Statistica Sinica Preprint No: SS-2015-0293R2	
Title	Singular prior distributions in Bayesian D-optimal design for nonlinear models
Manuscript ID	SS-2015-0293R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0293
Complete List of Authors	Timothy Waite
Corresponding Author	Timothy Waite
E-mail	timothy.waite@manchester.ac.uk

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

Statistica Sinica

SINGULAR PRIOR DISTRIBUTIONS AND ILL-CONDITIONING IN BAYESIAN D-OPTIMAL DESIGN FOR SEVERAL NONLINEAR MODELS

Timothy W. Waite

University of Manchester

Abstract: For Bayesian D-optimal design, we define a *singular prior distribution* for the model parameters as 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 several multi-parameter generalized linear models, we establish sufficient conditions for singularity of a prior distribution. For the generalized linear models we also obtain sufficient conditions for non-singularity. In the existing literature, weakly informative prior distributions are commonly recommended as a default choice for inference in logistic regression. Here it is shown that some of the recommended prior distributions are singular, and hence should not be used for Bayesian D-optimal design. Additionally, methods are developed to derive and assess Bayesian D-efficient designs when numerical evaluation of the objective function fails due to ill-conditioning, as often occurs for heavy-tailed prior distributions. These numerical methods are illustrated for logistic regression.

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

1. Introduction

Much effort has been devoted to the development of D-optimal design methods for nonlinear problems; for example, nonlinear models (e.g. Yang and Stufken (2009, 2012), Yang (2010)), generalized linear models (Khuri, Mukherjee, Sinha, and Ghosh (2006), Woods, Lewis, Eccleston, and Russell (2006), Yang, Zhang, and Huang (2011), Yang and Mandal (2015)), and linear models with mixed effects (Jones and Goos (2009)). In each of these areas, the choice of a D-optimal design depends on the unknown vector of model parameters, $\theta \in \Theta \subseteq \mathbb{R}^p$.

One approach to choosing a design is to make a 'best guess' of the parameter values, and calculate a corresponding locally D-optimal design (Chernoff, 1953), $\xi_\theta^* \in \arg\max_{\xi\in\Xi} |M(\xi;\theta)|$, where $M(\xi;\theta)$ is the Fisher information matrix for design $\xi \in \Xi$, and Ξ is the set of all competing designs. 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

$$\phi(\xi; \mathcal{P}) = \int_{\Theta} \log |M(\xi; \theta)| \, d\mathcal{P}(\theta) \,, \tag{1.1}$$

see, for example, 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 when $g : \Theta \to \mathbb{R}$ is a non-negative Σ-measurable function (Capinski and Kopp (2004)). When $g : \Theta \to \mathbb{R}$ is a general Σ-measurable function, it is said that $\int_{\Theta} g(\theta) \, d\mathcal{P}(\theta) = -\infty$ if and only if $\int_{\Theta} g^{+}(\theta) \, d\mathcal{P}(\theta) < \infty$ and $\int_{\Theta} g^{-}(\theta) \, d\mathcal{P}(\theta) = \infty$, where $g^{+}(\theta) = \max\{0, g(\theta)\}$ and $g^{-}(\theta) = \max\{0, -g(\theta)\}$.

A design that maximizes (1.1) 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 (1.1) is equivalent to maximization of an asymptotic approximation to the Shannon information gain from prior to posterior (Chaloner and Verdinelli (1995)).

In nonlinear problems, a *singular parameter vector* is a θ such that $M(\xi;\theta)$ has determinant zero for *any* design $\xi \in \Xi$. For such 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. The analogue of a singular parameter vector for Bayesian D-optimality is defined as follows:

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

(b) Given a prior distribution, $\mathcal{P}$, $\mathcal{P}$ is a *singular prior distribution* if *all* $\xi \in \Xi$ are Bayesian singular with respect to $\mathcal{P}$. Equivalently, if the geometric mean of $|M(\xi;\theta)|$ under $\mathcal{P}$ is zero for all $\xi \in \Xi$, $\mathcal{P}$ is a singular prior distribution. Here the geometric mean of a non-negative random variable X is $E^{\mathcal{G}}(X) = \exp[E\{\log(X)\}]$, with $E^{\mathcal{G}}(X) = 0$ if $E\log(X) = -\infty$ (Feng et al. (2017)).

For a singular prior distribution $\mathcal{P}$, Bayesian D-optimality can therefore not be used to select a design. In many models, such as the exponential decay model and logistic regression, while it is straightforward to detect singular parameter vectors, θ, by inspection of the information matrix, 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 (1.1). For several models, this is likely to occur for a heavy-tailed prior distribution $\mathcal{P}$, even if $\mathcal{P}$ is theoretically non-singular. Such ill-conditioning causes failure of numerical selection of Bayesian D-optimal designs.

In this paper, we clarify and extend the set of prior distributions for which Bayesian D-optimal design is feasible for three 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 several multi-factor generalized linear models. 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 generalized linear models, sufficient conditions for a prior distribution to be non-singular are established. These conditions are easily checked. In Section 3, methods are developed that enable the selection of highly Bayesian D-efficient designs for logistic regression when the quadrature approximation to (1.1) is ill-conditioned, thereby facilitating design for heavy-tailed prior distributions. In Section 4 we discuss alternative approaches to finding efficient designs when $\mathcal{P}$ is a singular prior

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

distribution.

2. Singularity of prior distributions for some standard models

2.1. Exponential decay model

We derive necessary and sufficient conditions for a prior distribution to be singular for the exponential decay model used, for example, to model the concentration of a chemical compound over time. This model is commonly used as an illustrative example of a nonlinear model in the optimal design of experiments literature, e.g. Dette and Neugebauer (1997), Atkinson (2003). Here, two parameterizations are considered: by rate, $\beta > 0$, and by 'lifetime', $\theta = 1/\beta > 0$. For the former, the model for the response, y, in terms of explanatory variable, $x > 0$, is

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

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

Assume that $\Xi = \mathcal{X}^n$, where $\mathcal{X} = [0, \infty)$. Then design $\xi = (x_1, \ldots, x_n) \in \Xi$ has information matrix

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

If at least one $x_i > 0$ and $S_{xx} = \sum_{i=1}^{n} x_i^2$, then

$$-2\beta \max_{i=1,\ldots,n} \{x_i\} \leq \log |M_\beta(\xi; \beta)| - \log S_{xx} \leq -2\beta \min_{i\,:\,x_i>0} \{x_i\} \,. \qquad (2.1)$$

By taking expectations, we have the following.

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

Here the prior, $\mathcal{P}$, can be singular if the distribution of β is heavy-tailed with infinite mean, e.g. if β is half-Cauchy (cf. Polson and Scott (2012)).

For the θ-parameterization, a change-of-variable argument shows that

$$\log |M_\theta(\xi; \theta)| = \log |M_\beta(\xi; \beta)| - 4 \log \theta \,. \tag{2.2}$$

The proof of the following is in the supplementary material.

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 that maximize $\phi(\xi; \mathcal{P})$ for nonlinear models, Chaloner and Verdinelli (1995) refer to prior distributions with unbounded support with $M(\xi; \theta)$ arbitrarily close to being singular. 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.*

Under the prior for θ in Corollary 1, the corresponding implied prior distribution for β has a proper density, $p(\beta) = 1/(a\beta^2)$ for $\beta \geq 1/a$. How-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

ever, this distribution for β has unbounded support and is heavy tailed, such that $E(\beta) = \infty$.

2.2. Compartmental model

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

$$y_i = \theta_3\{e^{-\theta_1 x_i} - e^{-\theta_2 x_i}\} + \epsilon_i\,, \qquad \epsilon_i \sim N(0, \sigma^2)\,, \tag{2.3}$$

where $x_i \geq 0$, $i = 1, \ldots, n$, $\theta_2 > \theta_1 > 0$, $\theta_3 > 0$ and $\sigma > 0$. Here, $\Xi = [0, \infty)^n$. Here the response y_i may be the concentration of a compound in a system, and the x_i the observation times. For example, Atkinson et al. (1993) consider a theophylline kinetics experiment on horses, finding optimal sampling times for (2.3) under several different (pseudo-)Bayesian criteria.

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 (i) $\theta_1 = \theta_2$ or (ii) $\theta_3 = 0$, and $|M(\xi;\theta)| \to 0$ when (iii) $\theta_1 \to \infty$. Physically, conditions (i) and (iii) correspond to situations where the flow rates in and out of

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

the compartment are either exactly balanced, or both very rapid. Each of these parameter scenarios results in a similar response profile, in which the concentration is close to zero throughout the duration of the experiment; if such a profile is observed, it is difficult to ascertain which values of the parameters generated the data.

Proofs of the following are given in the supplementary material. Let $\delta = \theta_2 - \theta_1 > 0$.

Lemma 1. *We have*

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

where $\tilde{M}_{\delta,1}$ is $\tilde{M}_{\delta,\theta_3}$ evaluated at $\theta_3 = 1$, and

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

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

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 $\int_{\theta_3>1}\log\theta_3\, d\mathcal{P}(\theta) < \infty$. For model (2.3), the prior parameter distribution $\mathcal{P}$ is singular if $E_{\mathcal{P}}(\theta_1) = \infty$, $\int_{\theta_3<1}\log\theta_3\, d\mathcal{P}(\theta) = -\infty$, or $\int_{\delta<1}\log\delta\, d\mathcal{P}(\theta) = -\infty$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

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)). For model (2.3), $\mathcal{P}$ is singular if θ_1 is half-Cauchy distributed, although for physiological compartmental models more specific prior information is often used (Gelman et al. (1996)).

2.3. Generalized linear models

Suppose there are n design points $x_i = (x_{i1}, \ldots, x_{iq})^{\mathrm{T}} \in \mathcal{X}$, with responses y_i, $i = 1, \ldots, n$. We assume a generalized linear model (GLM; McCullagh and Nelder (1989)), thus y_i has an exponential family distribution with mean $\mu_i = \mu(x_i; \beta)$ and variance $\gamma v(\mu_i)$, where μ satisfies

$$h[\mu(x;\beta)] = \eta(x;\beta) = f^{\mathrm{T}}(x)\beta\,, \tag{2.4}$$

with h the link function, γ a dispersion parameter, v the variance function, and $\eta_i = \eta(x_i;\beta)$ the linear predictor. Here $f(x) = (f_0(x), \ldots, f_{p-1}(x))^{\mathrm{T}}$ contains regression functions $f_j : \mathcal{X} \to \mathbb{R}$, $j = 0, \ldots, p-1$, and $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^{\mathrm{T}} \in \Theta$ is a vector of p regression parameters. We let $\mathcal{X} = [-1,1]^q$ and $\Xi = \mathcal{X}^n$.

For design $\xi = (x_1, \ldots, x_n)$ and model (2.4)

$$\begin{aligned} M(\xi;\beta) &= \sum_{i=1}^{n} w_i\, f(x_i) f^{\mathrm{T}}(x_i) \\ w(\eta) &= \frac{1}{\gamma v(\mu)} \left(\frac{\partial \mu}{\partial \eta}\right)^2, \end{aligned} \tag{2.5}$$

with $w_i = w(\eta_i)$, $i = 1, \ldots, n$ (e.g. Khuri et al. (2006), Atkinson and Woods (2015), Yang and Mandal (2015)).

The proofs of the following are straightforward; the details are omitted. Let F be the model matrix with rows $f^{\mathrm{T}}(x_i)$, so $\sum_{i=1}^{n} f(x_i) f^{\mathrm{T}}(x_i) = F^{\mathrm{T}} F$ is the information matrix of ξ under a linear model with regressors specified by f. The Loewner partial ordering on real symmetric matrices has $M_1 \preceq M_2$ if and only if $M_2 - M_1$ is non-negative definite (e.g. Pukelsheim (1993)).

Lemma 3. *For a generalized linear model, the information matrix satisfies*

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

Thus, since the log-determinant respects the Loewner ordering,

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

Lemma 4. *Suppose ξ is non-singular for the linear model with regressors given by f, so $\left|F^{\mathrm{T}} F\right| > 0$.*

(i) *If $E_{\mathcal{P}}\{\log \min_i w_i\} > -\infty$, then $\phi(\xi;\mathcal{P}) > -\infty$.*

(ii) *If $E_{\mathcal{P}}\{\log \max_i w_i\} = -\infty$, then $\phi(\xi;\mathcal{P}) = -\infty$.*

Lemma 4 can often be used to identify clear conditions on the prior distribution that lead to singularity (or non-singularity). However, to do so it is necessary to analyse the tail behaviour of the GLM weight function, $w(\eta)$, as $|\eta| \to \infty$, in order to establish whether (i) or (ii) holds.

2.3.1. Logistic regression

For logistic regression, $y_i \mid \beta \sim \text{Bernoulli}(\pi_i)$, where $\pi_i = \Pr(y_i = 1 \mid \beta) = \mu(x_i; \beta)$. The link function is the logit, $h(\pi) = \log\{\pi/(1-\pi)\}$, and

$$
\begin{aligned}
w(\eta) &= \exp(-|\eta|)\,\text{expit}(|\eta|)^2 \\
&\sim \exp(-|\eta|) \text{ as } |\eta| \to \infty\,.
\end{aligned}
\tag{2.6}
$$

Here $\text{expit}(\eta) = 1/\{1 + e^{-\eta}\}$. Lemma 4 is now 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, then $\phi(\xi; \mathcal{P}) > -\infty$.*

There is then no requirement for $\mathcal{P}$ to have bounded support and Bayesian D-optimality can be used to select a design with a normal or log-normal prior on the parameters. There is also no requirement for the parameters to be independent a priori. For example, the result applies to a normal-mixture hierarchical variable selection prior distribution (Chipman et al. (1997)).

Other prior distributions do not satisfy the conditions of Theorem 1; for example that proposed by Gelman, Jakulin, Pittau, and Su (2008), who recommend first rescaling the explanatory variables in observational studies to have mean zero and standard deviation $\frac{1}{2}$; for designed experiments, we suggest rescaling to the range $[-\frac{1}{2}, \frac{1}{2}]$.

It is possible to obtain a partial inverse result to Theorem 1.

Proposition 4. *If $j \in \{0, \ldots, p-1\}$, and if*

(i) *$\mathcal{P}$ is such that $\Pr(\beta_j > 1) > 0$;*

(ii) *$\mathcal{P}$ is such that, for all $\delta > 0$,*

$$\Pr(|\beta_k| < \delta \textit{ for all } k \neq j \,|\, \beta_j > 1) > 0\,;$$

(iii) *$\mathcal{P}$ is such that, for all $\delta > 0$,*

$$E_{\mathcal{P}}[\beta_j \,|\, \beta_j > 1,\, |\beta_k| < \delta,\; \textit{for all } k \neq j] = \infty\,;$$

(iv) *ξ is such that $\min_{i=1,\ldots,n} |f_j(x_i)| > 0$;*

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

A more intuitive understanding of the reason that these conditions lead to a singular prior distribution can be obtained by considering locally optimal design. There, we have that $|M(\xi; \beta)| \approx 0$ if for all design points the success probability $\Pr(y_i = 1 \mid \beta)$ is close to either 0 or 1. In that case, there is also a high probability of separation (Albert and Anderson (1984)) and thus the non-existence of maximum likelihood estimates. For Bayesian design, a heavy-tailed prior satisfying the conditions of Proposition 4 leads to similarly extreme values of the success probability, which is now a random variable owing to dependence on β.

Proposition 5. *Under the conditions of Proposition 4, there exists an event $\mathcal{E} \subseteq \Theta$, with $\Pr(\mathcal{E}) > 0$, conditional upon which either $\Pr(y_i = 1 \mid \beta)$ or $1 - \Pr(y_i = 1 \mid \beta)$ has prior geometric mean zero, according to whether $f_j(x_i) < 0$ or $f_j(x_i) > 0$, respectively.*

The proofs of Propositions 4 and 5 both rest on the identification of a region, $\mathcal{E}$, of parameter space where the linear predictor η_i can be approximated by the contribution, $\beta_j f_j(x_i)$, from the jth predictor.

The Gelman prior distribution, $\mathcal{P}_G$, places independent standard Cauchy distributions on $(1/10)\beta_0, (2/5)\beta_1, \ldots, (2/5)\beta_{p-1}$. Thus, the prior distributions for the regression coefficients are heavy-tailed, with undefined prior mean. The parameters a priori have $E|\beta_k| = \infty$, $k = 0, \ldots, p-1$. For a model with an intercept term, $f_0(x) = 1$, and Proposition 4 may be applied with $j = 0$; conditions (ii) and (iii) follow since β_0 is both heavy-tailed and independent of the other parameters.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

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

Often prior independence of parameters is not a reasonable assumption. For example, Chipman et al. (1997) define a hierarchical variable selection prior in which the probability of an interaction term being active is dependent on whether the parent terms are active, thereby satisfying the weak heredity principle. Proposition 4 can be used to show that, for logistic regression, a prior with this hierarchical structure is singular if the prior distribution of the intercept parameter is a mixture of two scaled zero-mode Cauchy distributions. In this case, the intercept is again both heavy-tailed and (typically) independent of the other parameters.

For logistic models with a single controllable variable, scalar $x \in \mathcal{X}$, $\mathcal{X} = \mathbb{R}$, Bayesian D-optimal design has also been studied for a different parameterization (for example, Chaloner and Larntz (1989)):

$$h(\pi_i) = \beta_1(x_i - \mu)\,, \tag{2.7}$$

which can be obtained from (2.4) via $\beta_0 = -\beta_1\mu$. When $\beta_1 = 0$, μ is not identifiable and $|M_\theta(\xi;\theta)| = 0$ for all $\xi \in \Xi$, with $\theta = (\mu, \beta_1)^{\mathrm{T}}$, $\Xi = \mathcal{X}^n$. The following result is straightforward to prove using Theorem 1.

Proposition 6. *For the* (μ, β_1)*-parameterization in* (2.7)*, if (i)* $E_{\mathcal{P}}(|\mu\beta_1|) <$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

∞, (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 $\mathcal{P}$. 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).

2.3.2. Probit regression

For probit regression, $y_i \,|\, \beta \sim \text{Bernoulli}(\pi_i)$, $\pi_i = \mu(x_i; \beta)$, with link $h(\pi) = \Phi^{-1}(\pi)$, where Φ is the standard normal c.d.f.. Here,

$$w(\eta) = \frac{\varphi(\eta)^2}{\Phi(\eta)(1-\Phi(\eta))}\,,$$

where $\varphi(\eta) = \frac{1}{\sqrt{2\pi}} e^{-\eta^2/2}$ is the standard normal p.d.f.. According to Abramowitz and Stegun (1964),

$$1 - \Phi(\eta) \sim \frac{1}{\eta\sqrt{2\pi}} e^{-\eta^2/2} \quad \text{as } \eta \to \infty\,.$$

Also, as $\eta \to \infty$, $\Phi(\eta) \to 1$, and so by symmetry of $w(\eta)$

$$w(\eta) \sim \frac{1}{\sqrt{2\pi}} |\eta| e^{-\eta^2/2} \quad \text{as } |\eta| \to \infty\,. \tag{2.8}$$

This asymptotic approximation can be used with Lemma 4 to obtain analogues of the results for logistic regression, with different conditions on the prior distribution.

Theorem 2. *If $E_{\mathcal{P}}|\beta_k\beta_l| < \infty$, for $k, l = 0, 1, \ldots, p-1$, then $\mathcal{P}$ is non-singular for the probit regression model.*

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

Proposition 7. *Given $j \in \{0, \ldots, p-1\}$, suppose that:*

(i) $\mathcal{P}$ *is such that* $\Pr(\beta_j > 1) > 0$ *and, for all* $\delta > 0$,

$$\Pr(|\beta_k| < \delta \textit{ for all } k \neq j \mid \beta_j > 1) > 0$$

(ii) $\mathcal{P}$ *is such that, for all* $\delta > 0$,

$$E_{\mathcal{P}}[\,|\beta_j|^2 \,\big|\, \beta_j > 1, |\beta_k| < \delta \textit{ for all } k \neq j] = \infty$$

(iii) ξ *is such that* $\min_i |f_j(x_i)| > 0$.

Then, for the probit link the design ξ is Bayesian singular with respect to $\mathcal{P}$.

Corollary 3. *For a probit model with an intercept term, the prior distribution $\mathcal{P}_G$ is singular.*

Again, a heavy-tailed prior on the intercept parameter results in the prior being singular for Bayesian D-optimality. The intuitive interpretation is similar to that for the logistic model. Here $\mathcal{P}_G$ would remain singular even if it were made somewhat less heavy-tailed, for example by replacing the Cauchy prior on β_0 with a $t(2)$ prior. In this case, condition (ii) above still holds because β_0 has infinite variance.

2.3.3. Poisson regression

Consider the model $y_i \mid \beta \sim \text{Poisson}(\lambda_i)$, with $\mu_i = \lambda_i$ and $h(\mu) = \log \mu$.

Optimal designs for this model were considered by Russell et al. (2009) and McGree and Eccleston (2012). Here, $w(\eta) = \exp(\eta)$ and we have the following results.

Theorem 3. *For the Poisson regression model with log link, if* $E_{\mathcal{P}}|\beta_k| < \infty$, $k = 0, \ldots, p-1$, *and* $|F^{\mathrm{T}}F| > 0$ *then* $\mathcal{P}$ *is non-singular.*

Proposition 8. *Given* $j \in \{0, \ldots, p-1\}$, *suppose that*

(i) $\mathcal{P}$ *is such that* β_j *is supported on* $(-\infty, 0)$,

(ii) $\mathcal{P}$ *is such that* $\Pr(\beta_j < -1) > 0$ *and, for all* $\delta > 0$,

$$\Pr(|\beta_k| < \delta \text{ for all } k \neq j \,|\, \beta_j < -1) > 0\,,$$

(iii) $\mathcal{P}$ *is such that for all* $\delta > 0$,

$$E_{\mathcal{P}}[\beta_j \,|\, \beta_j < -1,\, |\beta_k| < \delta \text{ for all } k \neq j] = -\infty\,,$$

(iv) $\mathcal{P}$ *is such that* $E_{\mathcal{P}}|\beta_k| < \infty$, $k \neq j$,

(v) ξ *is such that* $f_j(x_i) > 0$ *for* $i = 1, \ldots, n$.

Then ξ *is singular for the Poisson model with log link under Bayesian D-optimality.*

Corollary 4. *For a Poisson model with log link containing an intercept,* $f_0(x) = 1$, *if* β_0 *has a negative half-Cauchy prior independently of* β_k, $k = 1, \ldots, p-1$, *with* $E_{\mathcal{P}}|\beta_k| < \infty$, *then* $\mathcal{P}$ *is singular.*

Here, a heavy-tailed negative intercept parameter can result in a singular prior. Intuitively, it is clear that large negative values of β_0 lead to experiments where most of the responses are zero, leading to difficulties obtaining precise estimates of β_0 and the other parameters.

3. Numerical methods to overcome ill-conditioning

3.1. Objective function approximation

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

$$\phi(\xi;\mathcal{P}) \approx \phi(\xi;\mathcal{Q}) = \sum_{l=1}^{N_\mathcal{Q}} v_l \log\left|M(\xi;\beta^{(l)})\right|, \tag{3.1}$$

over a weighted sample,

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

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

A problem with approximation (3.1) is that for multi-parameter models numerical evaluation of $\phi(\xi;\mathcal{Q})$ can fail due to the presence of ill-conditioned matrices $M(\xi;\beta^{(l)})$, whose determinant will be estimated numerically as zero. This can occur even for non-singular $\mathcal{P}$; for singular $\mathcal{P}$ there is little point in evaluating $\phi(\xi;\mathcal{Q})$ since $\phi(\xi;\mathcal{P}) = -\infty$. When numerical evaluation of $\phi(\xi;\mathcal{Q})$ fails for all $\xi \in \Xi$, we say that $\mathcal{Q}$ is an *ill-conditioned quadrature scheme*. In principle, $\mathcal{Q}$ can be ill-conditioned for any prior distribution. However, for the models considered here, such as logistic regression, ill-conditioning of $\mathcal{Q}$ is more likely if the underlying prior distribution is heavy-tailed. In that case, there is high probability of large β, and so also of $M(\xi;\beta)$ being ill-conditioned. For any prior, even without heavy tails, other circumstances that may lead to ill-conditioning of $\mathcal{Q}$ include: use of a quadrature scheme, such as the Gotwalt method, which oversamples the tails of $\mathcal{P}$; use of a large number of quadrature points. In most integration problems, an increased number of quadrature points leads to improved approximation of the integral; paradoxically, in Bayesian D-optimal design this may cause numerical evaluation to fail due to ill-conditioning.

3.2. Objective function bounds for logistic regression

For some models, it is possible to obtain bounds that allow approxima-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

tion of $\phi(\xi; \mathcal{Q})$ when $\mathcal{Q}$ is ill-conditioned, as often occurs for heavy-tailed priors. These bounds may be applied to enable straightforward selection of Bayesian D-efficient designs for such priors (see Section 3.3). Here we focus on the case of logistic regression; a similar approach can be used for the compartmental model (using Lemma 1), and other GLMs. From Lemma 3 and (2.6), we see that $\phi(\xi; \beta) = \log |M(\xi; \beta)|$ lies in $[\phi_L(\xi; \beta), \phi_U(\xi; \beta)]$, where

$$\phi_L(\xi; \beta) = \log |F^{\mathrm{T}}F| + p \min_{i=1,\dots,n} \{-|\eta_i| + 2 \log \operatorname{expit} |\eta_i|\}$$

$$\phi_U(\xi; \beta) = \log |F^{\mathrm{T}}F| + p \max_{i=1,\dots,n} \{-|\eta_i| + 2 \log \operatorname{expit} |\eta_i|\} .$$

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

$$\phi_L(\xi; \mathcal{Q}) \le \phi(\xi; \mathcal{Q}) \le \phi_U(\xi; \mathcal{Q}) , \tag{3.2}$$

where

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

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 (3.2) a relatively narrow interval for $\phi(\xi;\mathcal{Q})$. Here (3.2) 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 an example to show how the bounds enable an extension of the set of prior distributions for which Bayesian D-efficient designs can be obtained in practice. 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 authors' models,

$$f(x) = (1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3)^{\mathrm{T}}$$

$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_{12}, \beta_{13}, \beta_{23})^{\mathrm{T}},$$

where $q=3$, $x=(x_1,x_2,x_3)^{\mathrm{T}}$. We adopt a different prior distribution: $\log\beta_0 \sim N(-1,2)$, $\beta_1\sim N(2,2)$, $\beta_2\sim N(1,2)$, $\beta_3\sim N(-1,2)$, and $\beta_{12},\beta_{13},\beta_{23}\sim N(0.5,2)$, independently. The log-normal prior for the intercept parameter is heavy-tailed. However, from Theorem 1, the joint 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 five radial points and four random rotations. Direct numerical evaluation of $\phi(\xi;\mathcal{Q})$ failed, since $\mathcal{S}$ contained 39 parameter vectors. However, from (3.2), $\phi(\xi;\mathcal{Q}) \in [-6.85, -6.78]$.

3.3. Use of bounds in design optimization and assessment

The bounds from (3.2) may also be used within an optimization algorithm to help find Bayesian D-efficient designs. The *Bayesian D-efficiency* of ξ is

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

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

When $\mathcal{Q}$ is well-conditioned, the Bayesian D-efficiency may be approximated by numerical search for a $\xi^*_{\mathcal{Q}} \in \arg\max_{\xi\in\Xi}\phi(\xi;\mathcal{Q})$ that maximizes the quadrature approximated objective function, and substitution of 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, for example if $\mathcal{P}$ is heavy-tailed, then this method fails since $\phi(\xi;\mathcal{Q})$ cannot be evaluated directly, and $\xi^*_{\mathcal{Q}}$ cannot be found using a numerical search. We may nonetheless use numerical methods to find designs $\xi^*_{\mathcal{Q},L}$ and $\xi^*_{\mathcal{Q},U}$ so that $\xi^*_{\mathcal{Q},L} \in \arg\max_{\xi\in\Xi} \phi_L(\xi;\mathcal{Q})$ and $\xi^*_{\mathcal{Q},U} \in \arg\max_{\xi\in\Xi} \phi_U(\xi;\mathcal{Q})$. Then a lower bound for the Bayesian efficiency of $\xi^*_{\mathcal{Q},L}$ can be approximated, via substitution of the designs found, into

$$\text{Bayes-eff}(\xi^*_{\mathcal{Q},L};\mathcal{Q}) \geq \exp\{[\phi_L(\xi^*_{\mathcal{Q},L};\mathcal{Q}) - \phi_U(\xi^*_{\mathcal{Q},U};\mathcal{Q})]/p\} \times 100\%\,. \quad (3.3)$$

To find exact designs that maximize 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 three 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, is similar to $\xi^*_{\mathcal{Q},U}$: to 2 d.p. the two are identical. For this $\mathcal{Q}$, the objective function $\phi(\xi;\mathcal{Q})$ cannot be computed exactly due to ill-conditioning. Thus, given an alternative design ξ', e.g. a 16-run combination of $\xi^*_{\mathcal{Q},L}$ and $\xi^*_{\mathcal{Q},U}$, it is not possible to evaluate whether ξ' has higher Bayesian D-efficiency than $\xi^*_{\mathcal{Q},L}$. However,

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

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, $\xi^*_{\mathcal{Q},L}$, that maximizes the lower bound $\phi_L(\xi;\mathcal{Q})$.

the lower bound on the Bayesian D-efficiency is Bayes-eff$(\xi^*_{\mathcal{Q},L};\mathcal{Q}) \gtrsim 99.4\%$, so any improvement to be gained by using a different design will be small.

The computation of the numerical value of the lower bound in (3.3) is approximate since we cannot be certain to have found the global optimum $\xi^*_{\mathcal{Q},U}$, although here 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

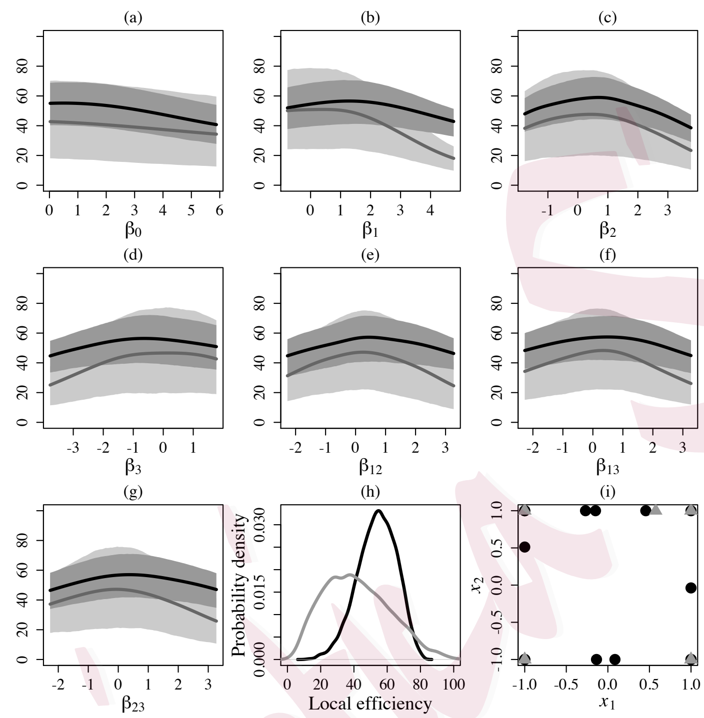

Figure 1: Robustness comparison of the Bayesian D-efficient design $\xi^*_{\mathcal{Q},L}$ (black lines/points) versus the EW-optimal design (grey lines/points). *Panels* (a)–(g): conditional distribution, given β_k, of the local efficiency, $\text{eff}(\xi;\beta)$, induced by the prior on β (solid line, conditional mean; shaded region, 10% and 90% quantiles). *Panel* (h): marginal distribution of the local efficiency. *Panel* (i): 2-dimensional projection of the design points.

the local D-efficiency,

$$\text{eff}(\xi;\beta) = \{|M(\xi;\beta)|/|M(\xi^*_\beta;\beta)|\}^{1/p}\,, \tag{3.4}$$

where $\xi^*_\beta \in \arg\max_{\xi\in\Xi} |M(\xi;\beta)|$ is a locally D-optimal design. 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 ξ^*_β, and substituting the design found into (3.4). For other β, $M(\xi;\beta)$ is ill-conditioned for all $\xi \in \Xi$. Then approximate bounds on the efficiency can be derived by numerical search for the designs $\xi^*_{U,\beta} \in \arg\max_{\xi\in\Xi} \phi_U(\xi;\beta)$ and $\xi^*_{L,\beta} \in \arg\max_{\xi\in\Xi} \phi_L(\xi;\beta)$, and from the fact that

$$\exp\frac{1}{p}[\phi_L(\xi;\beta) - \phi_U(\xi^*_{U,\beta};\beta)] \le \text{eff}(\xi;\beta) \le \exp\frac{1}{p}[\phi_U(\xi;\beta) - \phi_L(\xi^*_{L,\beta};\beta)]\,. \tag{3.5}$$

To visualize the dependence of the local efficiency on the individual parameters, for each regression coefficient β_j we plot the approximate mean and 10% and 90% quantiles of the conditional distribution of $\text{eff}(\xi;\beta)$ given β_j. Owing to the need to search for a locally D-optimal design, evaluation of $\text{eff}(\xi;\beta)$ is computationally intensive. Thus, before computing the conditional mean and quantiles it is advantageous to first build a statistical emulator of $\text{eff}(\xi;\beta)$ as a function of β, using Gaussian process interpolation. This is analogous to the approach followed in the computer experiments

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

literature when the main effects of a computationally expensive simulator are visualized (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, $\xi^*_{\mathcal{Q},L}$, that maximizes the lower bound for $\phi(\xi;\mathcal{Q})$. The support points of the quadrature scheme are used to train the emulator of $\mathrm{eff}(\xi^*_{\mathcal{Q},L};\beta)$. In the example, only three 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 (3.5) gave no additional information beyond $\mathrm{eff}(\xi^*_{\mathcal{Q},L};\beta) \in [0\%, 100\%]$. Thus we decided to omit them from the training set, as including the bounds $[0\%, 100\%]$ would not substantively reduce uncertainty about their efficiency. Figure 1 shows approximations to the conditional mean and conditional quantiles (given β_j) of the local efficiency, obtained using the emulator and Monte Carlo sampling. Also shown is a kernel density estimate for the marginal 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 local efficiency of $\xi^*_{\mathcal{Q},L}$ is in the range 55-60%. The lower and upper quartiles of the local efficiency distribution are approximately 46% and 62%. Although at first glance the typical local efficiencies may appear fairly low, it is important to remember that due to the large amount of prior uncertainty here, there is no design whose local efficiency is significantly higher than $\xi^*_{\mathcal{Q},L}$ uniformly across the entire parameter space. The design $\xi^*_{\mathcal{Q},L}$ achieves a near-optimal trade-off in performance, as quantified by the high estimated Bayesian D-efficiency obtained earlier, across the very different parameter scenarios that are possible under the prior for β. The design is thus relatively robust. There appear to be no significant areas of the parameter space where the design performance is poor and, for example, the prior probability that $\mathrm{eff}(\xi^*_{\mathcal{Q},L};\beta) < 0.2$ appears negligible.

For comparison, results are included for the EW-optimal design, ξ^*_{EW}, advocated by Yang et al. (2016), which maximizes

$$\psi_{\mathrm{EW}}(\xi) = \log|EM_\beta(\xi;\beta)| = \log\left|\sum_{i=1}^{n} E[w(\eta_i)]\, f(x_i)f^{\mathrm{T}}(x_i)\right| .$$

In factorial experiments for logistic regression, Yang et al. (2016) found EW-optimal designs to be of comparable statistical efficiency to Bayesian D-optimal designs, while requiring less computational effort to obtain. Here, as shown in Figure 1(a)–(h), the EW-optimal design is much less robust than the Bayesian D-efficient design, at least in this case; its local efficiency $\mathrm{eff}(\xi^*_{\mathrm{EW}};\theta)$ has generally lower mean, both conditionally on β_j

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

and marginally, and the local efficiency also exhibits higher variability. The smaller difference in robustness between Bayesian D-optimal and EW-optimal designs observed by Yang et al. (2016) may be due to their restriction to a factorial design space; the greater performance of the Bayesian D-efficient design appears to be due to the inclusion of a greater number of factor settings in the interior of $(-1, 1)$ (see Figure 1(i)).

In addition to the lesser statistical performance of the EW-optimal design in this case, here the computational convenience of EW-optimal designs over Bayesian D-optimal designs is much reduced. For factorial problems, a reduction in computational cost is achieved by precomputing $E[w(\eta)]$ for every point in the finite design space, enabling faster evaluation of $\psi_{\mathrm{EW}}(\xi)$. Here, precomputation is not possible since we have continuous factors. One computational benefit of the EW criterion is that it successfully avoids problems with ill-conditioning, but this offers only a minor advantage over Bayesian D-optimality, for which ill-conditioning problems can now be overcome using the bounds developed in Section 3.2.

4. Discussion

The central tenet of this paper is that it is not permissible to use a singular prior distribution in conjunction with Bayesian D-optimality as a design selection criterion. One rough intuitive interpretation of this is

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

that parameter uncertainty under such a prior is so great that any design will have low (local) efficiency across a significant portion of the parameter space. In this case, there are two possibilities: consider a different prior distribution; or adopt a different design selection criterion. These alternatives are considered in Sections 4.1 and 4.2, respectively.

We identified a prominent class of default prior distributions for logistic regression that should not be used for Bayesian D-optimal design. Future work could seek to develop results on singular prior distributions for population pharmacokinetic models, for which optimal sampling times are more commonly sought (e.g. Mentré et al. (1997)). Such models extend (2.3) by allowing subject-specific kinetic parameters.

4.1. Alternative prior distributions

Often there are multiple plausible candidates for a suitable prior distribution. In the subjectivist framework, informative priors are elicited from expert knowledge by obtaining summaries to which a probability distribution may be fitted (e.g. Garthwaite et al. (2005), Oakley and O'Hagan (2007)). If using uninformative or weakly informative priors there are still often multiple candidate priors. If design selection fails because $\mathcal{P}$ is singular but there exists an alternative candidate prior $\mathcal{P}'$ that is non-singular, then Bayesian D-optimality could be used with $\mathcal{P}'$ instead. In all, one must

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0293

avoid selecting prior distributions for analytical convenience if they do not accurately represent the available expert belief or knowledge. For further discussion of this, and the role of the prior in design of experiments, see the supplementary material and Woods et al. (2016).

4.2. Alternative selection criteria

If all candidate prior distributions that agree with the elicited prior knowledge or beliefs are singular, then it is necessary to use an alternative design selection criterion that suffers from fewer problems with singularities. One such criterion is EW D-optimality (Yang et al. (2016)). The numerical results in Section 3.3 suggest that if the problem is with an ill-conditioned $\mathcal{Q}$ rather than a singular $\mathcal{P}$, then the EW D-optimal design may be less robust than a Bayesian D-efficient design found using the numerical methods developed here. Another alternative selects ξ to maximize the *mean local efficiency*,

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

This is fairly insensitive to the presence of θ with $|M(\xi;\theta)| \approx 0$, and is a special case of the objective function discussed by Dette and Wong (1996) (Φ_1 in their notation). Unlike Bayesian D-optimality, these alternative criteria do not have the interpretation of approximate equivalence to the maximization of Shannon information gain.

Another approach to design selection under parameter uncertainty is to consider maximin designs. In the case of greatest interest in this paper, Θ is such that $\inf_{\theta\in\Theta}|M(\xi;\theta)| = 0$ for all $\xi \in \Xi$, so design selection fails using the unstandardized maximin D-criterion (Imhof (2001)). Often design selection also fails when using the standardized maximin D-criterion. It is clear that the Bayesian approach, under suitable prior distributions, benefits from greater robustness to the presence of singular θ than the use of maximin criteria. For related results see Braess and Dette (2007).

Supplementary Materials

The online supplementary material for this paper contains additional discussion and proofs of the analytical results described in the text.

Acknowledgements

I am grateful to two anyonymous referees whose insightful and helpful comments prompted several improvements to the paper, and to Professors David Woods and Susan Lewis for several invaluable discussions and comments. This work was supported by 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.

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover, New York. 10th printing.

Albert, A. and Anderson, J. (1984). 'On the existence of maximum likelihood estimates in logistic regression models'. *Biometrika* **71**, 1–10.

Atkinson, A. C. (2003). 'Horwitz's rule, transforming both sides and the design of experiments for mechanistic models'. *J. Roy. Stat. Soc. C* **52**, 261–278.

Atkinson, A. C., Chaloner, K., Herzberg, A. M., and Juritz, J. (1993). 'Optimum experimental designs for properties of a compartmental model'. *Biometrics* **49**, 325–337.

Atkinson, A. C. and Woods, D. C. (2015). 'Designs for generalized linear models'. In Dean, A., Morris, M., Stufken, J., and Bingham, D. (Eds.), 'Handbook of Design and Analysis of Experiments', Chapter 13, pages 471–548. Chapman and Hall/CRC.

Braess, D. and Dette, H. (2007). 'On the number of support points of maximin and Bayesian optimal designs'. *Ann. Statist.* **35**, 772–792.

Capinski, M. and Kopp, P. E. (2004). *Measure, Integral and Probability*. Springer, New York. 2nd edn.

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.

Chernoff, H. (1953). 'Locally optimal designs for estimating parameters'. *Ann. Math. Statist.* **24**, 586–602.

Chipman, H., Hamada, M., and Wu, C. (1997). 'A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing'. *Technometrics* **39**, 372–381.

Dette, H. and Neugebauer, H.-M. (1997). 'Bayesian D-optimal designs for exponential regression models'. *J. Statist. Plann. Infer.* **60**, 331–349.

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

Feng, C., Wang, H., Zhang, Y., Han, Y., Liang, Y., and Tu, X. M. (2017). 'Generalized definition of the geometric mean of a nonnegative random variable'. *Comm. Statist. Th. Meth.* **46**, 3614–3620.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). 'Statistical methods for eliciting probability distributions'. *J. Amer. Statist. Assoc.* **100**, 680–701.

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.

Imhof, L. A. (2001). 'Maximin designs for exponential growth models and heteroscedastic polynomial models'. *Ann. Statist.* **29**, 561–576.

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*. Chapman and Hall, Boca Raton, FL. 2nd edn.

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

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

Oakley, J. E. and O'Hagan, A. (2007). 'Uncertainty in prior elicitations: a nonparametric approach'. *Biometrika* **94**, 427–441.

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

Pukelsheim, F. (1993). *Optimal design of experiments*. SIAM, Philadelphia.

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

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

Waite, T. W. and Woods, D. C. (2015). 'Designs for generalized linear models with random block effects via information matrix approximations'. *Biometrika* **102**, 677–693.

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.

Woods, D. C., Overstall, A. M., Adamou, M., and Waite, T. W. (2016). 'Bayesian design of experiments for generalised linear models and dimensional analysis with industrial and scientific application (with discussion)'. *Quality Engineering* **29**, 91–118.

Yang, J. and Mandal, A. (2015). 'D-optimal factorial designs under generalized linear models'. *Comm. Statist. Simul. Comput.* **44**, 2264–2277.

Yang, J., Mandal, A., and Majumdar, D. (2016). 'Optimal designs for 2^k factorial experiments with binary response'. *Statist. Sinica* **26**, 385–411.

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

Yang, M. and Stufken, J. (2009). 'Support points of locally optimal designs for nonlinear models with two parameters'. *Ann. Statist.* **37**, 518–541.

Yang, M. and Stufken, J. (2012). 'Identifying locally optimal designs for nonlinear models: a simple extension with profound consequences'. *Ann. Statist.* **40**, 1665–1681.

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