Computational Statistics and Data Analysis 126 (2018) 112–124

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

An induced natural selection heuristic for finding optimal Bayesian experimental designs

David J. Price [a,b,c,*], Nigel G. Bean [d,e], Joshua V. Ross [d,e], Jonathan Tuke [d,e]

[a] Disease Dynamics Unit, Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge CB3 0ES, United Kingdom
[b] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, VIC 3010, Australia
[c] Victorian Infectious Diseases Reference Laboratory Epidemiology Unit at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne and Royal Melbourne Hospital, VIC 3000, Australia
[d] School of Mathematical Sciences, University of Adelaide, SA 5005, Australia
[e] ARC Centre of Excellence for Mathematical & Statistical Frontiers, School of Mathematical Sciences, University of Adelaide, SA 5005, Australia

ARTICLE INFO

Article history:
Received 8 March 2017
Received in revised form 26 April 2018
Accepted 26 April 2018
Available online 5 May 2018

Keywords:
Bayesian optimal design
Optimisation heuristic
Stochastic models
Sampling windows

ABSTRACT

Bayesian optimal experimental design has immense potential to inform the collection of data so as to subsequently enhance our understanding of a variety of processes. However, a major impediment is the difficulty in evaluating optimal designs for problems with large, or high-dimensional, design spaces. An efficient search heuristic suitable for general optimisation problems, with a particular focus on optimal Bayesian experimental design problems, is proposed. The heuristic evaluates the objective (utility) function at an initial, randomly generated set of input values. At each generation of the algorithm, input values are "accepted" if their corresponding objective (utility) function satisfies some acceptance criteria, and new inputs are sampled about these accepted points. The new algorithm is demonstrated by evaluating the optimal Bayesian experimental designs for the previously considered death, pharmacokinetic and logistic regression models. Comparisons to the current "gold-standard" method are given to demonstrate the proposed algorithm as a computationally-efficient alternative for moderately-large design problems (i.e., up to approximately 40-dimensions).

1. Introduction

Optimising the design of an experiment is an important consideration in many areas of science, including, but not limited, to: biology (Faller et al., 2003), clinical trials (Berry, 2004) and epidemiology (Pagendam and Pollett, 2013). The theory of optimal experimental design is a statistical framework that allows us to determine the optimal experimental protocol to gain the most information about model parameters, given constraints on resources.

In evaluating an optimal Bayesian design, there are two main components: the search across the design space, and the evaluation of the utility. There have been many approaches to improving the efficiency of both aspects, summarised by Ryan et al. (2015a). Recently, Overstall and Woods (2017) proposed the Approximate Coordinate Exchange (ACE) algorithm to address the search aspect of the Bayesian experimental design problem. The method utilises a coordinate

* Corresponding author at: Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, VIC 3010, Australia.
E-mail address: david.j.price@alumni.adelaide.edu.au (D.J. Price).

https://doi.org/10.1016/j.csda.2018.04.011
0167-9473/

exchange algorithm to update one dimension of the design at a time, coupled with a Gaussian process in order to search each dimension efficiently. It has been asserted that the future of optimal Bayesian experimental design lies in the ability to evaluate the optimal designs for large-scale problems (*i.e.*, large or high-dimensional design spaces), in a computationally-efficient manner (Ryan et al., 2015a). In this paper, we address this by proposing a new search algorithm targeted at finding optimal Bayesian experimental designs.

The search heuristic we present performs targeted sampling of the design space to find high utility designs, without making any assumptions about the shape of the utility function. An initial population of random designs is generated — synonymous with multiple algorithm runs from random initial conditions as in other optimisation routines. Our method borrows the idea of targeting regions of high utility, as per the MCMC approach of Müller (1999), by sampling new designs at each iteration around the "best" designs; chosen according to some acceptance criteria. We describe this algorithm using the notion of "survival-of-the-fittest", as the "fittest" individuals –according to their objective (utility) function value – survive at each iteration (generation) based on a user-defined acceptance criteria, to produce offspring for the next generation. Hence, we propose this as a new type of evolutionary algorithm (*e.g.*, Goldberg, 1989), and refer to it herein as the Induced Natural Selection Heuristic (INSH).

By independently sampling new designs around each accepted design, we aim to avoid the pitfalls associated with some other optimisation routines. For example, INSH is able to sample multiple regions of high utility at a time, thus exploring multiple local optima simultaneously, rather than potentially being stuck at a single local optima. Furthermore, by not combining the retained designs in any way, INSH avoids the potential to move to a region of low utility that is at the "centre" of multiple local optima — as may occur in a cross-entropy or genetic algorithm. By taking a sampling approach, as opposed to trying to approximate the function, INSH makes no assumptions about the shape of the utility function — thus, it is not limited to utility functions that are, for example, smooth. Utilising (embarrassingly) parallel computation tools, the method can efficiently evaluate the utility for a large number of designs in each iteration.

The ACE algorithm has allowed the consideration of Bayesian optimal designs for a larger, more-complex class of statistical models and experiments than was possible with previous algorithms. There are a number of drawbacks to ACE, however. By searching in one-dimension at a time, ACE risks missing the globally-optimal design, and instead may find only local optima. An approach to avoid this is to re-run the algorithm from a number of randomly generated initial designs (Overstall and Woods, 2017). Similarly, as noted by the authors, by searching in one-dimension at a time, the algorithm will be inefficient in scenarios where there is a large correlation between the design variables — a problem which adds to the difficulty in choosing a suitable number of iterations for each phase of the algorithm. The algorithm requires a sufficiently-good estimate of the utility when determining whether to accept the candidate design — spurious estimates may lead to sub-optimal candidate designs being accepted, and thus push the algorithm away from regions of high utility. Alternatively, a large improvement in the computation time arises from the estimation of the utility surface in each dimension in the form of a Gaussian process based on a number of candidate points. This approximation to the utility surface based on noisy evaluations of the utility aims to provide a smooth approximation to the surface. When the surface is not smooth, or has a discontinuity (*e.g.*, as exists in the utility surface for the death model in Fig. 2(a) at $\boldsymbol{t} \approx (2.75, t_2)$ and $\boldsymbol{t} \approx (t_1, 2.75)$), this has the potential to cause problems for the ACE algorithm.

In the following, we present the INSH search algorithm in a general framework, and we note that efficient evaluation of the utility is another problem that needs to be addressed. We consider two existing approaches to evaluating the utility: an Approximate Bayesian Computation (ABC) approach used by Price et al. (2016), in a scenario where the benefits of this approach are realised; and a nested Monte-Carlo approximation using code from the `acebayes` package (Overstall et al., 2017), otherwise.

We consider the problem of finding the optimal design for the death model, a pharmacokinetic (PK) model tracking the concentration of a drug or treatment in the blood, and a four-factor logistic regression model. In the death and PK examples, a design d consists of n sampling times $(t_1, \ldots, t_n)$, subject to some problem-specific constraints. First, we address the question of when to observe the stochastic process in order to gain the most information about the model parameters governing the death model. The Markovian death model has been considered previously in a Bayesian framework by Cook et al. (2008), Drovandi and Pettitt (2013), and Price et al. (2016). We compare the optimal designs for 1–4 observation times in order to demonstrate the efficacy of the method. Second, we consider the question of sampling times for a PK model – a process where the design space is higher-dimensional – in order to demonstrate the efficiency of the INSH algorithm for larger design spaces. The optimal designs are compared to those evaluated using the "gold-standard" Approximate Coordinate Exchange (ACE) algorithm of Overstall and Woods (2017). We also consider the idea of sampling windows for this example, which have been considered previously by Green and Duffull (2003), Chenel et al. (2005), Graham and Aarons (2006), McGree et al. (2012), and Duffull et al. (2012), for example. Finally, we compare the results of the INSH algorithm to those of the ACE algorithm for a standard four-factor logistic regression model (Overstall and Woods, 2017) — a considerably higher-dimensional problem. We consider examples with $n = 6$, 10, 24, and 48 (independent) replicates in each experiment; corresponding to a design space with up to 192 dimensions (i.e., when $n = 48$ replicates).

1.1. Bayesian optimal experimental design

The aim of optimal experimental design is to determine the best experimental protocol in order to maximise some utility of the experiment. To achieve this aim, we specify a utility function $U(\boldsymbol{\theta}, \boldsymbol{y}, d)$ representing how we 'value' the experimental