ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Optimal Bayesian design for discriminating between models with intractable likelihoods in epidemiology

Mahasen B. Dehideniya [a,*], Christopher C. Drovandi [a,b], James M. McGree [a,b]

[a] School of Mathematical Sciences, Queensland University of Technology (QUT), GPO Box 2434, Brisbane, QLD 4001, Australia
[b] ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), Australia

ARTICLE INFO

Article history:
Received 21 January 2017
Received in revised form 6 March 2018
Accepted 6 March 2018
Available online xxxx

Keywords:
Approximate Bayesian computation
Ds-optimality
Model discrimination
Mutual information
Parameter estimation
Coordinate exchange algorithm
Zero–one utility

ABSTRACT

A methodology is proposed to derive Bayesian experimental designs for discriminating between rival epidemiological models with computationally intractable likelihoods. Methods from approximate Bayesian computation are used to facilitate inference in this setting, and an efficient implementation of this inference framework for approximating the expectation of utility functions is proposed. Three utility functions for model discrimination are considered, and the performance of these utilities is explored in designing experiments for discriminating between three epidemiological models; the death model, the Susceptible–Infected model, and the Susceptible–Exposed–Infected model. The challenge of efficiently locating optimal designs is addressed by an adaptation of the coordinate exchange algorithm which exploits parallel computational architectures.

1. Introduction

Epidemiological studies are important for understanding how a disease is transmitted, and for the development of preventative measures which might reduce or limit the spread of the disease. Informative data collection is crucial in developing this understanding, and can be achieved by conducting an experiment according to an optimal design that provides the maximum amount of information to address the aim of the experiment which could include model selection, parameter estimation and prediction. However, the derivation of optimal designs in epidemiological experiments is a challenging task as most epidemiological models contain likelihoods which are computationally expensive to evaluate (Becker, 1993). Consequently, only a few attempts have been made in both the frequentist literature (Pagendam and Pollett, 2013) and the Bayesian literature (Cook et al., 2008; Drovandi and Pettitt, 2013) to derive optimal designs for experiments in epidemiology.

In the frequentist literature, the design of epidemiological experiments has been facilitated via an approximation to the likelihood. Pagendam and Pollett (2013) used a Gaussian diffusion approximation in deriving D-optimal experimental designs to estimate parameters of the SI (Susceptible–Infected), SIS (Susceptible–Infected–Susceptible) and SIR (Susceptible–Infected–Recovered) epidemic models. The designs derived in this work were dependent upon point estimates of the parameter values, and are thus termed locally optimal designs. In contrast, the Bayesian approach provides a framework to account for the uncertainty in parameters when deriving optimal designs (Ryan, 2003). This was demonstrated in the work of Cook et al. (2008) who derived optimal observation times for parameter estimation of the death model and the SI model. In their work, the moment closure method was used to approximate the likelihood of the SI model.

Recent developments in approximate Bayesian computation (ABC) provide a comprehensive framework to undertake Bayesian inference and design when the likelihood is intractable. Drovandi and Pettitt (2013) presented a likelihood-free

* Corresponding author.
E-mail address: d.mahasen@hdr.qut.edu.au (M.B. Dehideniya).

https://doi.org/10.1016/j.csda.2018.03.004
0167-9473/

method to derive Bayesian designs for parameter estimation of Markov process models of epidemics and macroparasite population evolution using the ABC rejection method (Beaumont et al., 2002). In the work of Price et al. (2016), ABC rejection was used to approximate a utility function based on the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) in designing experiments for parameter estimation of epidemiological models.

Previous work in the design of epidemiological experiments has focussed on estimating model parameters of an assumed true model to describe the process of interest (Cook et al., 2008; Drovandi and Pettitt, 2013; Pagendam and Pollett, 2013). However, in reality, there may be uncertainty about the true epidemiological process (see Lee et al., 2015), and indeed, the purpose of the experiment could be to determine how a disease spreads. Hence, the lack of knowledge about the true model should be taken into account when designing efficient experiments. Thus, the need for the development of new methods to design efficient epidemiological experiments for model discrimination motivates the work described in this article. Here, we consider the design problem of locating a set of observation times which yields information to efficiently discriminate between competing models. Moreover, previous work on designing experiments for model discrimination (Atkinson and Fedorov, 1975; Cavagnaro et al., 2010; Drovandi et al., 2014; Woods et al., 2017; Overstall et al., 2018) were limited to models where the likelihood can be easily computed. Thus, this is the first paper to propose methods for finding Bayesian optimal designs for discriminating between models with intractable likelihoods.

Finding the optimal design for an experiment requires the maximisation of an expected utility over all possible designs, and it is a challenging optimisation problem because the utility surface is noisy and may be relatively flat around its maximum. Further, it can be computationally prohibitive to undertake the optimisation even for experiments with a moderate number of design variables (see the review by Ryan et al. (2016)). Müller (1999) proposed a simulation-based approach that converts the optimisation problem to a problem of sampling from a target distribution for which the mode is the optimal design. First, samples are drawn from the target distribution $h(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})$ (joint distribution of the parameters, data, and design) using Markov chain Monte Carlo (MCMC) simulations, and then the estimated multivariate mode of the marginal distribution of $\boldsymbol{d}$ is deemed the optimal design. The Müller algorithm has been widely used in the Bayesian experimental design literature (Stroud et al., 2001; Cook et al., 2008; Drovandi and Pettitt, 2013; Ryan et al., 2014). However, in practice, this method suffers from slow convergence. Moreover, sampling from the joint distribution $h(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})$ using an MCMC method and determining the multivariate mode for a large number of design variables are computationally expensive tasks (Drovandi and Pettitt, 2013).

Alternatively, existing local search optimisation methods can be used to locate the optimal design. For instance, the coordinate exchange (CE) algorithm of Meyer and Nachtsheim (1995) has been used to find D-optimal designs in screening experiments by Goos and Jones (2011) and Palhazi Cuervo et al. (2016). Further, Gotwalt et al. (2009) used the coordinate exchange algorithm in constructing pseudo-Bayesian optimal designs for parameter estimation of non-linear models. The coordinate exchange algorithm starts from a given initial design and iteratively maximises the utility function by changing one design variable at a time while keeping all other variables fixed. This iterative procedure continues until there is little or no improvement in the value of the utility. In practice, this may require a large number of utility evaluations, especially when continuous design variables are involved in the experiment. Recent work of Overstall and Woods (2017) extends the idea of the coordinate exchange algorithm by considering an approximation of the expected utility as a function of a single design variable conditional on the remaining fixed variables. This approximation is facilitated by fitting a Gaussian process emulator based on a relatively small number of utility evaluations. This emulator is then used to approximate the utility function across the entire range of the considered variable and to estimate the maximum at each iteration.

In this work, evaluating the approximate utility of a given design is highly computational as it requires a large number of simulations from the model in order to approximate a posterior distribution via ABC methods. Consequently, finding Bayesian optimal designs for models with intractable likelihoods in a continuous design space could be computationally prohibitive. However, the use of a discrete design space to locate optimal designs significantly reduces the required computational effort as it allows the use of pre-simulated data for the posterior approximations in utility evaluations (discussed later). This idea has been used by Drovandi and Pettitt (2013) within the Müller algorithm and by Price et al. (2016) who finds the optimal design using an exhaustive search. However, these methods quickly become computationally intensive as the number of design dimensions increases. Hence, in this setting, it would be advantageous to reduce the required number of utility evaluations when searching for the optimal design. For this purpose, we propose using the refined coordinate exchange algorithm where, at each iteration of the exchange algorithm, the coordinate space reduces and becomes more refined. Further, the algorithm is structured such that parallel computational architectures can be exploited. As will be seen, through using this algorithm, we are able to efficiently locate Bayesian designs in higher dimensions than previously explored in the design literature related to models with intractable likelihoods.

The paper is organised as follows. In the next section, the problem of model choice in the Bayesian framework is described. Section 3 presents the utility functions used in this work, and Section 4 describes the ABC methods that are used for inference and in estimating the expected utility of a given design. An adapted version of the coordinate exchange algorithm which exploits parallel computational architectures is presented in Section 5. In Section 6, the design for a pharmacokinetic model is considered to explore and demonstrate the performance of our proposed optimisation algorithm. Following this, two epidemiological examples are considered to demonstrate the performance of three utility functions, namely the mutual information utility, the Ds-optimal utility and the Zero–One (0–1) utility for model discrimination. The paper concludes with a discussion and suggestions for further research.

Please cite this article in press as: Dehideniya M.B., et al., Optimal Bayesian design for discriminating between models with intractable likelihoods in epidemiology. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.03.004.