prior for a global scale parameter', *Bayesian Anal.* **7**, 887–902.

Santner, T. J., Williams, B. J. and Notz, W. (2003), *The design and analysis of computer experiments*, Springer.

Waite, T. W. and Woods, D. C. (2015), 'Designs for generalized linear models with random block effects via information matrix approximations', *Biometrika* (in press).

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

Yang, M. (2010), 'On the de la Garza Phenomenon', *Ann. Statist.* **38**, 2499–2524.

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

School of Mathematics, University of Manchester, Manchester, M13 9PL, U.K.
E-mail: timothy.waite@manchester.ac.uk