A Note on Bayesian D-optimality

Matt Simpson

September 8, 2016

1 Introduction

This is a quick note outlining the class of problems we need to look for in order to use PSO for finding Bayesian D-optimal designs, or nearly optimial designs. Strictly speaking, D-optimality is more narrow than necessary.

2 General Framework

Suppose we wish to design an experiment with n experimental units in order to determine the effect of some covariates on some response y_i, $i = 1, 2, \ldots, n$. Let $\boldsymbol{y} = \boldsymbol{y}_{1:n} = (y_1, y_2, \ldots, y_n)'$. The covariates are split into two types — fixed covariates and adjustable covariates. Fixed covariates are immutable features of the experimental units, while adjustable covariates are features which are chosen by the experimenter. For example, any treatment applied to the experimental unity is an adjustable covariate. Let $\boldsymbol{X}^f$ denote the $n \times p$ matrix of fixed covariates, $\boldsymbol{X}^a$ denote the $n \times q$ matrix of adjustable covariates, and $\boldsymbol{X} = [\boldsymbol{X}^f, \boldsymbol{X}^a]$ the full $n \times (p+q)$ matrix of covariates. Finally let $\boldsymbol{\theta}$ denote a vector of model parameters. Then the generic model is

$$\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta} \sim [\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}], \qquad \boldsymbol{\theta}|\boldsymbol{X} \sim [\boldsymbol{\theta}]$$

where $[.|.]$ denotes the probability density or mass function of the enclosed random variables. Often $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$, $\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta} \sim [\boldsymbol{y}|\boldsymbol{\mu}, \phi]$ with $g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}$ and $g(.)$ is a known link function.

The design problem is to choose $\boldsymbol{X}^a \in \mathcal{X}$ in order to maximize some design criterion. The space $\mathcal{X}$ contains all constraints on the elements of $\boldsymbol{X}^a$, for example if $x_{i2} = x_{i1}^2$ in a quadratic regression model. What is being chosen here is not which covariates to include, but rather the values of the covariates. For example in a simple linear regression model with $\boldsymbol{X}^f = (1, 1, \ldots, 1)'$ correpsonding to the intercept and $\boldsymbol{X}^a = (x_1, x_2, \ldots, x_n)$ corresponding to the slope, the design problem is to choose the values of $x_1, x_2, \ldots, x_n$ for a fixed n, typically where x_i lives in a constrained space.

A standard Bayesian design criterion is the expected shannon information gain:

$$\mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}}\left\{\log\frac{[\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X}]}{[\boldsymbol{\theta}]}\right\} = \iint \log\frac{[\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X}]}{[\boldsymbol{\theta}]}[\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}][\boldsymbol{\theta}]d\boldsymbol{\theta}d\boldsymbol{y}.$$

Maximizing this in $\boldsymbol{X}^a$ is equivalent to maximizing

$$U(\boldsymbol{X}^a) = \iint \log\frac{[\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}]}{[\boldsymbol{y}|\boldsymbol{X}]}[\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}][\boldsymbol{\theta}]d\boldsymbol{\theta}d\boldsymbol{y}.$$

When $\boldsymbol{\theta} \sim N(\boldsymbol{\theta}_0, \sigma^2\boldsymbol{S}_0)$ and $\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta} \sim N(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2\boldsymbol{I}_n)$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and σ^2 is known, then maximizing $U(\boldsymbol{X}^a)$ is equivalent to maximizing the Bayesian D-optimality criterion:

$$D(\boldsymbol{X}^a) = |\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{S}_0^{-1}|.$$

However in general, maximizing U and D are not equivalent, and when they are not, typically $U(\boldsymbol{X}^a)$ must be approximated via monte carlo simulation as well as an approximation for the model's marginal likelihood, $[\boldsymbol{y}|\boldsymbol{X}]$.

Whether the model is U or D, finding the optimal design is a difficult optimization problem, and near-optimal designs are often desireable. As such, this is a perfect problem for heuristic optimization algorithms such as particle swarm optimization. The problem for us, then, is to find an example that fits into this framework and ideally uses federal data.

3 Bayesian D-Optimality for Choice Experiments

A wide range of fields from economics to marketing use discrete choice experiments in order to evaluate the preferences of individuals about various items in a choice set. Economists might be interested in, for example, nonmarket valuation — e.g. attaching a dollar value to environmental benefits that do not have a market price, while marketers might be interesting in finding the optimal product mix in some industry.

The discrete choice model framework these experiments use is as follows. Let $i = 1, 2, \ldots, I$ indicate the individual choosers, $j = 1, 2, \ldots, J_i$ indicate the choice sets that chooser i sees, and $k = 1, 2, \ldots, K_{ij}$ indicate the options available to chooser i in choice set j. Then we represent the utility of option k of set j for individual i as $u_{ijk} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \varepsilon_{ijk}$ Here $\boldsymbol{x}'_{ijk}$ is a vector of covariates containing information about the option available to the chooser in this choice, and also possibly containing information about the chooser such as demographics. Typically $\varepsilon_{ijk} \overset{iid}{\sim} F$ where F is some known cdf, often the Gumbel cdf, or the normal cdf. Especially in the case of a normal cdf, the independence assumption can be relaxed though this can cause identifiability issues. The utility of each option is unobserved, but option i chooses is observed. If option k^* is chosen, this implies that $u_{ijk^*} = \max_k u_{ijk}$. Under *iid* Gumbel errors, the probability that option k is chosen is given by

$$p_{ijk} = \frac{e^{\boldsymbol{x}'_{ijk}\boldsymbol{\beta}}}{\sum_{t=1}^{K_{ij}} e^{\boldsymbol{x}'_{ijt}\boldsymbol{\beta}}},$$

which gives rise to the multinomial logit (MNL) model. Let $y_{ijk} = 1$ if chooser i chooses option k of set j, and $y_{ijk} = 0$ if they choose any other option in the set. Then letting $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijK_{ij}})$ and similarly for $\boldsymbol{p}_{ij}$, we have $\boldsymbol{y}_{ij} \overset{ind}{\sim} \text{Multinomial}(1, \boldsymbol{p}_{ij})$.

The problem, then, is to choose which choice sets are presented to which choosers in order to maximize some design criterion, typically in order to maximize the information learned about the parameter $\boldsymbol{\beta}$. Most design criterion used in the literature are based on the Fisher information (FI) matrix. The FI matrix can be written as

$$I(\boldsymbol{X}, \boldsymbol{\beta}) = -\sum_{i=1}^{I}\sum_{j=1}^{J_i} \boldsymbol{X}'_{ij}(\boldsymbol{P}_{ij} - \boldsymbol{p}_{ij}\boldsymbol{p}'_{ij})\boldsymbol{X}_{ij}$$

where $\boldsymbol{X}_{ij} = (\boldsymbol{x}_{ij1}, \boldsymbol{x}_{ij2}, \ldots, \boldsymbol{x}_{ijK_{ij}})'$, $\boldsymbol{p}_{ij} = (p_{ij1}, p_{ij2}, \ldots, p_{ijK_{ij}})'$, and $\boldsymbol{P}_{ij} = \text{diag}(\boldsymbol{p}_{ij})$. The matrix does not depend on the values of the y_{ijk}s, so the expected information matrix is the same. Typically the goal is to maximize $\det[-I(\boldsymbol{X}, \boldsymbol{\beta})]$ or some function of it. The FI matrix is defined for a specfic value of $\boldsymbol{\beta}$, so there are two approaches: choose a reasonable value of $\boldsymbol{\beta}$

and find the optimal design for that particular value, or put a prior on $\boldsymbol{\beta}$ and maximize the expected value of the FI with respect to that prior. A Bayesian entropy based design criterion can also be used, though I do not think this corresponds to maximizing the a priori expected FI.

The main problem with these design problems, from our perspective, is that the choice sets live in a discrete space. Typically it involves varying one or more factors on a finite number of levels. Any continuous factors, such as price, are reduced to a small number of possible factors. So if we use this sort example, we might not have a ready-made example.