D-optimal designs for multinomial experiments

Gwenda Thompson and Ken Russell*

Australian Bureau of Statistics, Canberra, AUSTRALIA

and

Charles Sturt University, Wagga Wagga, AUSTRALIA

kerussell@csu.edu.au

Where's Wagga Wagga ("Wogga Wogga")?

Outline

- Modelling a multinomial experiment
- Approximate (continuous) designs
- D-optimality
- Generalized Equivalence Theorem
- Examples
- Comments

Multinomial experiment

The response of an individual can fall in one of k classes, which need not form a hierarchy (contrast Zocchi & Atkinson, 1999).

$$\begin{aligned} \pi_{ij} &= \text{P(response of } i\text{th experimental unit falls in } j\text{th class)} \\ & \quad i = 1, \dots, n; \; j = 1, \dots, k. \end{aligned}$$

For each individual, observe values of m covariates $x_1, \dots, x_m$. Write $\mathbf{x} = (x_1, \dots, x_m)^\top$.

Want to model $\pi_1, \dots, \pi_k$ using $\mathbf{x}$.

Modelling the category probabilities

As $\pi_1 + \cdots + \pi_k = 1$, choose one probability (say π_1) as a 'baseline'. Model $\ln(\pi_j/\pi_1)$ by a linear combination of functions of the covariates. Write

$$\ln(\pi_j/\pi_1) = \eta_j(\mathbf{x}) = f_j^\top(\mathbf{x})\,\boldsymbol{\theta}, \quad (j = 2, \ldots, k)$$

with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ to be estimated.

Implies

$$\begin{aligned} \pi_1 &= \frac{1}{1 + \exp(\eta_2) + \cdots + \exp(\eta_k)}, \\ \pi_j &= \frac{\exp(\eta_j)}{1 + \exp(\eta_2) + \cdots + \exp(\eta_k)} \quad (j = 2, \ldots, k). \end{aligned}$$

Special case: $k = 2$ categories

$$\pi_2 = \frac{\exp(\eta_2)}{1 + \exp(\eta_2)} \quad \Leftrightarrow \quad \ln\left(\frac{\pi_2}{1 - \pi_2}\right) = \eta_2.$$

This is logistic regression, with category $2 =$ 'success'.

Example

For $k = 3$ and $m = 2$ covariates x_1, x_2, consider model

$$\begin{aligned} \ln(\pi_2/\pi_1) &= \theta_1 + \theta_2 x_1 + \theta_3 x_2 \\ \ln(\pi_3/\pi_1) &= \theta_4 + \theta_5 x_1 \end{aligned}$$

$$\text{i.e.,} \quad \boldsymbol{\eta} = \begin{pmatrix} \eta_2 \\ \eta_3 \end{pmatrix} = \begin{bmatrix} 1 & x_1 & x_2 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_5 \end{bmatrix} = \mathbf{F}\boldsymbol{\theta},$$

where

$$\mathbf{F} = \begin{pmatrix} f_2^\top(\mathbf{x}) \\ \vdots \\ f_k^\top(\mathbf{x}) \end{pmatrix}.$$

The design

To estimate $\boldsymbol{\theta}$, take observations at s support points $\mathbf{x}_1, \ldots, \mathbf{x}_s$ in the design space, $\mathcal{X}$.

Write

$$\xi = \left\{ \begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_s \\ w_1 & w_2 & \ldots & w_s \end{array} \right\}$$

for the approximate ('continuous') design to be used, where w_i represents the proportion of observations to be taken at $\mathbf{x}_i$ $(w_1 + \cdots + w_s = 1)$.

D-optimality

D-optimality: minimize volume of confidence ellipsoid for $\boldsymbol{\theta}$.

D-optimal design, $\boldsymbol{\xi}^*$, maximizes determinant of information matrix, $M(\xi, \boldsymbol{\theta})$, where

$$M(\xi, \boldsymbol{\theta}) = \sum_{i=1}^{s} \mathbf{F}_i^\top \mathbf{V} \mathbf{F}_i$$

and

$$\mathbf{V} = \begin{pmatrix} \pi_2(1-\pi_2) & -\pi_2\pi_3 & \dots & -\pi_2\pi_k \\ -\pi_2\pi_3 & \pi_3(1-\pi_3) & \dots & -\pi_3\pi_k \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_2\pi_k & -\pi_3\pi_k & \dots & \pi_k(1-\pi_k) \end{pmatrix}.$$

Locally optimal designs

In Generalized Linear Models, the information matrix $M(\xi, \boldsymbol{\theta})$ is a function of the parameters in $\boldsymbol{\theta}$ that we want to estimate.

Generally use an estimate, $\boldsymbol{\theta}_0$, of $\boldsymbol{\theta}$ in the expression for $M(\xi, \boldsymbol{\theta})$.

Call the design, ξ^*, that maximizes $|M(\xi, \boldsymbol{\theta}_0)|$ over some class of designs Ξ a locally D-optimal design. That is,

$$\xi^* = \arg\max_{\xi \in \Xi} |M(\xi, \boldsymbol{\theta}_0)|.$$

Generalized Equivalence Theorem

Is $|M(\xi, \boldsymbol{\theta})|$ at a local or global maximum?

Check using the Generalized Equivalence Theorem:

The design ξ^* is D-optimal if and only if

$$\begin{aligned} d(\xi^*, \mathbf{x}) &\leq p \quad \text{for all values of } \mathbf{x} \in \mathcal{X} \\ &= p \quad \text{for each support point } \mathbf{x} \text{ of } \xi^*, \end{aligned}$$

where $d(\xi^*, \mathbf{x})$ is the standardized variance, given by

$$d(\xi, \mathbf{x}) = \text{tr}\left(\mathbf{V}\mathbf{F}^\top M(\xi, \boldsymbol{\theta}_0)^{-1} \mathbf{F} \right)$$

for the multinomial experiment.

Example 1

For $k = 3$, $m = 1$, suppose we have

$$\eta_2 = 0 + x \quad \text{and} \quad \eta_3 = 0 - x;$$

$$\text{i.e., } \boldsymbol{\theta} = (0, 1, 0, -1)^\top.$$

Want to find locally D-optimal design over $\mathcal{X} = [-1, 1]$.

Example 1

For $k = 3$, $m = 1$, suppose we have

$$\eta_2 = 0 + x \quad \text{and} \quad \eta_3 = 0 - x;$$

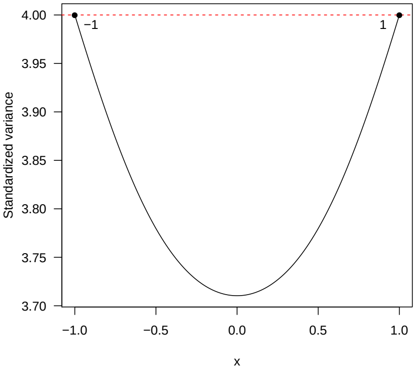
$$\text{i.e., } \boldsymbol{\theta} = (0, 1, 0, -1)^\top.$$

Want to find locally D-optimal design over $\mathcal{X} = [-1, 1]$.

Could it be

$$\xi_1 = \left\{ \begin{array}{cc} -1 & 1 \\ 0.5 & 0.5 \end{array} \right\}?$$

Standardized variance for ξ_1



So $d(\xi_1, \mathbf{x}) \leq p$ for all values of $\mathbf{x} \in \mathcal{X}$ and equals p at each support point of ξ_1. Hence ξ_1 is locally D-optimal.

The maximum values of $d(\xi_1, \mathbf{x})$ occurred at the extremities of $\mathcal{X}$. This suggests that, outside $\mathcal{X}$, $d(\xi_1, \mathbf{x})$ might keep increasing. Perhaps ξ_1 wouldn't be locally D-optimal for wider design regions.

Consider $\mathcal{X}_2 = [-2, 2]$.

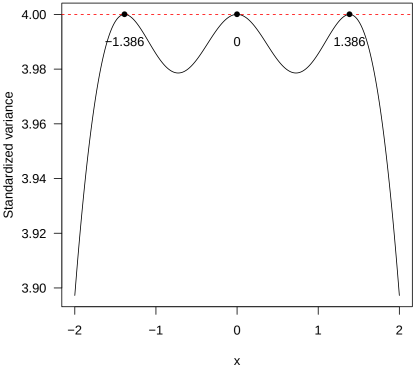
The maximum values of $d(\xi_1, \mathbf{x})$ occurred at the extremities of $\mathcal{X}$. This suggests that, outside $\mathcal{X}$, $d(\xi_1, \mathbf{x})$ might keep increasing. Perhaps ξ_1 wouldn't be locally D-optimal for wider design regions.

Consider $\mathcal{X}_2 = [-2, 2]$.

The locally D-optimal design for $\mathcal{X}_2$ is

$$\xi_2 = \left\{ \begin{array}{ccc} -1.386 & 0 & 1.386 \\ 0.3756 & 0.2488 & 0.3756 \end{array} \right\}.$$

Standardized variance for ξ_2

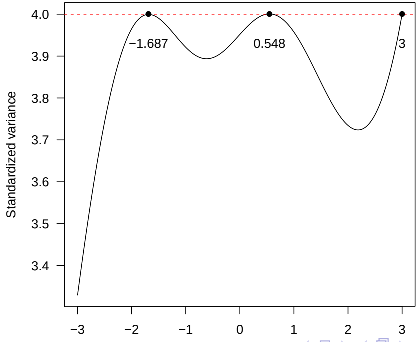


Support points of ξ_2 do not occur at the extremities of $\mathcal{X} = [-2, 2]$. So ξ_2 is also locally D-optimal for $\mathcal{X} = [-3, 3]$.

Example 2

$k = 3$, $m = 1$, $\eta_2 = 0 + 1x$, $\eta_3 = 0 + 1x$; i.e., $\boldsymbol{\theta} = (0, 1, 0, 1)^\top$.
The locally D-optimal design on $\mathcal{X} = [-3, 3]$ is

$$\xi_2 = \left\{ \begin{array}{ccc} -1.687 & 0.548 & 3 \\ 0.4180 & 0.2168 & 0.3652 \end{array} \right\}.$$



Broaden $\mathcal{X}$ to $[-5, 5]$

The locally D-optimal design on $\mathcal{X} = [-5, 5]$ is

$$\xi_3 = \left\{ \begin{array}{ccc} -1.867 & 0.758 & 5 \\ 0.3562 & 0.3799 & 0.2639 \end{array} \right\}.$$

Compare

$$\xi_2 = \left\{ \begin{array}{ccc} -1.687 & 0.548 & 3 \\ 0.4180 & 0.2168 & 0.3652 \end{array} \right\}$$

on $-[3, 3]$.

Example 3

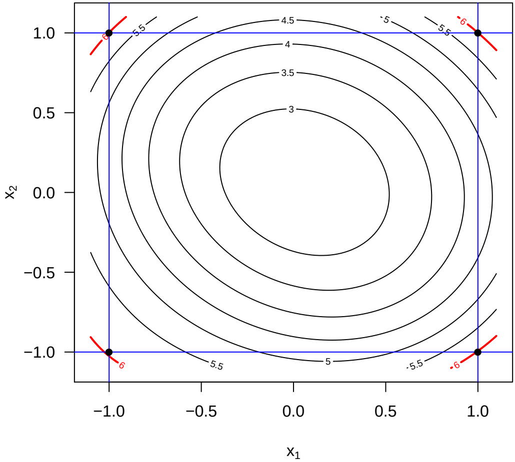
$k = 3$ categories, $m = 2$ variables, $\eta_2 = 0 + 1x_1 + 1x_2$, $\eta_3 = 0 + 1x_1 + 1x_2$; i.e. $\boldsymbol{\theta} = (0, 1, 1, 0, 1, 1)^\top$.

Take $\mathcal{X} = [-1, 1]^2$.

The locally D-optimal design is

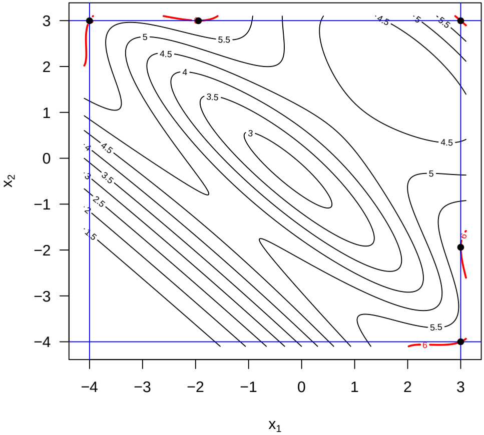
$$\xi_2 = \left\{ \begin{array}{cccc} (-1,-1) & (-1,1) & (1,-1) & (1,1) \\ 0.1944 & 0.2934 & 0.2917 & 0.2205 \end{array} \right\}.$$

Contour plot of Standardized Variance



Extend the design space to $[-4, 3]^2$

Note: $s = 5$ support points



Comments

- For experiments with a single response category (e.g. Poisson or logistic regression), the optimal design has s support points, where $p \leq s \leq p(p+1)/2$. Conjecture that, for the multinomial, have

$$\lceil p/(k-1) \rceil \leq s \leq \lceil p(p+1)/\{2(k-1)\} \rceil.$$

- Often obtained faster searches for optimal designs by rescaling problem; e.g., instead of $\eta = 1 + 2x$ for $x \in [-1, 1]$, write $\eta = 1 + 0.2x$ for $x \in [-10, 10]$. Increases magnitude of elements of information matrix.
- Using optimization routines in R, often found Nelder-Mead better than gradient procedures.

Postscript

Anthony Atkinson (yesterday): *Why does everyone estimate parameters . . . using the information matrix?*

Postscript

Anthony Atkinson (yesterday): *Why does everyone estimate parameters . . . using the information matrix?*

Have obtained designs which optimize the estimation of the probabilities $\pi_2, \ldots, \pi_k$ over all $\mathbf{x} \in \mathcal{X}$, using an Integrated Mean Square Error criterion.

Postscript

Anthony Atkinson (yesterday): *Why does everyone estimate parameters . . . using the information matrix?*

Have obtained designs which optimize the estimation of the probabilities $\pi_2, \ldots, \pi_k$ over all $\mathbf{x} \in \mathcal{X}$, using an Integrated Mean Square Error criterion.

But that's for another time and place.