Constructing efficient exact designs of experiments using a branch-and-bound method

Mgr. Eva Bednáriková

Supervisor: doc. Mgr. Radoslav Harman, PhD.

Faculty of Mathematics, Physics and Informatics,
Comenius University in Bratislava, Slovakia

January 20, 2014

Table of contents

Linear regression model

$$y(x) = \mathbf{f}^\top(x)\beta + \varepsilon(x)$$

where

- $x \in \mathfrak{X}$, and $\mathfrak{X} = \{x_1, x_2, \ldots, x_n\}$ is a finite design space,
- $y(x)$ is an observation in design point x,
- $\mathbf{f}(x) \in \mathbb{R}^m$ is a known regression function,
- $\beta \in \mathbb{R}^m$ is a vector of unknown parameters,
- $\mathrm{E}(\varepsilon(x)) = 0$, $\mathrm{Var}(\varepsilon(x)) = \sigma^2 < \infty$ and trials are performed independently.

Approximate and exact design of experiment

Approximate design ξ
is a probability measure on $\mathfrak{X}$

$$\xi = \left\{ \begin{array}{cccc} x_1 & x_2 & \cdots & x_n \\ \xi_1 & \xi_2 & \cdots & \xi_n \end{array} \right\}$$

- ξ_i is a weight of design point $x_i \in \mathfrak{X}$
- $\sum_i \xi_i = 1$

Exact design ξ_N of size N
is a probability measure on $\mathfrak{X}$, such that $N\xi_N(x) \in \mathbb{Z}$ for all $x \in \mathfrak{X}$

$$\xi_N = \left\{ \begin{array}{cccc} x_1 & x_2 & \cdots & x_n \\ \frac{N(x_1)}{N} & \frac{N(x_2)}{N} & \cdots & \frac{N(x_n)}{N} \end{array} \right\}$$

- N - the total number of trials
- $N(x_i)$ - the number of trials in the design point $x_i \in \mathfrak{X}$
- $\sum_i N(x_i) = N$

D-optimality

Information matrix of a design ξ

$$\mathbf{M}(\xi) = \sum_{x \in \mathfrak{X}} \xi(x)\mathbf{f}(x)\mathbf{f}^\top(x).$$

D-optimality criterion

$$\Phi : \mathbf{M}(\xi) \mapsto \Phi(\mathbf{M}(\xi)) = \det{}^{1/m}(\mathbf{M}(\xi)).$$

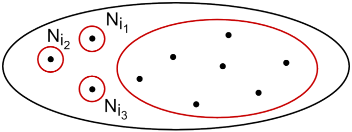
- D-optimal design ξ^* maximizes the D-optimality criterion.
- Statistical interpretation: D-optimal design minimizes the volume of the confidence ellipsoid for β.
- Our main goal is to find D-optimal exact designs, using results from the theory of D-optimal approximate design.

Partially fixed design

- Let $\{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, n\}$ be a set of indices.
- An approximate design ξ is called **partially fixed**, if

$$\xi(x_{i_l}) = \frac{N_{i_l}}{N} \quad \text{for} \ \ l = 1, 2, \ldots, k,$$

  where
  - N_{i_l} is a given number of trials in the design point x_{i_l}
  - N is the total number of trials; $\sum_{l=1}^{k} N_{i_l} \le N$.
- D-optimal partially fixed design maximizes the determinant of the information matrix among all partially fixed designs.



Relative D-efficiency

- The D-efficiency of design ξ relative to design η is given by

$$\mathrm{eff}_D(\xi|\eta) = \frac{\det^{1/m}(\mathbf{M}(\xi))}{\det^{1/m}(\mathbf{M}(\eta))}.$$

- The D-efficiency represents statistical "quality" of design ξ compared to design η.
- We usually compare designs to a D-optimal design ξ^*; in this case, $\mathrm{eff}_D(\xi|\xi^*) \leq 1$.

Re-normalization multiplicative heuristic

To compute a D-optimal partially fixed design, it is possible to use the **re-normalization multiplicative heuristic**, which consists of three parts:

1. choice of the initial design ξ_0,
2. transformation of design ξ_j to design ξ_{j+1},
3. the stopping rule.

In the following, we will use notation $\mathfrak{X}_0 := \mathfrak{X} \setminus \{x_{i_1}, \ldots, x_{i_k}\}$ and $N_0 := N - \sum_{l=1}^{k} N_{i_l}$.

Re-normalization multiplicative heuristic

- **Initial design** ξ_0 is
  - fixed to the values $N_{i_1}, \ldots, N_{i_k}$ for points $x_{i_1}, \ldots, x_{i_k}$,
  - uniformly distributed over all other points $x \in \mathfrak{X}_0$.
- At the step j of the algorithm, the **next design** ξ_{j+1} is
  - fixed to the values $N_{i_1}, \ldots, N_{i_k}$ for points $x_{i_1}, \ldots, x_{i_k}$,
  - for all $x \in \mathfrak{X}_0$ given by

$$\xi_{j+1}(x) = \xi_j(x) d(x, \xi_j) \frac{N_0/N}{\sum_{y \in \mathfrak{X}_0} \xi_j(y) d(y, \xi_j)},$$

    where $d(x, \xi) = \mathbf{f}^\top(x) \mathbf{M}(\xi)^{-1} \mathbf{f}(x)$ is a variance function.
- The **stopping rule** is based on the D-efficiency of the current design relative to the D-optimal design, as specified in the following theorem.

Illustration of the re-normalization multiplicative heuristic

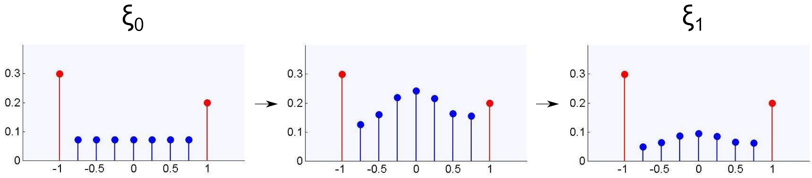


Figure: Illustration of one step of the re-normalization multiplicative heuristic. Points fixed to given values are marked in red, while points that remain “free” are marked in blue. An initial design ξ_0 is multiplied at first by the corresponding variance function, and then by a normalization constant (so that the new design ξ_1 is a probability measure on $\mathfrak{X}$).

Theoretical results of partially fixed designs

Theorem

Let ξ be a partially fixed design, let ξ^* be a D-optimal partially fixed design, and let $\epsilon = N^{-1}\left(\sum_{l=1}^{k} N_{i_l} d(x_{i_l}, \xi) + N_0 \max_{x \in \mathfrak{X}_0} d(x, \xi)\right) - m$. Then

$$\mathit{eff}_D(\xi|\xi^*) \geq \frac{m}{m+\epsilon}.$$

Moreover, let $h_m(\epsilon) = m\left(1 + \epsilon/2 - \sqrt{\epsilon(4+\epsilon-4/m)}/2\right)$. Then $\xi^*(x) = 0$ for any $x \in \mathfrak{X}_\circ$ satisfying

$$d(x, \xi) < \max_{y \in \mathfrak{X}_0} d(y, \xi) - \frac{N}{N_0}\left(m + \epsilon - h_m(\epsilon)\right).$$

This theorem follows from Harman [2014], and its second part can be used to accelerate the multiplicative algorithm mentioned before.

Branch-and-bound (B&B)

- a general enumerative method used to solve discrete optimization problems,
- for design problems, some versions of this method were used in, e.g., Welch [1982] and Ucinski and Patan [2007],
- each node of the branch-and-bound tree represents a class of partially fixed designs, associated with restrictions given by a couple $(\mathbf{I}, \mathbf{N})$, where $\mathbf{I} = (i_1, \ldots, i_k)$ and $\mathbf{N} = (N_{i_1}, \ldots, N_{i_k})$ for some $k \in \mathbb{Z}_0^+$,
- the **root node** of the tree corresponds to $\mathbf{I} = \emptyset, \mathbf{N} = \emptyset$ (no design point is fixed)

Construction step of the B&B tree

At each **construction step**, one node $(\mathbf{I} = (i_1, \dots, i_k), \mathbf{N} = (N_{i_1}, \dots, N_{i_k}))$ is divided into $N_0 + 1$ descendants with one more point $x_{i_{k+1}}$ fixed to the values $0, 1, \dots, N_0$:

$$(\mathbf{I}, \mathbf{N}) \begin{cases} ((\mathbf{I}, i_{k+1}) & , (\mathbf{N}, 0)) \\ ((\mathbf{I}, i_{k+1}) & , (\mathbf{N}, 1)) \\ & \vdots \\ ((\mathbf{I}, i_{k+1}) & , (\mathbf{N}, N_0)) \end{cases}$$

Illustration of the B&B tree

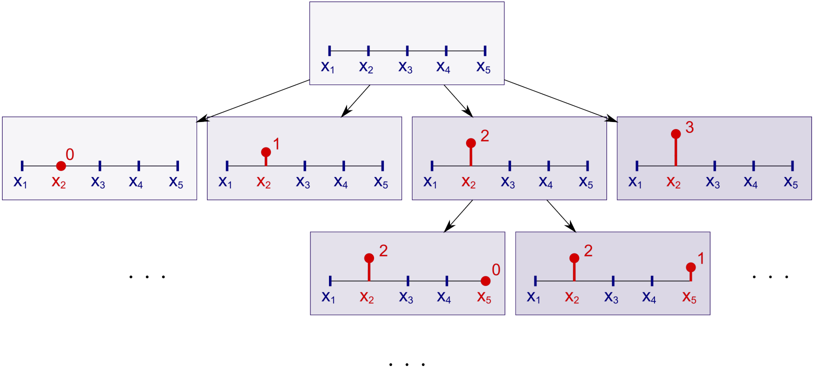


Figure: Illustration of a part of the branch-and-bound tree for $N = 3$ and $n = 5$. A blue tick represents a "free" design point, while a red dot represents a design point fixed to the particular value.

Termination of the nodes of the B&B tree

- The **global lower bound** *LB* is the maximum value of the D-optimality criterion among all information matrices of exact designs found so far.
- The **upper bound** $UB(\mathbf{I}, \mathbf{N})$ for each node $(\mathbf{I}, \mathbf{N})$ is a value of the D-optimality criterion of the corresponding D-optimal partially fixed design, that can be computed using the multiplicative algorithm. The node is discarded, if

  $$\alpha\, UB(\mathbf{I}, \mathbf{N}) < LB,$$

  where $\alpha > 0$ is a constant chosen so that the final design resulting from the B&B algorithm has the *required D-efficiency* relative to the D-optimal exact design.

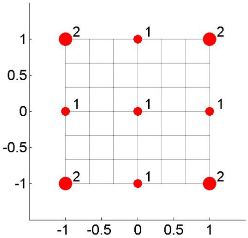
Example 1: Quadratic regression

- Quadratic regression in two variables:

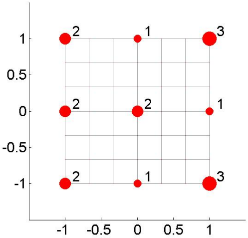
  $$y(x_1, x_2) = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 + \beta_6 x_2^2 + \varepsilon(x_1, x_2),$$

  where
- $(x_1, x_2)^\top \in \mathfrak{X} = \tilde{\mathfrak{X}} \times \tilde{\mathfrak{X}}$,
- $\tilde{\mathfrak{X}} = \left\{-1, -\frac{2}{3}, -\frac{1}{3}, 0, \frac{1}{3}, \frac{2}{3}, 1\right\}$.
- That is, the number of design points is $n = 49$.

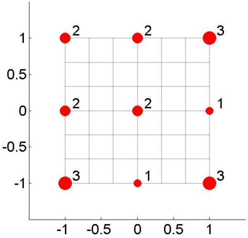
Example 1: Results



(a) $N = 13$



(b) $N = 17$



(c) $N = 19$

Figure: Exact designs for the quadratic regression model with D-efficiency at least 0.95 found by the branch-and-bound algorithm.

Example 2: Trigonometric regression

- Trigonometric regression in one variable:

$$\begin{aligned} y(x) = &\beta_1 + \beta_2 \sin x + \beta_3 \cos x + \beta_4 \sin 2x + \beta_5 \cos 2x \\ &+ \beta_6 \sin 3x + \beta_7 \cos 3x + \varepsilon(x), \end{aligned}$$

where

- $x \in \mathfrak{X} = \{0, d, 2d, 3d, \ldots, 29d\}$, where $d = \frac{2\pi}{30}$.
- That is, the number of design points is $n = 30$.

Example 2: Results

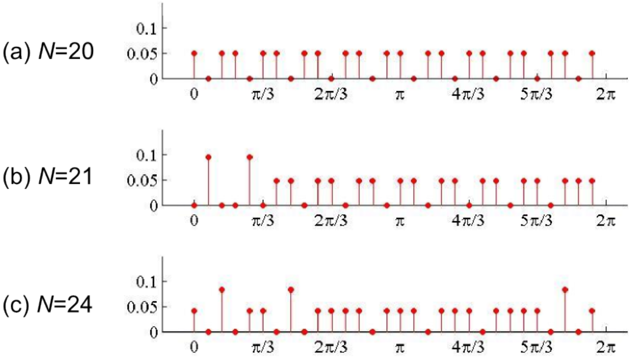


Figure: Exact designs for the trigonometric regression model with D-efficiency at least 0.95 found by the branch-and-bound algorithm.

References

R. Harman. Multiplicative methods for computing d-optimal stratified designs of experiments. *Journal of Statistical Planning and Inference*, 146:82–94, 2014.

D. Ucinski and M. Patan. D-optimal design of a monitoring network for parameter estimation of distributed systems. *Journal of Global Optimization*, 39(2):291–322, 2007.

W. J. Welch. Branch-and-bound search for experimental designs based on d optimality and other criteria. *Technometrics*, 24(1):41–48, 1982.

Thank you for attention.