Integrability and Bayesian $D$-optimality

Tim Waite

t.w.waite@southampton.ac.uk

Statistical Sciences Research Institute
University of Southampton

Supported by the UK Engineering and Physical Sciences Research Council

PODE2013 - 15 June 2013
Lilly UK, Surrey
Outline

- Bayesian design and approximations
- Parameter singularities and non-integrability
- Suggestions for addressing non-integrability
- Illustration: how badly things can go wrong
In recent years, there have been many developments in optimal experimental designs for more sophisticated models:

- generalized linear mixed models
- nonlinear mixed effects models

For these models, $D$-optimal designs etc. depend on the values of the parameters.

**Approaches**

- locally optimal ‘best guess’
- Bayesian design
- maximin designs
- response-adaptive/sequential design
Bayesian design

(Chaloner & Verdinelli, 1995)

Notation

- $\mathbf{y} \in \mathcal{Y}$ responses
- $\mathbf{\theta} \in \Theta$ parameter vector
- prior knowledge summarized by $f(\mathbf{\theta})$
- $\xi$ finitely-supported approximate design

Idea: choose $\xi$ to maximize the expected ‘distance’ / information gain

$$f(\mathbf{\theta}) \text{ prior } \rightarrow f(\mathbf{\theta}|\xi, \mathbf{y}) \text{ posterior}$$

‘Distance’ measured via Kullback-Leibler divergence / Shannon information gain

$$\psi_{KL}(\xi) = \int_{\mathcal{Y}} \int_{\Theta} \log \frac{f(\mathbf{\theta}|\xi, \mathbf{y})}{f(\mathbf{\theta})} f(\mathbf{y}, \mathbf{\theta}|\xi) d\mathbf{\theta} d\mathbf{y}$$
In practice, optimization of expected information gain is usually too hard.

Instead we typically optimize a surrogate objective function

\[
\phi(\xi) = \mathbb{E}_\theta \log |nM(\xi, \theta)| \tag{1}
\]

\[
\phi_2(\xi) = \mathbb{E}_\theta \log |nM(\xi, \theta) + R| \tag{2}
\]

- \(M(\xi, \theta)\) is the Fisher information matrix, \(\mathbb{E}_y \left[ -\frac{\partial^2 \log f(y|\xi,\theta)}{\partial \theta \partial \theta} \right] \)
- \(R = \frac{\partial^2 \log f(\theta)}{\partial \theta \partial \theta}, \text{ or } R = \text{var}(\theta)^{-1} \)

Objective function (1) is the most common

Also sometimes used when a Bayesian analysis will not be conducted (pseudo-Bayesian design)

Focus of the talk - sometimes the approximation \(\phi\) can fail badly
Singularities

Intuitive definition
A **parameter singularity** is a combination of parameter values where all designs are 'uninformative'.

Formal definition
\( \theta_0 \) is a parameter singularity if, for any \( \xi \), \( |M(\xi, \theta_0)| = 0 \)

Example
Exponential regression model

\[
\begin{align*}
y_i & \sim N[\eta(x_i), \sigma^2] \\
\eta(x) & = e^{-x/\theta}
\end{align*}
\]

parameterized by lifetime \( \theta > 0 \)

Parameter singularities \( \{0, \infty\} \)
Singularities

**Intuitive definition**

A parameter singularity is a combination of parameter values where all designs are 'uninformative'.

**Formal definition**

\( \theta_0 \) is a parameter singularity if, for any fixed \( \xi \), \( |M(\xi, \theta)| \to 0 \) as \( \theta \to \theta_0 \).

**Example**

Exponential regression model

\[
y_i \sim N[\eta(x_i), \sigma^2]
\]

\[
\eta(x) = e^{-x/\theta}
\]

parameterized by lifetime \( \theta > 0 \)

Parameter singularities \( \{0, \infty\} \)
Why are there parameter singularities at 0 and $\infty$?
Example II

Logistic regression

Binary response (0/1) - event occurs or does not occur. Controllable variable, $x$, is usually a (log)-dose

$$y_i \sim \text{Bernoulli}(p_i)$$

$$p_i = \frac{1}{1 + \exp\{-\beta(x_i - \mu)\}}$$

- $\mu$ is the dose at which there is a 50% chance of the event occurring
- Chaloner & Larntz (1989) studied Bayesian design for this model

Parameter singularities at $\beta = 0, \beta = \infty$

Why? When $\beta = 0$, $\mu$ is not identifiable
Parameter singularity at $\beta = \infty$

\[ p(x) \]

$\beta = 6$
$\beta = 1$
The punchline

Reconsider

\[ \phi(\xi) = \int_{\Theta} \log |M(\xi, \theta)| f(\theta) \, d\theta \]

- near parameter singularities, \(|M(\xi, \theta)| \to 0\)
- the integrand above \(\to -\infty\) in parts of the domain of integration

So the integral

- may not exist (Riemann)
- may equal \(-\infty\) (Lebesgue)

(in the same sense that \(\int_{[0,1]} \frac{-1}{x^2} \, dx\) either doesn’t exist or is \(-\infty\))
Chaloner & Verdinelli (1995) discussed integrability, but only highlighted the issue where the prior has unbounded support.

Tsutakawa (1972) gave a problem where the integral is $-\infty$.

The issue is not restricted to unbounded supports.
What can be done?

1. Use a different approximation
   - not $\phi$, instead e.g. $\phi_2$, or a more sophisticated computational approximation

2. Use a different design selection criterion
   - if Bayesian analysis will not be used, the principled justification of Shannon information gain breaks down

3. Use a different prior

4. Density designs?
Alternative criteria

Efficiency distribution

Consider the $D$-efficiency function

\[
\text{eff}(\xi|\theta) = \left\{ \frac{|M(\xi, \theta)|}{\sup_{\xi'} |M(\xi', \theta)|} \right\}^{1/p}
\]

- prior on $\theta$ induces a distribution on $\text{eff}(\xi|\theta)$
- Woods et al. (2006) used efficiency function & distribution to assess designs

From pseudo-Bayesian viewpoint, optimization of e.g. $\phi$ is a device to obtain satisfactory efficiency distribution
If analysis non-Bayesian, and $\phi$ is degenerate, makes sense to use a criterion which is well-behaved

**Mean local efficiency**

One approach is to maximize

$$\Psi(\xi) = E_{\theta} \text{eff}(\xi|\theta)$$

- has an interpretation as minimizing an expected *cost regret*
- (amount of overspend due to inefficiency, when compared with other equally informative designs)
Density designs

One suggestion is, instead of finitely-supported designs,

\[
\left\{ \begin{array}{c}
    x_1 \\
    \vdots \\
    x_k \\
    w_1 \\
    \vdots \\
    w_k
\end{array} \right\}
\]

define a design using a probability density function, \( g(x) \), on \( \mathcal{X} \)

In some sense such designs ‘get everywhere’ in \( \mathcal{X} \), and are infinitely-supported

Have been considered, e.g. by Wiens (1992) in context of model robustness

Information matrix formed as

\[
M(\xi, \theta) = \int_{\mathcal{X}} M(x, \theta) g(x) dx
\]
Detailed example

Exponential regression

\[ y_i \sim N[\eta(x_i), \sigma^2] \]
\[ \eta(x) = e^{-x/\theta} \]

Assume a priori that

\[ \theta \sim U(0, a) \]

Parameter singularities \( \{0, \infty\} \), but for \( \theta > 0 \) only singular design is \( x = 0 \)

- for fixed \( \theta > 0 \), \( |M(\xi, \theta)| = 0 \) only when \( \xi \) puts unit mass on \( x = 0 \)

Lemma all single-point designs have \( \phi(\xi) = -\infty \)

Theorem all finitely-supported designs have \( \phi(\xi) = -\infty \)
Conclusion: here $\phi$ is useless in helping us make a choice between designs

- despite the fact the prior support is bounded
- numerical methods - spurious comparisons
To prove the Lemma, consider

$$\log |M(x, \theta)| = -\frac{2x}{\theta} - 4 \log \theta + 2 \log x$$

To prove the Theorem, make use of the following inequality. For $x > 0$, $y \geq 0$

$$\log(x + y) \leq \log(x) + y/x$$

Can be used to show that

$$\log |M(\xi, \theta)| \leq \log w_1 + \log M(x_1, \theta) + T(\theta)$$

WLOG $x_1 \leq x_2 \leq \ldots x_n$, in which case it is true that $0 \leq T(\theta) \leq 1$

$$E \log |M(\xi, \theta)| \leq \log w_1 + E \log M(x_1, \theta) + E T(\theta) = -\infty$$
Locally optimal and maximum mean efficiency designs can be computed analytically.

**Proposition** the locally $D$-optimal design at $\theta$ is the single-point design $x = \theta$.

**Proposition** the design which maximizes the mean local efficiency under $\theta \sim U(0, a)$ is the single-point design $x = a/2$

- The mean local efficiency of this design is 67%, regardless of $a$. 

Tim Waite (U. Southampton)
Properties of the $\Psi$-optimal design

(a) Efficiency vs. $\theta$

(b) Density vs. Efficiency

Integrability and Bayesian $D$-optimality
Density designs

Consider the design $\xi_U$ defined by a uniform probability density on $(0, a)$

$$g(x) = a^{-1}1(0 < x < a)$$

Recall that the information matrix is

$$M(\xi_U, \theta) = \frac{1}{a} \int_0^a M(x, \theta) \, dx$$

It can be shown that

$$\phi(\xi_U) = \mathbb{E} \log |M(\xi_U, \theta)| > -\infty$$

so the uniform design is not degenerate with respect to $\phi$

Can also compute the $D$-efficiency

$$\text{eff}(\xi_U|\theta) = \left\{ \frac{|M(\xi_U, \theta)|}{\sup_{\xi'} |M(\xi', \theta)|} \right\}^{1/p}$$

and the efficiency distribution
Properties of the uniform density design

(c)

(d)
Density designs cannot be used directly in practice. How about finite (random) samples from the distribution?

Let $X_n = (x_1, \ldots, x_n)$ be such a sample.

**Proposition**

As $n \to \infty$, $\text{eff}(X_n|\theta) \to \text{eff}(\xi_U|\theta)$ almost surely

Moreover we can produce ‘95% performance limits’
Sampling properties of uniform design, $n = 100$
For any $\theta$, we have a positive probability of obtaining a reasonably efficient design

This must be traded off with the probability of obtaining a design which is inefficient for most values of $\theta$

Moreover, the sampled design will have $\phi(X_n) = -\infty$
Conclusions

- when producing Bayesian designs, be cautious about integrability
- if parameter singularities can’t be avoided, consider alternative approximations/criteria

Future work

- development of further explicitly pseudo-Bayesian criteria
- other situations where random designs may be helpful
References