GreedyExperimentalDesign: Finding Experimental Designs using Greedy Search with Random Restarts

Adam Kapelner
Queens College, City University of New York
Department of Mathematics

Abstract

Keywords: experimental design, greedy search, optimization, R, Java.

1. Introduction

Assume a randomized controlled two-arm experiment with n subjects and treatment (T) and control (C) denoted by the n-binary vector $\mathbf{1}_T$ where entries of 1 in location i indicates subject i was administered T and entries of 0 indicates C. Define the number of treatments $n_T := \sum_{i=1}^n \mathbf{1}_{T,i}$ and the number of controls $n_C := n - n_T$. For each subject, p covariates $\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$ are measured. Define $\bar{\boldsymbol{X}}_T$ as the p-vector of sample averages for each of the covariates in subjects where $\mathbf{1}_T = 1$ (the treatments) and $\bar{\boldsymbol{X}}_C$ as the p-vector of sample averages for each of the covariates in subjects where $\mathbf{1}_T = 0$ (the controls). The investigator will eventually measure one response for each subject collected in the n-vector $\boldsymbol{y}$, but this is not our current interest. We assume that each of the p covariates is standardized.

There are many functions of $\mathbf{1}_T$ and $\boldsymbol{X}$ that will yield higher efficiency when testing null hypotheses about effects of the treatment. Below are a few:[1]

1. n_T/n which measures the balance of treatment allocations. 0.5 is the optimal value.

2. $\sum_{j=1}^p \left|\bar{\boldsymbol{X}}_{T,j} - \bar{\boldsymbol{X}}_{C,j}\right|$ which is a measure of balance between the covariate distributions. Covariate distribution permitting, zero is the optimal value.

3. $\frac{n_T n_C}{n} \left(\bar{\boldsymbol{X}}_T - \bar{\boldsymbol{X}}_C\right)^\top \boldsymbol{S}_{\boldsymbol{X}}^{-1} \left(\bar{\boldsymbol{X}}_T - \bar{\boldsymbol{X}}_C\right)$ is a Mahalanobis-like distance metric. Covariate distribution permitting, zero is the optimal value.

For many of our proposals below we will fix n_T/n to be 0.5 and then minimize one of the other two objective functions.

[1] There are also metrics which measure the similarity between the two joint densities f_T and f_C which we may want to explore later.

2. Greedy Switches Algorithm

Draw one vector from the space of $\binom{n}{n/2}$ possible balanced $\mathbf{1}_T$ vectors. Create a list of the indices of size $n/2$ corresponding to where $\mathbf{1}_T = 1$ (call it I_T). Create a list of the indices of size $n/2$ corresponding to where $\mathbf{1}_T = 0$ (call it I_C). For every pair in $I_T \times I_C$, switch the 0 and 1 within $\mathbf{1}_T$ and record the resulting value of the objective function. For all possible $n^2/4$ possible switches (of which all preserve $n_T/n = 0.5$), find the switch which yielded the minimum value of the objective function. Make that switch inside $\mathbf{1}_T$. Continue in this fashion until you can no longer improve the objective value.

Replication

The stable version of **GreedyExperimentalDesign** will be soon on CRAN and the development version is located at `https://github.com/kapelner/GreedyExperimentalDesign`. The package code is under the GPL3 and LGPL licenses. Results, tables, and figures found in this paper can be replicated via the scripts located in the `GreedyExperimentalDesign/vignettes` folder within the `git` repository.

Acknowledgements

We thank Abba Krieger and David Azriel for helpful discussions. We thank Simon Urbanek for his very generous help with **rJava**.

Affiliation:

Adam Kapelner
Department of Mathematics
Queens College, City University of New York
64-19 Kissena Blvd Room 325
Flushing, NY, 11367
E-mail: `kapelner@qc.cuny.edu`
URL: `http://kapelner.com`