Introduction to Optimal Design of Experiments

Weng Kee Wong

Department of Biostatistics **Fielding School of Public Health**

UCLA

An Overview of Optimal Design of Experiments
Biostatistics 279
Spring 2021

3.1 Approximate designs (Kiefer, 1958-1982)

Suppose X, $f(x)$, N and an optimality criterion ϕ are given.

Formulation for Optimal Approximate Design Problem:
How many points are needed to optimize the criterion? Find k
Where are the optimal design (or support) points? Find
$x_1, x_2, \cdots, x_k \in X$
What is the optimal proportion of the total observations to take at each of these points? Find $w_1, w_2, , w_k$ such that

$$0 < w_i < 1, \ i = 1, 2, \cdots, k$$

$$w_1 + w_{2+} \cdots + w_k = 1.$$

The implemented design takes $n_i = [Nw_i]$ observations at design points $x_i, i = 1, 2, \cdots, k$ such that $n_1 + n_2 + \cdots + n_k = N$.

3.5 Locally D-optimal Designs for the Logistic Model on $X = [-1, 1]$ (Ford's PhD thesis, 1972)

$$log \frac{\pi(x)}{1-\pi(x)} = \theta_1 + \theta_2 x, \quad \theta^T = (\theta_1, \theta_2), \;\; \theta_1 > 0 \;\; \& \;\; \theta_2 > 0.$$

- Let a solve $exp(z) = (z+1)/(z-1)$, i.e. $a = 1.54$ and let u^* solve

$$exp(\theta_1 + \theta_2 u) = \frac{2 + (u+1)\theta_2}{-2 + (u+1)\theta_2}.$$

-

condition	locally D-optimal design
$\{\theta : \theta_2 - \theta_1 \geq a\}$	$\{\frac{a-\theta_1}{\theta_2}, \frac{-a-\theta_1}{\theta_2}; \frac{1}{2}, \frac{1}{2}\}$
$\{\theta : \theta_2 - \theta_1 < a, exp(\theta_1 + \theta_2) \leq \frac{\theta_2+1}{\theta_2-1}\}$	$\{-1, u^*; \frac{1}{2}, \frac{1}{2}\}$
$\{\theta : exp(\theta_1 + \theta_2) > \frac{\theta_2+1}{\theta_2-1}\}$	$\{-1, 1; \frac{1}{2}, \frac{1}{2}\}$

- Corrected results in Sebastiani and Settimi (JSPI, 1997)

3.6 Information Matrices for linear models

$$y(x) = f^T(x)\theta + e(x)/\sqrt{\lambda(x)}, \qquad x \in X.$$

- $f(x)$ = a given $d \times 1$ vector of regression functions
- $Ee(x) = 0$; $var(e(x))= \sigma^2$; $\lambda(x)$ = known positive function

If errors are normally and independently distributed, Fisher information matrix for a k-point design ξ is proportional to

$$M(\xi) = \sum_{i=1}^{k} \lambda(x_i) w_i f(x_i) f^T(x_i), \qquad k \geq d$$

and $cov(\widehat{\theta}) = M(\xi)^{-1}$(apart from a multiplicative constant).

For a nonlinear model, we have $E(y) = f(x, \theta)$; replace above $f(x)$ by gradient of $f(x, \theta)$ (with respect to θ).

3.7 Optimality Design Criteria

Formulate objective as a function of the information matrix:

- D-optimality: $\Phi(\xi) = \mathit{ln}\,|M(\xi)^{-1}|$

3.7 Optimality Design Criteria

Formulate objective as a function of the information matrix:

- D-optimality: $\Phi(\xi) = ln\,|M(\xi)^{-1}|$
- L-optimality: $\Phi(\xi) = tr\ L\ M(\xi)^{-1}$ for a user-selected matrix L

If L = identity matrix, we have $\Phi(\xi) = tr\ M(\xi)^{-1}$ (A-optimality)

If L = $f(z)f(z)^T, \Phi(\xi) = tr\ f(z)f(z)^T M(\xi)^{-1} = f(z)^T M(\xi)^{-1} f(z)$,
=var(z, ξ)
(the variance of y at z).

If L = $\frac{\partial g(\theta)}{\partial \theta} \frac{\partial g^T(\theta)}{\partial \theta^T}$ for a given function $g(\theta)$, $\Phi(\xi)$ is proportional to the asymptotic variance of the estimated $g(\theta)$.
(use a first order Taylor's expansion)

3.8 Convex Analysis

Each of the above Φ is convex; so find ξ^* that minimizes $\Phi(\xi)$ over ALL designs ξ on X.

3.8 Convex Analysis

Each of the above Φ is convex; so find ξ^* that minimizes $\Phi(\xi)$ over ALL designs ξ on X.

- The Frechet derivative of Φ at $M(\xi^*)$ in the direction of $M(\xi)$ is

$$F_\Phi(M(\xi^*), M(\xi)) = \lim_{\alpha \to 0} \frac{\Phi((1-\alpha)M(\xi^*) + \alpha M(\xi)) - \Phi(M(\xi^*)))}{\alpha}.$$

- If Φ is convex on the set of information matrices and differentiable at $M(\xi^*)$, then ξ^* is Φ-optimal if and only if

$$F_\Phi(M(\xi^*), f(x)f^T(x))) \geq 0$$

for all $x \in$ X with equality at the design points of ξ^*.

Each of the above Φ is convex; so find ξ^* that minimizes $\Phi(\xi)$ over ALL designs ξ on X.

- The Frechet derivative of Φ at $M(\xi^*)$ in the direction of $M(\xi)$ is

$$F_\Phi(M(\xi^*), M(\xi)) = \lim_{\alpha \to 0} \frac{\Phi((1-\alpha)M(\xi^*) + \alpha M(\xi)) - \Phi(M(\xi^*)))}{\alpha}.$$

- If Φ is convex on the set of information matrices and differentiable at $M(\xi^*)$, then ξ^* is Φ-optimal if and only if

$$F_\Phi(M(\xi^*), f(x)f^T(x))) \geq 0$$

  for all $x \in$ X with equality at the design points of ξ^*.

- If Φ is non-differentiable, work with sub-gradient; see Wong (Biometrika, 1992), Wong and Cook (JRSSB, 1993) and Wong (JSPI, 1994), Berger, King & Wong (Psychometrika, 2001).

3.9 Checking Conditions or Equivalence Theorems

Aim: To verify optimality of a design among ALL designs on X.

Let $(X, f(x), \lambda(x))$ be given. The design ξ^* is D-optimal if and only if

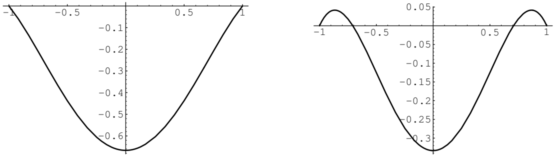
$$-F_\Phi(M(\xi^*), f(x)f^T(x))) = \lambda(x)f^T(x)M(\xi^*)^{-1}f(x) - d \leq 0 \quad \textit{for } \forall\, x \in X,$$

with equality at the support points of ξ^*. Here d is the dimension of f(x). Draw a picture to verify optimality!

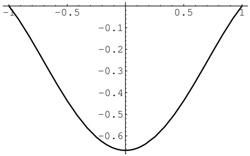
An Example:

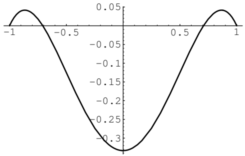
Suppose $f^T(x) = (1, x), X = [-1, 1]$ and $\lambda(x) = c - x^2$, $c > 1$. Is the design equally supported at ± 1 D-optimal for all $c > 1$?

3.10: Plots of the directional derivative for the equally weighted design at ± 1 when $\lambda(x) = 4 - x^2$ (left) and when $\lambda(x) = 2.5 - x^2$ on $X = [-1, 1]$.



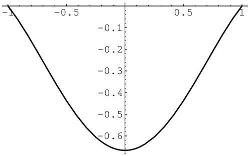
3.10: Plots of the directional derivative for the equally weighted design at ± 1 when $\lambda(x) = 4 - x^2$ (left) and when $\lambda(x) = 2.5 - x^2$ on $X = [-1, 1]$.

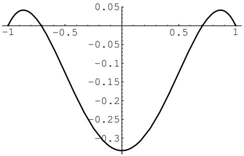




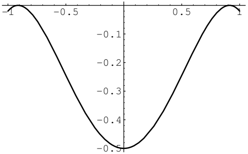
- The D-optimal design is equally supported at ± 1 if $c \geq 3$; otherwise it is equally supported at $\pm\sqrt{c/3}$ ($c = 2.5$ below).

3.10: Plots of the directional derivative for the equally weighted design at ± 1 when $\lambda(x) = 4 - x^2$ (left) and when $\lambda(x) = 2.5 - x^2$ on $X = [-1, 1]$.





- The D-optimal design is equally supported at ± 1 if $c \geq 3$; otherwise it is equally supported at $\pm\sqrt{c/3}$ ($c = 2.5$ below).



3.11 Basic Design Strategy for Multiple Objectives

- Constrained Optimal Designs
  i.e. design that satisfies a set of user-specified efficiency requirements; eg. minimize $\phi_2(\xi)$ subject to $\phi_1(\xi) \leq c$.

- Compound Optimal Designs
  i.e. design that minimizes a fixed convex combination of convex functionals: $\phi(\xi|\lambda) = \lambda\phi_1(\xi) + (1-\lambda)\phi_2(\xi)$.

- Compound Optimal Designs are equivalent to Constrained Optimal Designs: Plot efficiencies of each compound optimal versus λ, $\lambda \in [0, 1]$.

- In practice, prioritize the importance of the objectives and apply theory for single-objective study, Cook & Wong , JASA (1994), Wong, Statistica Neerlandica (1999), Huang & Wong, Drug Information Journal (2004)

3.12 Efficiency Plots for Dual-Objective Optimal Designs

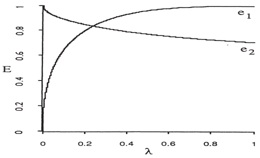


Figure 1. Efficiencies E Versus λ for Example 1.

Further, using (9), we obtain the values given in the first two lines of table 1 in Studden (1982).

To connect explicitly the constrained design problems and the compound design problems, the relationship between a constraint expressed as $E_1(\xi) \geq e_1$ (or $E_2(\xi) \geq e_2$) and the corresponding value of λ must be established. Using results from Fedorov (1980, thm. 1), it can be shown that ξ_λ maximizes $\phi(\xi \mid \lambda)$ if and only if

$$2(1-\lambda)d_1(x, \xi_\lambda) + \lambda\{b^T M_2^{-1}(\xi_\lambda) f_2(x)\}^2 - 4(1-\lambda) - \lambda b^T M_2^{-1}(\xi_\lambda) b \leq 0 \quad (11)$$

for all x in $[-1, 1]$. Further (11) becomes an equality at the support points for ξ_λ. In that case, substituting (9) into (11) yields

$$\lambda = \frac{e_1^2}{4 + 4(1-e_1)^{1/2} - 4e_1 + e_1^2}. \quad (12)$$

Figure 1, constructed using (10) and (12), shows the relationship between optimal designs with efficiency constraints and compound optimal designs. For example, a design that maximizes ϕ_2 subject to the constraint $E_1(\xi) \geq .6$ can be found by maximizing $\phi(\xi \mid \lambda)$ with $\lambda \approx .1$. Figure 1 contains useful information on the interpretation of λ as well. In particular, it might be felt that setting $\lambda = .5$ would yield a design in which equal interest is placed on the two criteria. But from Figure 1, the compound design problem with $\lambda = .5$ is equivalent to the constrained problem in which we maximize ϕ_2 subject to the constraint that $E_1(\xi) \geq .96$. The resulting constrained design has $E_2(\xi_{\lambda=.5}) \approx .78$. In terms of the efficiencies, placing equal interest on the two criteria would seem to require $\lambda = .25$, because at that point $E_1(\xi_\lambda) \approx E_2(\xi_\lambda) \approx .84$. Finally, reconstructing the plot in Figure 1 so that the horizontal axis is $1 - \lambda$ rather than λ provides the corresponding plot for maximizing ϕ_1 subject to a constraint on ϕ_2.

Example 2. For the simple linear regression model $f_1^T(x) = (1, x)$ on $\chi = [-1, 1]$, consider balancing A optimality with precise estimation of the response at the point $z = .75$:

$$\phi_1(\xi) = -d_1(z, \xi)/d_1(z, \xi^1) = -[E_1(\xi)]^{-1}$$

and

$$\phi_2(\xi) = -\mathrm{tr}\, M_1^{-1}(\xi)/\mathrm{tr}\, M_1^{-1}(\xi^2) = -[E_2(\xi)]^{-1}.$$

The design ξ^1 is optimal for ϕ_1 and has the minimum variance possible for a fitted value at the point z. This minimum variance is the same as that obtained under the design that places mass 1 at z. The design ξ^2 is the A-optimal design. Both ξ^1 and ξ^2 are supported at ± 1, with the masses at 1 being $\frac{7}{8}$ and $\frac{1}{2}$. Thus $d_1(z, \xi^1) = 1$ and $\mathrm{tr}[M_1^{-1}(\xi^2)] = 2$.

Define $w(x) = (9x^2 - 25x + 16)^{1/2}$ and $g(x) = (3x + 4 - w(x))/8$ for $0 \leq x \leq 1$. Then

$$\xi_\lambda(1) = g(E_1(\xi_\lambda)), \quad (13)$$

which is the analog of (9) for this example. Next, again from Fedorov (1980, thm. 1) ξ_λ satisfies

$$\lambda(f_1^T(x) M_1^{-1}(\xi_\lambda) f_1(z))^2 + (1-\lambda) f_1(x) M_1^{-2}(\xi_\lambda) f_1(x) = \lambda d_1(z, \xi_\lambda) + (1-\lambda)\,\mathrm{tr}\, M_1^{-1}(\xi_\lambda)$$

at the support points $x = \pm 1$. Substituting (13) into this equation yields

$$\lambda = \frac{8(w(e_1) - 3e_1)}{34g(e_1) - 17 - 48g^2(e_1)},$$

which is the analog of (12) for $e_1 \geq .64$. From this we constructed Figure 2. The relationships in Figure 2 are, of course, qualitatively similar to those in Figure 1. But note that the value of λ at which the efficiencies are equal is much larger than that for Figure 1. Generally, the interpretation of λ depends heavily on the functionals involved. Useful inter-

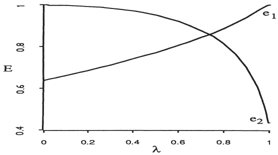


Figure 2. Efficiencies E Versus λ for Example 2.

3.13 Design Efficiency - what it means exactly?

Efficiency of a design ξ is defined relative to the optimal design ξ^*.

For estimating θ_1 in the simple linear model: $eff(\xi) = \frac{var_{\xi^*}(\hat{\theta}_1)}{var_{\xi}(\hat{\theta}_1)}$

Interpretation: If the efficiency of ξ is 0.5, ξ needs to be replicated twice to do as well as ξ^*.

3.13 Design Efficiency - what it means exactly?

Efficiency of a design ξ is defined relative to the optimal design ξ^*.

For estimating θ_1 in the simple linear model: $eff(\xi) = \frac{var_{\xi^*}(\hat{\theta}_1)}{var_{\xi}(\hat{\theta}_1)}$

- For estimating both parameters, define efficiency by

Interpretation: If the efficiency of ξ is 0.5, ξ needs to be replicated twice to do as well as ξ^*.

3.13 Design Efficiency - what it means exactly?

Efficiency of a design ξ is defined relative to the optimal design ξ^*.

For estimating θ_1 in the simple linear model: $eff(\xi) = \frac{var_{\xi^*}(\hat{\theta}_1)}{var_{\xi}(\hat{\theta}_1)}$

- For estimating both parameters, define efficiency by
- A-efficiency = $\frac{var_{\xi^*}(\hat{\theta}_0) + var_{\xi^*}(\hat{\theta}_1)}{var_{\xi}(\hat{\theta}_0) + var_{\xi}(\hat{\theta}_1)}$

Interpretation: If the efficiency of ξ is 0.5, ξ needs to be replicated twice to do as well as ξ^*.

3.13 Design Efficiency - what it means exactly?

Efficiency of a design ξ is defined relative to the optimal design ξ^*.

For estimating θ_1 in the simple linear model: $eff(\xi) = \frac{var_{\xi^*}(\hat{\theta}_1)}{var_{\xi}(\hat{\theta}_1)}$

- For estimating both parameters, define efficiency by
- A-efficiency = $\frac{var_{\xi^*}(\hat{\theta}_0) + var_{\xi^*}(\hat{\theta}_1)}{var_{\xi}(\hat{\theta}_0) + var_{\xi}(\hat{\theta}_1)}$
- D-efficiency = some ratio of the areas of the two confidence regions from the two designs

Interpretation: If the efficiency of ξ is 0.5, ξ needs to be replicated twice to do as well as ξ^*.

4 Methods for Searching Optimal Designs

- Brute force and guess work
- A purely theoretical approach
- Algorithms

4.1 By brute force: find a minimax D-optimal design

Consider the Logistic Model on a given interval X. Let $\theta^T = (\theta_1, \theta_2) \in \Theta$, Θ known and let Ξ be the set of all designs on X.

Design Criterion: Find ξ^* = arg $\min_{\xi \in \Xi} \max_{\theta \in \Theta}$ *log* $|M(\xi, \theta)|^{-1}$.

4.1 By brute force: find a minimax D-optimal design

Consider the Logistic Model on a given interval X. Let $\theta^T = (\theta_1, \theta_2) \in \Theta$, Θ known and let Ξ be the set of all designs on X.

Design Criterion: Find ξ^* = arg $\min_{\xi \in \Xi} \max_{\theta \in \Theta}$ *log* $|M(\xi, \theta)|^{-1}$.

- No known algorithms that will generate a minimax optimal design.

4.1 By brute force: find a minimax D-optimal design

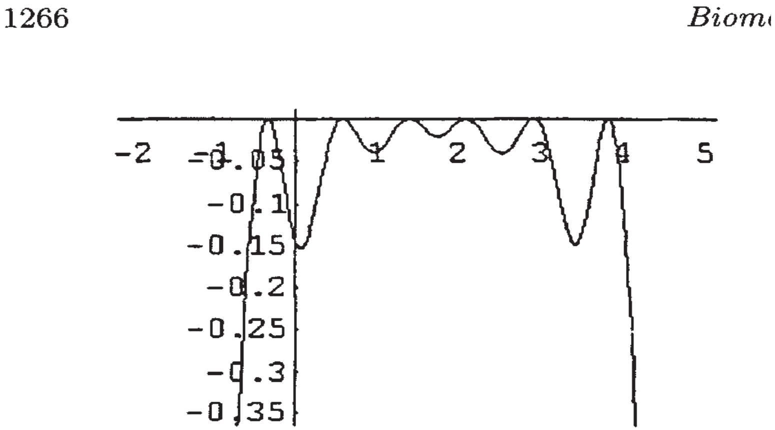
Consider the Logistic Model on a given interval X. Let $\theta^T = (\theta_1, \theta_2) \in \Theta$, Θ known and let Ξ be the set of all designs on X.

Design Criterion: Find ξ^* = arg $\min_{\xi \in \Xi} \max_{\theta \in \Theta}$ *log* $|M(\xi, \theta)|^{-1}$.

- No known algorithms that will generate a minimax optimal design.
- King & Wong (Biometrics, 2002) found minimax D-optimal designs when $\Theta = [0, 3.5] \times [1, 3.5]$ and X is unrestricted:

x_i	-0.35	0.62	1.39	2.11	2.88	3.85
w_i	0.18	0.21	0.11	0.11	0.21	0.18

4.1 Equivalence Plot confirms the Minimax D-Optimal Design for the Logistic Model



4.3 Need for Algorithms To Search for Optimal Designs

- Derivation of optimal designs for nonlinear models is tedious, difficult and method for one model does not usually generalize to another

Knowledge in matrix algebra and matrix derivatives are helpful. Exemplary tools include results from the simultaneous diagonalization theorem and typical results from matrix derivatives:

Suppose the elements of each of the matrices A and B are functions of a scalar t. Assume A is non-singular.

(i) $\frac{dAB}{dt} = \frac{dA}{dt} + A\frac{dB}{dt}$

(ii) $\frac{dA^{-1}}{dt} = -A^{-1}\frac{dA}{dt}A^{-1}$

(iii) $\frac{d\ \ln|A|}{dt} = -trA^{-1}\frac{dA}{dt}$

For proofs of these results and many other interesting related results, see the monograph by Rogers, G. S, 1980: Matrix Derivatives. Lecture Notes in Statistics, Vol. 2, Dekker.

In Praise of Simplicity not Mathematistry! Ten Simple Powerful Ideas for the Statistical Scientist

Roderick J. LITTLE

Ronald Fisher was by all accounts a first-rate mathematician, but he saw himself as a scientist, not a mathematician, and he railed against what George Box called (in his Fisher lecture) "mathematistry." Mathematics is the indispensable foundation of statistics, but for me the real excitement and value of our subject lies in its application to other disciplines. We should not view statistics as another branch of mathematics and favor mathematical complexity over clarifying, formulating, and solving real-world problems. Valuing simplicity, I describe 10 simple and powerful ideas that have influenced my thinking about statistics, in my areas of research interest: missing data, causal inference, survey sampling, and statistical modeling in general. The overarching theme is that statistics is a missing data problem and the goal is to predict unknowns with appropriate measures of uncertainty.

KEY WORDS: Calibrated Bayes; Causal inference; Measurement error; Missing data; Penalized spline of propensity.

1. INTRODUCTION: THE UNEASY RELATIONSHIP BETWEEN STATISTICS AND MATHEMATICS

American Statistical Association President, Sastry Pantula, recently proposed renaming the Division of Mathematical Sciences at the U.S. National Science Foundation as the Division of Mathematical and Statistical Sciences. Opponents, who viewed statistics as a branch of mathematics, questioned why statistics should be singled out for special treatment.

Data can be assembled in support of the argument that statistics *is* different—for example, the substantial number of academic departments of statistics and biostatistics, the rise of the statistics advanced placement examination, and the substantial number of undergraduate statistics majors. But the most important factor for me is that statistics is not just a branch of mathematics. It is an inductive method, defined by its applications to the sciences and other areas of human endeavor, where we try to glean information from data.

The relationship between mathematics and statistics is somewhat uneasy. Since the mathematics of statistics is often viewed as basically rather pedestrian, statistics is rather low on the totem pole of mathematical subdisciplines. Statistics needs its mathematical parent, since it is the indispensable underpinning of the subject. On the other hand, unruly statistics has ambitions to reach beyond the mathematics fold; it comes alive in applica- and medicine, and with increasing influence recently on the hard sciences such as astronomy, geology and physics.

The scientific theme of modern statistics fits the character of its most influential developer, the great geneticist, R. A. Fisher, who seemed to revolutionize the field of statistics in his spare time! Fisher's momentous move to Rothampsted Experimental Station rather than academia underlined his dedication to science. Though an excellent mathematician, Fisher viewed himself primarily as a scientist, and disparaged rivals like Neyman and Pearson as mere "mathematicians."

George Box's engaging Fisher lecture focused on the links between statistics and science (Box 1976). He wrote:

> My theme then will be first to show the part that [Fisher] being a good scientist played in his astonishing ingenuity, originality, inventiveness, and productivity as a statistician, and second to consider what message that has for us now.

Box attributed Fisher's hostility to mathematicians to distaste for what he called "mathematistry," which he defined as

> [. . .] the development of theory for theory's sake, which, since it seldom touches down with practice, has a tendency to redefine the problem rather than solve it. Typically, there has once been a statistical problem with scientific relevance but this has long since been lost sight of. (Box 1976)