Bayesian design of experiments for intractable likelihood models using coupled auxiliary models and multivariate emulation

Antony M. Overstall
Southampton Statistical Sciences Research Institute,
University of Southampton,
Southampton, UK
(A.M.Overstall@soton.ac.uk)

James M. McGree
School of Mathematical Sciences,
Queensland University of Technology,
Brisbane, Australia
(james.mcgree@qut.edu.au)

Abstract

A Bayesian design is given by maximising the expected utility over the design space. The utility is chosen to represent the aim of the experiment and its expectation is taken with respect to all unknowns: responses, parameters and/or models. Although straightforward in principle, there are several challenges to finding Bayesian designs in practice. Firstly, the expected utility is rarely available in closed form and requires approximation. Secondly, the expected utility needs to be maximised over a, potentially, high-dimensional design space. In the case of intractable likelihood models, these problems are compounded by the fact that the likelihood function, whose evaluation is required to approximate the expected utility, is not available in closed form. A strategy is proposed to find Bayesian designs for intractable likelihood models. It relies on the development of new methodology involving auxiliary modelling to approximate the expected utility, under an intractable likelihood model, applied with the latest approaches to maximising approximated expected utilities.

1 Introduction

1.1 Background

Often, the dynamics underpinning a complex physical phenomenon can be modelled by a stochastic process or processes. It is commonly the situation that these processes (or models) depend on unknown parameters and, potentially, other controllable variables. In this paper, we consider the case where an experiment is to be performed to learn about the phenomenon by either estimating the unknown parameters of a particular model or comparing different models. That is, the physical phenomenon of interest is observed at a series of time points, after the specification of any controllable variables, and the stochastic model (models) is (are) *fitted* to the observed responses. In particular, we focus on the optimal choice of time points and controllable variables (collectively referred to as design variables) to best learn about the unknown process or processes.

A key characteristic of the stochastic models studied in this paper is that, although the dynamics behind each process can be relatively simple, the probability model linking parameters and design variables to responses is typically only defined implicitly. This is usually termed an intractable likelihood model. The

development of new statistical methodology, so called *likelihood-free* methodology, to analyse observed responses under an intractable likelihood model has received much attention in recent years, e.g. approximate Bayesian computation (Tavaré et al. 1997); synthetic likelihood (Wood, 2010); variational Bayes (Tran et al., 2017) and auxiliary modelling (Gourieroux et al. 1993).

The task of designing the experiment, i.e. specifying the design variables, has received significantly less attention. Under the frequentist approach to statistical inference, Pagendam and Pollett (2013) and Parker et al. (2015) used numerical approximations to the Fisher information to find D-optimal designs (e.g., Atkinson et al., 2007, Chapter 11) for stochastic epidemic and queueing models, respectively. In this paper, the Bayesian approach to statistical inference is used. Under such an approach, a Bayesian optimal design (Chaloner and Verdinelli, 1995) is found by maximising the expectation of a utility function with respect to all unknown quantities (i.e. models, parameters and unobserved responses) where the utility function is chosen to represent the aim of the experiment, e.g. parameter estimation or model comparison. Finding optimal designs under the Bayesian approach, even for tractable likelihood models, is a significant computational challenge; see recent reviews of the field by Ryan et al. (2016b) and Woods et al. (2017). For intractable likelihood models, approaches for finding optimal designs have been proposed that use approximate Bayesian computation (Drovandi and Pettitt 2013, Hainy et al. 2013, Price et al. 2016) and auxiliary modelling (also known as indirect inference; Ryan et al. 2016a). A common feature of these methodologies is that they have only been applied for examples of experiments with very small number of design variables rendering them of limited practical relevance.

In this paper, we aim to overcome the restrictions of these approaches. The contribution is to combine the auxiliary modelling approach with multivariate statistical emulators and the latest approaches for finding Bayesian optimal designs under tractable likelihood models. The result is an approach that can find approximate Bayesian optimal designs for intractable likelihood models for experiments with design spaces of realistic dimensionality. We illustrate the methodology on a series of illustrative examples from the biological sciences which include experiments with aims of parameter estimation and model comparison.

1.2 Setup

Suppose the experiment consists of n runs. For $k = 1, \ldots, n$, the kth run involves the specification of a $w \times 1$ vector of design variables $\mathbf{d}_k \in \mathcal{D}$. Let y_k be the corresponding observed response from the phenomenon for the kth run. It is assumed that, independently,

$$y_k \sim \mathcal{F}\left(\boldsymbol{\theta}, \mathbf{d}_k\right), \tag{1}$$

where $\mathcal{F}$ is a distribution depending on a $p \times 1$ vector of unknown parameters $\boldsymbol{\theta} \in \Theta$, with Θ the parameter space. Let $f(y|\boldsymbol{\theta}, \mathbf{d})$ denote the probability density function (pdf) or probability mass function (pmf) of the distribution $\mathcal{F}$. The likelihood is then given by

$$\pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}) = \prod_{k=1}^{n} f(y_k|\boldsymbol{\theta}, \mathbf{d}_k),$$

where $\mathbf{y} = (y_1, \ldots, y_n)$ is the $n \times 1$ vector of responses and $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_n) \in \Delta = \mathcal{D}^n$ is the $q \times 1$ vector giving the design, with $q = nw$ the total number of design variables and Δ the design space. For the models considered in this paper, $\mathcal{F}$ is only defined implicitly with the result that $f(y|\boldsymbol{\theta}, \mathbf{d})$ and $\pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D})$ are not available in closed form.

1.3 Example: Lotka-Volterra System

We now present the Lotka-Voltera (LV) system as a simple example of an intractable likelihood model. The LV system is frequently used to describe the dynamics of population sizes when two species (prey and predators) interact. Let $R(t)$ and $S(t)$ denote the population sizes of prey and predators, respectively, at

time t. The dynamics are given by the following equations

$$
\begin{aligned}
\mathrm{P}\left(R(t+\delta t)=R(t)+1, S(t+\delta t)=S(t)\right) &= \theta_1 R(t)\delta t + o(\delta t), \\
\mathrm{P}\left(R(t+\delta t)=R(t)-1, S(t+\delta t)=S(t)+1\right) &= \theta_2 R(t)S(t)\delta t + o(\delta t), \\
\mathrm{P}\left(R(t+\delta t)=R(t), S(t+\delta t)=S(t)-1\right) &= \theta_3 S(t)\delta t + o(\delta t),
\end{aligned}
$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ are the unknown parameters. The prey are assumed to have unlimited food supply at all times and, unless subject to predation, experience exponential growth with rate $\theta_1 R(t)$. The rate of predation is assumed to be proportional to the rate at which prey and predators meet, i.e. $\theta_2 R(t)S(t)$. Finally, the predators die with rate $\theta_3 S(t)$. Suppose that the experiment consists of observing the number of prey at a series of times $t_1, \ldots, t_n$, i.e. we observe $y_i = S(t_i)$, for $i = 1, \ldots, n$, with the aim of estimating the parameters $\boldsymbol{\theta}$. Using the terminology from Section 1.2, in this example, $w = 1$, $\mathbf{d}_k = t_k$, $q = n$ and $\mathbf{D} = (t_1, \ldots, t_n)$. For given values of the initial population sizes, $R(0)$ and $S(0)$, parameters $\boldsymbol{\theta}$ and time t, the probability distribution of $S(t)$, i.e. $\mathcal{F}(\boldsymbol{\theta}, t)$, is not available in closed form.

1.4 Bayesian optimal design of experiments

We initially describe the concept of Bayesian optimal design of experiments for the experimental aim of parameter estimation. Extension to model comparison is conceptually straightforward (see Section 3.4). Bayesian optimal design of experiments begins with the specification of a utility function denoted by $u(\boldsymbol{\theta}, \mathbf{y}, \mathbf{D})$ which represents the *utility* of estimating $\boldsymbol{\theta}$ using observed responses $\mathbf{y}$ generated via design $\mathbf{D}$. A Bayesian optimal design is given by maximising (over Δ) the expected utility function given by

$$
U(\mathbf{D}) = \int u(\boldsymbol{\theta}, \mathbf{y}, \mathbf{D}) \mathrm{dP}_{\boldsymbol{\theta}, \mathbf{y} | \mathbf{D}}, \tag{2}
$$

where the expectation is with respect to the joint distribution of all unknown quantities; $\boldsymbol{\theta}$ and $\mathbf{y}$. In this paper we consider a class of utility functions which we term *likelihood-based*. This is where the utility function is a functional of the likelihood, $\pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D})$, and/or the marginal likelihood (also known as evidence) given by

$$
\pi(\mathbf{y}|\mathbf{D}) = \int_\Theta \pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}) \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.
$$

We write the likelihood-based utility as $u(\boldsymbol{\theta}, \mathbf{y}, \mathbf{D}) = u^*(\pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}), \pi(\mathbf{y}|\mathbf{D}))$. This class of utility function includes many commonly-employed utilities for both parameter estimation and model comparison. As an example, consider the Shannon information gain (SIG; Lindley 1956) given by

$$
\begin{aligned}
u_S(\boldsymbol{\theta}, \mathbf{y}, \mathbf{D}) &= u_S^*(\pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}), \pi(\mathbf{y}|\mathbf{D})) \\
&= \log \pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}) - \log \pi(\mathbf{y}|\mathbf{D}). \tag{3}
\end{aligned}
$$

The Bayesian design under the SIG utility is equivalently the design that maximises the expected (with respect to the marginal distribution of $\mathbf{y}$) Kullback-Liebler divergence between the prior and posterior distributions of $\boldsymbol{\theta}$.

Although conceptually straightforward, there are at least two hurdles to finding Bayesian designs in practice (even for tractable likelihood models). Firstly, the integration required to evaluate $U(\mathbf{D})$ is rarely analytically tractable and will require approximation. It is also usually the case that the utility function itself is analytically intractable. Secondly, the design space Δ can be of high dimensionality, i.e. q can be relatively large.

The methods proposed in the literature for approximately maximising the expected utility can be broadly classified into simulation- or smoothing-based. The simulation-based approach of Müller (1999) places an artificial joint distribution on $\boldsymbol{\theta}$, $\mathbf{y}$ and $\mathbf{D}$ such that the marginal pdf of $\mathbf{D}$ is proportional to $U(\mathbf{D})$. Simulation methods are then used to generate a sample from this joint distribution which is then used to estimate the marginal mode of $\mathbf{D}$, i.e. the optimal design. This method has been further refined, for example, by Müller

et al. (2004) and Amzal et al. (2006). However due to the difficulties in efficiently sampling over a high dimensional space, the typical limit of dimensionality for the design space under these methods is considered to be $q = 4$ (e.g. Ryan et al. 2016b).

Smoothing-based approaches are based on the following Monte Carlo approximation to the expected utility

$$
\begin{aligned}
\tilde{U}(\mathbf{D}) &= \frac{1}{B} \sum_{i=1}^{B} u(\boldsymbol{\theta}_i, \mathbf{y}_i, \mathbf{D}), \\
&= \frac{1}{B} \sum_{i=1}^{B} u^*(\pi(\mathbf{y}_i | \boldsymbol{\theta}_i, \mathbf{D}), \pi(\mathbf{y}_i | \mathbf{D}))
\end{aligned}
\tag{4}
$$

where $\{\boldsymbol{\theta}_i, \mathbf{y}_i\}_{i=1}^{B}$ is a sample of size B generated from the joint distribution of $\boldsymbol{\theta}$ and $\mathbf{y}$ (given $\mathbf{D}$). Due to the stochastic nature of the Monte Carlo approximation, application of standard optimisation methods (e.g. Lange, 2013) is difficult. Instead, Müller and Parmigiani (1995) proposed a method whereby $\tilde{U}(\mathbf{D})$ is evaluated at a series of designs and a statistical model (a smoother or emulator) is fitted that is able to predict $\tilde{U}(\mathbf{D})$ (and therefore $U(\mathbf{D})$) at any $\mathbf{D} \in \Delta$. This predictor is then maximised over the design space, Δ. Müller and Parmigiani (1995) were able to consider design spaces with dimensionality of $q = 2$. This method has been further refined by Weaver et al. (2016) (with maximum $q = 3$) and Jones et al. (2016) (with maximum $q = 9$).

To aid in the applicability to design spaces of higher dimensionality, Overstall and Woods (2017) proposed the approximate coordinate exchange (ACE) algorithm. Here a cyclic ascent algorithm (usually referred to as coordinate exchange in the design of experiments literature; see Meyer and Nachtsheim 1995) is used to maximise the expected utility. At each of the q elements (coordinates) of the design, $\tilde{U}(\mathbf{D})$ is evaluated at a series of designs which only differ in that coordinate and a Gaussian process smoother is fitted to the resulting evaluations and used to predict $U(\mathbf{D})$ for any design. This prediction is then maximised over the one-dimensional design space of the coordinate under study. By using this methodology, Overstall and Woods (2017) were able to find approximately optimal designs for experiments in examples with design spaces of up to $q = 192$ dimensions, i.e. nearly two orders of magnitude greater than existing methods. A brief description of the ACE algorithm is given in Appendix A. It is implemented in the `acebayes` (Overstall et al., 2017) `R` package. The ACE algorithm is currently the state of the art in computing Bayesian designs and, for this reason, we use it in all examples. However the methodology we propose is suitable to use with any optimisation method which only requires the evaluation of the Monte Carlo approximation to the expected utility given by (4).

To apply any optimisation method relying on evaluation of the Monte Carlo approximation to the expected utility given by (4), it is a requirement to be able to evaluate the utility function $u(\boldsymbol{\theta}, \mathbf{y}, \mathbf{D})$. However, it is usually the case that the utility function itself is analytically intractable. Specifically, likelihood-based utilities depend on the marginal likelihood, e.g. the SIG utility given by (3), which is typically not available in closed form. The obvious approach is to use a further (inner) Monte Carlo approximation resulting in a nested or double-loop Monte Carlo approximation to the expected utility (Ryan, 2003; Huan and Marzouk, 2013; Overstall and Woods, 2017). For example, to approximate the SIG utility, we generate a further sample, $\left\{\tilde{\boldsymbol{\theta}}_j\right\}_{j=1}^{C}$, of size C from the prior distribution of $\boldsymbol{\theta}$. The SIG utility is then approximated by

$$
\begin{aligned}
\tilde{u}_S(\boldsymbol{\theta}, \mathbf{y}, \mathbf{D}) &= u_S^*(\pi(\mathbf{y} | \boldsymbol{\theta}, \mathbf{D}), \tilde{\pi}(\mathbf{y} | \mathbf{D})), \\
&= \log \pi(\mathbf{y} | \boldsymbol{\theta}, \mathbf{D}) - \log \tilde{\pi}(\mathbf{y} | \mathbf{D}),
\end{aligned}
$$

where the inner Monte Carlo approximation to the marginal likelihood is

$$
\tilde{\pi}(\mathbf{y} | \mathbf{D}) = \frac{1}{C} \sum_{j=1}^{C} \pi(\mathbf{y} | \tilde{\boldsymbol{\theta}}_j, \mathbf{D}).
\tag{5}
$$

4

1.5 Bayesian design for intractable likelihood models

Finding Bayesian designs becomes impossible under an intractable likelihood model using the methods described in Section 1.4 which rely on a large number of evaluations of the likelihood $\pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D})$ to approximate the expected utility. In the Monte Carlo approximation to the expected utility given by (4), the utility function is evaluated B times where each evaluation of the utility needs at least C evaluations of the likelihood for the inner Monte Carlo approximation to the marginal likelihood given by (5).

We assume at this point that, although we are unable to evaluate the likelihood, it is possible to generate samples from the intractable likelihood model. All models considered in this paper are examples of Markov process models. It is straightforward to generate samples from these models using the Gillespie method (Gillespie, 1977). In recent years there has been an explosion of novel methodology to evaluate the posterior distribution under an intractable likelihood depending only on the ability to generate from the model. The most popular of these methods is approximate Bayesian computation (ABC; Tavaré et al. 1997). Here, the likelihood is approximated by the ABC likelihood

$$\pi_{ABC}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}) = \int I(\delta(\tilde{\mathbf{y}}, \mathbf{y}) \leq \epsilon) \mathrm{d}\mathrm{P}_{\tilde{\mathbf{y}}|\boldsymbol{\theta}, \mathbf{D}},$$

where $I(A)$ is the indicator function for event A, $\delta(\tilde{\mathbf{y}}, \mathbf{y}) \geq 0$ is a discrepancy function (with $\delta(\tilde{\mathbf{y}}, \mathbf{y}) = 0$ if and only if $\tilde{\mathbf{y}} = \mathbf{y}$) and $\epsilon \geq 0$ is a specified tolerance. The corresponding ABC posterior distribution, given by $\pi_{ABC}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{D}) \propto \pi_{ABC}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D})\pi(\boldsymbol{\theta})$, can be sampled from by using a variety of methods such as ABC rejection (Beaumont et al., 2002), ABC MCMC (Marjoram et al., 2003), and ABC sequential Monte Carlo (Sisson et al., 2007). If $\epsilon = 0$, then the ABC posterior is equal to the true posterior distribution. However, for $\epsilon > 0$ there is typically a trade-off between choosing ϵ to be sufficiently small to ensure accurate inference and large enough for computational efficiency.

Some authors (e.g., Drovandi and Pettitt, 2013; Hainy et al., 2013; Price et al., 2016; Dehideniya et al., 2018) have used ABC to approximate the utility function when finding Bayesian designs. For example, define the ABC SIG utility as

$$
\begin{aligned}
u_{S:ABC}(\boldsymbol{\theta}, \mathbf{y}, \mathbf{D}) &= u_S^*(\pi_{ABC}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}), \pi_{ABC}(\mathbf{y}|\mathbf{D})), \\
&= \log \pi_{ABC}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}) - \log \pi_{ABC}(\mathbf{y}|\mathbf{D}),
\end{aligned}
$$

where

$$
\begin{aligned}
\pi_{ABC}(\mathbf{y}|\mathbf{D}) &= \int \pi_{ABC}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \\
&= \int I(\kappa(\tilde{\mathbf{y}}, \mathbf{y}) \leq \epsilon)\mathrm{P}_{\tilde{\mathbf{y}}, \boldsymbol{\theta}|\mathbf{D}}
\end{aligned}
$$

is the ABC marginal likelihood. The ABC SIG utility can then be approximated via

$$
\begin{aligned}
\tilde{u}_{S:ABC}(\boldsymbol{\theta}, \mathbf{y}, \mathbf{D}) &= u_S^*(\tilde{\pi}_{ABC}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}), \tilde{\pi}_{ABC}(\mathbf{y}|\mathbf{D})), \\
&= \log \tilde{\pi}_{ABC}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}) - \log \tilde{\pi}_{ABC}(\mathbf{y}|\mathbf{D}),
\end{aligned}
$$

where

$$\tilde{\pi}_{ABC}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}) = \frac{1}{C_1} \sum_{j=1}^{C_1} I(\kappa(\tilde{\mathbf{y}}_j^{\boldsymbol{\theta}}, \mathbf{y}) \leq \epsilon), \tag{6}$$

$$\tilde{\pi}_{ABC}(\mathbf{y}|\mathbf{D}) = \frac{1}{C_2} \sum_{j=1}^{C_2} I(\kappa(\tilde{\mathbf{y}}_j, \mathbf{y}) \leq \epsilon), \tag{7}$$

respectively. In (6) and (7), respectively, $\{\tilde{\mathbf{y}}_j^{\boldsymbol{\theta}}\}_{j=1}^{C_1}$ is a sample of size C_1 from the distribution of $\mathbf{y}$ conditional on $\boldsymbol{\theta}$ and $\{\tilde{\mathbf{y}}_j\}_{j=1}^{C_2}$ is a sample of size C_2 from the marginal distribution of $\mathbf{y}$.

The ABC methodology needs the successful simultaneous specification of a discrepancy function and tolerance. In an inferential setting with fixed observations $\mathbf{y}$ and design, $\mathbf{D}$, it is possible to tailor these choices. However for Bayesian design, $\mathbf{y}$ is unknown meaning the choice of discrepancy function and tolerance need to be suitable for all observations under the marginal distribution of $\mathbf{y}$ and any design $\mathbf{D} \in \Delta$. This means applying ABC techniques to find Bayesian designs for anything other than small n is difficult (e.g Dehideniya et al., 2018) and therefore Bayesian design for intractable likelihood models using ABC has been limited to design spaces of small dimensionality.

As mentioned in Section 1.1, an alternative methodology for inference under an intractable likelihood is auxiliary modelling (also known as indirect inference). This is a well established methodology for both frequentist (Gourieroux et al., 1993; Heggland and Frigessi, 2004) and Bayesian (Drovandi et al., 2011, 2015) inference. An auxiliary distribution $\mathcal{F}_X (\boldsymbol{\theta}, \mathbf{d}_k)$ is used to approximate the distribution, $\mathcal{F}$, of y_k given in (1). Suppose the pdf/pmf of $\mathcal{F}_X$ is denoted by $f_X(y|\boldsymbol{\theta}, \mathbf{d})$, then the auxiliary likelihood is given by

$$\pi_X(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}) = \prod_{k=1}^{n} f_X(y_k|\boldsymbol{\theta}, \mathbf{d}_k), \tag{8}$$

which can be used as an approximation to the assumed likelihood $\pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D})$. A naive approach would then be to replace evaluation of the assumed likelihood $\pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D})$ by evaluation of the auxiliary likelihood $\pi_X(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D})$, in all subsequent computations, including those in the nested Monte Carlo approximation to the expected utility. This approach has previously been used to find Bayesian designs under an intractable likelihood by Ryan et al. (2016a). However, the designs were found using sampling-based approaches (see Section 1.4) and so were restricted to design spaces of unrealistically small dimensionality. There also exists a further subtle disadvantage of using nested Monte Carlo to approximate the expected utility under the auxiliary modelling approach, which relates to the complexity of the auxiliary distribution. Consider the inner Monte Carlo approximation to the marginal likelihood given by (5). For all pairs of $i = 1, \ldots, B$ and $j = 1, \ldots, C$, we need to evaluate the auxiliary likelihood

$$\pi_X(\mathbf{y}_i|\tilde{\boldsymbol{\theta}}_j, \mathbf{D}) = \exp\left(\sum_{k=1}^{n} \log f_X(y_{ik}|\tilde{\boldsymbol{\theta}}_j, \mathbf{d}_k)\right), \tag{9}$$

where y_{ik} is the kth element of $\mathbf{y}_i$. In general, the exponent in the right hand side of (9) can be decomposed as

$$\sum_{k=1}^{n} \log f_X(y_{ik}|\tilde{\boldsymbol{\theta}}_j, \mathbf{d}_k) = \sum_{k=1}^{n} \alpha(y_{ik}, \mathbf{d}_k) + \sum_{k=1}^{n} \beta(\tilde{\boldsymbol{\theta}}_j, \mathbf{d}_k) + \sum_{k=1}^{n} \gamma(y_{ik}, \tilde{\boldsymbol{\theta}}_j, \mathbf{d}_k), \tag{10}$$

for functions α, β and γ whose form depend on the exact form of $f_X(y|\boldsymbol{\theta}, \mathbf{D})$. Therefore, to evaluate the nested Monte Carlo approximation to the expected SIG utility, α and β are evaluated $B \times n$ and $C \times n$ times each, respectively. However, γ is evaluated $B \times C \times n$ times, which can result in a high computational burden. In cases where the nested Monte Carlo approximation has been applied previously for tractable likelihood models (e.g Huan and Marzouk, 2013; Overstall and Woods, 2017), the actual (not auxiliary) response distribution is from the exponential family of distributions and γ can be decomposed as follows

$$\gamma(y, \boldsymbol{\theta}, \mathbf{d}) = \gamma_y(y, \mathbf{d})\gamma_\theta(\boldsymbol{\theta}, \mathbf{d}) \tag{11}$$

which significantly reduces the computational burden of evaluation. It transpires that γ_y and γ_θ need only be evaluated $B \times n$ and $C \times n$ times each, respectively. However, as we demonstrate in Section 3, the auxiliary distribution typically needs to possess characteristics which are not found in exponential family distributions. For example, in the case of the LV model, where the response given by the number of prey is a count, the negative binomial distribution provides a far more adequate auxiliary distribution (see Section 3.2.2) than the Poisson, the latter being an exponential family distribution. In these cases, the decomposition given by (11) will typically not hold. This apparently simple complication significantly increases the computational burden of evaluating the nested Monte Carlo approximation to the expected utility (see Section 3 for comparisons).

1.6 Organisation of the paper

In Section 2, we introduce a novel approach to finding Bayesian designs for intractable likelihood models using auxiliary modelling. The central platform of this approach is the development of a new method for approximating the expected likelihood-based utility which does not require nested Monte Carlo. This is based on an auxiliary model in combination with a copula to account for dependency in the elements of $\mathbf{y}$, induced by integrating out $\boldsymbol{\theta}$. In addition, we propose an automatic approach to creating and assessing auxiliary models based on ideas from computer experiments. We apply the latest Bayesian design optimisation methods and demonstrate our approach on examples from the biological sciences, where the experimental aims include parameter estimation and model comparison, and where the design spaces are of realistic dimensionality.

2 Methodology

2.1 Outline

To make the terminology in this section clearer, the auxiliary model, $\mathcal{F}_X(\boldsymbol{\theta}, \mathbf{d})$, introduced in Section 1.5, is termed the *conditional auxiliary model*, since it is conditional on $\boldsymbol{\theta}$. Once the conditional auxiliary model for the distribution $\mathcal{F}(\boldsymbol{\theta}, \mathbf{d})$ has been specified it is straightforward to obtain the auxiliary likelihood approximation to the likelihood using (8).

Now consider approximating the marginal likelihood. Let $\mathcal{G}(\mathbf{d})$ denote the marginal distribution of an element, y (with design variables $\mathbf{d}$), of $\mathbf{y}$ having marginalised over $\boldsymbol{\theta}$ and all elements of $\mathbf{y}$ apart from y. Let $g(y|\mathbf{d})$ and $G(y|\mathbf{d})$ be the pdf/pmf and distribution function of $\mathcal{G}(\mathbf{d})$, respectively. If the elements of $\mathbf{y}$ are continuous then, by Sklar's theorem (e.g. Nelson, 1998, Section 2.3), the marginal likelihood is uniquely given by

$$\pi(\mathbf{y}|\mathbf{D}) = c\left(G(y_1|\mathbf{d}_1), \ldots, G(y_n|\mathbf{d}_n)|\mathbf{D}\right) \times \prod_{k=1}^{n} g(y_k|\mathbf{d}_k), \tag{12}$$

where c is the pdf of the copula $\mathcal{C}$ of the marginal distribution of $\mathbf{y}$. Suppose that $\mathbf{u} = (G(y_1|\mathbf{d}_1), \ldots, G(y_n|\mathbf{d}_n))$, then the marginal distribution of each element of $\mathbf{u}$ is $\mathrm{U}[0,1]$ and the copula is the joint distribution of $\mathbf{u}$. Essentially a continuous multivariate probability distribution can be decomposed into the marginal distributions of each element and the copula which controls the dependency structure. Suppose we are able to find a suitable auxiliary model (termed the *marginal auxiliary model*), denoted by $\mathcal{G}_X(\mathbf{d})$, for $\mathcal{G}(\mathbf{d})$, with pdf $g_X(y|\mathbf{d})$ and distribution function $G_X(y|\mathbf{d})$, and an *auxiliary copula*, $\mathcal{C}_X$, with pdf $c_X(\mathbf{u}|\mathbf{D})$, then the marginal likelihood can be approximated using

$$\pi_X(\mathbf{y}|\mathbf{D}) = c_X\left(G_X(y_1|\mathbf{d}_1), \ldots, G_X(y_n|\mathbf{d}_n)|\mathbf{D}\right) \times \prod_{k=1}^{n} g_X(y_k|\mathbf{d}_k). \tag{13}$$

The decomposition given by (12) is only unique for continuous $\mathbf{y}$. However, this does not preclude its use for discrete $\mathbf{y}$ (see, e.g., Panagiotelis et al., 2012).

We thus need three auxiliary models: (a) the conditional auxiliary model, $\mathcal{F}_X$, for $\mathcal{F}$; (b) the marginal auxiliary model, $\mathcal{G}_X$, for $\mathcal{G}$ and (c) the auxiliary copula, $\mathcal{C}_X$, for $\mathcal{C}$. The combination of marginal auxiliary model and auxiliary copula create a, what we term, *coupled auxiliary model* for the marginal distribution of $\mathbf{y}$.

In the case of the conditional and marginal auxiliary models, we choose a family of distributions, $\mathcal{H}_X(\boldsymbol{\phi})$, indexed by a $v \times 1$ vector of *auxiliary parameters* $\boldsymbol{\phi}$. Let $h_X(y|\boldsymbol{\phi})$ and $H_X(y|\boldsymbol{\phi})$ be the pdf/pmf and distribution function, respectively, of $\mathcal{H}_X(\boldsymbol{\phi})$. Similar to all statistical modelling the choice of underlying family of distributions is subjective and the usual considerations of support and over-dispersion should guide the decision. Moreover, in Section 2.3 methods are described to assess the adequacy of auxiliary models which serve a similar purpose to diagnostic methods used in statistical modelling. In the case of the conditional auxiliary model, we allow $\boldsymbol{\phi}$ to be a function of $\boldsymbol{\theta}$ and $\mathbf{d}$, i.e. $\boldsymbol{\phi}_f(\boldsymbol{\theta}, \mathbf{d})$ and in the case of the marginal auxiliary model, a function of just $\mathbf{d}$, i.e. $\boldsymbol{\phi}_g(\mathbf{d})$. We find an estimate, $\hat{\boldsymbol{\phi}}$, for the function $\boldsymbol{\phi}$ using training

data generated from the assumed model in combination with an automatic approach based on computer experiments (see Section 2.2), and then

$$
\begin{aligned}
f_X(y|\boldsymbol{\theta}, \mathbf{d}) &= h_X(y|\hat{\boldsymbol{\phi}}_f(\boldsymbol{\theta}, \mathbf{d})), \\
g_X(y|\mathbf{d}) &= h_X(y|\hat{\boldsymbol{\phi}}_g(\mathbf{d})), \\
G_X(y|\mathbf{d}) &= H_X(y|\hat{\boldsymbol{\phi}}_g(\mathbf{d})).
\end{aligned}
\tag{14}
$$

For computational efficiency, we propose that the estimation of $\boldsymbol{\phi}$ (and therefore the specification of the conditional and marginal auxiliary models) be completed "off-line", i.e. prior to the implementation of the optimisation method chosen to maximise the Monte Carlo approximation, $\tilde{U}(\mathbf{D})$, to the expected utility.

Now consider specifying the copula. Similar to above, we choose a family for the copula indexed by an $r \times 1$ vector of *copula parameters*, $\boldsymbol{\gamma}$. In contrast to the conditional and marginal auxiliary models, the choice of copula family will be less intuitive. However, for all examples in this paper, the normal copula (e.g. Arbenz, 2013) sufficed to produce an adequate coupled auxiliary model. Unlike the conditional and marginal auxiliary models, we propose that the specification of the auxiliary copula be made "on-line", i.e. at evaluation of $\tilde{U}(\mathbf{D})$ within the chosen optimisation method. The reasoning for this difference is as follows. In the case of the conditional and marginal auxiliary models, the dimensionality of the arguments of the function $\boldsymbol{\phi}$ are $p + w$ and w, respectively, i.e. relatively small. However, were $\boldsymbol{\gamma}$ a function of $\mathbf{D}$, then the dimensionality of this argument is $q = nw$, i.e. relatively large, for which it may not be possible to estimate $\boldsymbol{\gamma}$ reliably for all $\mathbf{D} \in \Delta$. At each evaluation of $\tilde{U}(\mathbf{D})$, since $\mathbf{D}$ is fixed, $\boldsymbol{\gamma}$ is independent of $\mathbf{D}$ and its value estimated using a copula training sample generated from the assumed model. Therefore, we write the copula as $\mathcal{C}_X(\boldsymbol{\gamma})$ with pdf $c_X(\mathbf{u}|\boldsymbol{\gamma})$, i.e. independent of $\mathbf{D}$. For example, the pdf for the normal copula is given by

$$
c_X(\mathbf{u}|\boldsymbol{\gamma}) = \frac{1}{|\mathbf{R}(\boldsymbol{\gamma})|}^{\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{v}^T\left(\mathbf{R}(\boldsymbol{\gamma})^{-1} - \mathbf{I}_n\right)\mathbf{v}\right),
$$

where $\mathbf{v}$ is an $n \times 1$ vector with kth element $v_k = \Phi^{-1}(u_k)$, u_k is the kth element of $\mathbf{u}$, and $\mathbf{R}(\boldsymbol{\gamma})$ is an $n \times n$ correlation matrix with $r = \frac{1}{2}n(n-1)$ unique elements given by the elements of $\boldsymbol{\gamma}$.

We now summarise the steps required to approximate the expected utility given a design $\mathbf{D}$, a Monte Carlo sample size B and a copula training sample size L.

1. Generate sample, $\{\boldsymbol{\theta}_i\}_{i=1}^B$ from the prior distribution of $\boldsymbol{\theta}$. For $i = 1, \ldots, B$ and $k = 1, \ldots, n$, generate

$$
y_{ik} \sim \mathcal{F}(\boldsymbol{\theta}_i, \mathbf{d}_k),
$$

   and let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})$. Now $\{\mathbf{y}_i, \boldsymbol{\theta}_i\}_{i=1}^B$ is the Monte Carlo sample from the joint distribution of $\mathbf{y}$ and $\boldsymbol{\theta}$.

2. Generate sample, $\{\bar{\boldsymbol{\theta}}_i\}_{i=1}^L$ from the prior distribution of $\boldsymbol{\theta}$. For $l = 1, \ldots, L$ and $k = 1, \ldots, n$, generate

$$
\bar{y}_{lk} \sim \mathcal{F}(\bar{\boldsymbol{\theta}}_l, \mathbf{d}_k),
$$

   and let $\bar{\mathbf{y}}_l = (\bar{y}_{l1}, \ldots, \bar{y}_{ln})$. Now $\{\bar{\mathbf{y}}\}_{l=1}^L$ is the copula training sample from the marginal distribution of $\mathbf{y}$.

3. Calculate the maximum likelihood estimates, $\hat{\boldsymbol{\gamma}}$, of $\boldsymbol{\gamma}$ where

$$
\hat{\boldsymbol{\gamma}} = \arg\max_{\gamma} \prod_{l=1}^L c_X\left(G_X(\bar{y}_{l1}|\mathbf{d}_1), \ldots, G_X(\bar{y}_{ln}|\mathbf{d}_n)|\boldsymbol{\gamma}\right).
$$

   This maximisation will need to be computed numerically since closed form maximum likelihood estimates typically do not exist for copula parameters.

8

4. For $i = 1, \ldots, B$, calculate the following approximations to the likelihood and marginal likelihood

$$\pi_X(\mathbf{y}_i|\boldsymbol{\theta}_i, \mathbf{D}) = \prod_{i=1}^{n} f_X(y_{ik}|\boldsymbol{\theta}_i, \mathbf{d}_k),$$

$$\pi_X(\mathbf{y}_i|\mathbf{D}) = c_X \left(G_X(y_{i1}|\mathbf{d}_1), \ldots, G_X(y_{in}|\mathbf{d}_n)|\hat{\boldsymbol{\gamma}} \right) \prod_{i=1}^{n} g_X(y_{ik}|\mathbf{d}_k),$$

where $f_X(y|\boldsymbol{\theta}, \mathbf{d})$, $g_X(y|\mathbf{d})$ and $G_X(y|\mathbf{d})$ are given by (14).

5. For $i = 1, \ldots, B$, approximate the likelihood based utility using

$$u_X(\boldsymbol{\theta}_i, \mathbf{y}_i, \mathbf{D}) = u^*(\pi_X(\mathbf{y}_i|\boldsymbol{\theta}_i, \mathbf{D}), \pi_X(\mathbf{y}_i|\mathbf{D})).$$

The resulting approximation to the expected utility given by

$$\tilde{U}(\mathbf{D}) = \frac{1}{B} \sum_{i=1}^{B} u_X(\mathbf{y}_i, \boldsymbol{\theta}_i, \mathbf{D}),$$

is termed the *auxiliary Monte Carlo approximation* to the expected utility.

2.2 Conditional and marginal auxiliary models

To estimate the functions, ϕ_f and ϕ_g, we propose an automatic approach originating from the field of computer experiments (see, e.g., Dean et al. 2015, Section V). In this area, the goal is to approximate an unknown function (which is usually computationally expensive). To do this, the function is evaluated a "small" number of times at a specified meta-design of arguments and a statistical model (known as an emulator) is fitted to the output. The emulator provides a prediction of the unknown function for any argument. We use the multivariate Gaussian process (MGP; Conti and O'Hagan 2010) model as the emulator. This is a multivariate generalisation of the Gaussian process model which is a commonly employed emulator in computer experiments.

We begin by generating a training sample $\left\{ \mathbf{d}^{(i)} \right\}_{i=1}^{M}$ of size M from $\mathcal{D}$. We employ the usual design used for computer experiments, i.e. a space-filling Latin hypercube design (e.g. Santner et al., 2003, Chapter 5). We then generate a sample $\left\{ \boldsymbol{\theta}^{(i)} \right\}_{i=1}^{M}$ of size M from the prior distribution of $\boldsymbol{\theta}$. For $j = 1, \ldots, N$, we further generate samples $\left\{ \boldsymbol{\theta}^{(ij)} \right\}_{i=1}^{M}$ from the same prior distribution. Finally, for $i = 1, \ldots, M$ and $j = 1, \ldots, N$, independently, generate

$$y_f^{(ij)} \sim \mathcal{F}\left(\boldsymbol{\theta}^{(i)}, \mathbf{d}^{(i)} \right)$$
$$y_g^{(ij)} \sim \mathcal{F}\left(\boldsymbol{\theta}^{(ij)}, \mathbf{d}^{(i)} \right).$$

We have generated M samples of size N from two distributions. In the case of the $y_f^{(ij)}$'s, the sample is generated from $\mathcal{F}\left(\boldsymbol{\theta}^{(i)}, \mathbf{d}^{(i)} \right)$. In the case of the $y_g^{(ij)}$'s, the sample is generated from $\mathcal{G}\left(\mathbf{d}^{(i)} \right)$. Under each of these M samples we compute the maximum likelihood estimates of $\boldsymbol{\phi}$ under $\mathcal{H}_X(\boldsymbol{\phi})$, i.e. let

$$\hat{\boldsymbol{\phi}}_f^{(i)} = \arg\max_{\boldsymbol{\phi}} \prod_{j=1}^{N} h_X(y_f^{(ij)}|\boldsymbol{\phi}),$$

$$\hat{\boldsymbol{\phi}}_g^{(i)} = \arg\max_{\boldsymbol{\phi}} \prod_{j=1}^{N} h_X(y_g^{(ij)}|\boldsymbol{\phi}),$$

9

for $i = 1, \ldots, M$. Typically, the maximum likelihood estimates are not available in closed form so numerical methods will be used. We now describe how to use $\left\{ \hat{\phi}_f^{(i)} \right\}_{i=1}^M$ and $\left\{ \hat{\phi}_g^{(i)} \right\}_{i=1}^M$ to estimate the functions ϕ_f and ϕ_g, respectively.

Let $\mathbf{Z}$ be the $v \times M$ matrix where the ith column (for $i = 1, \ldots, M$) is given by $\mathbf{z}_i = \lambda(\hat{\phi}^{(i)})$ where $\hat{\phi}^{(i)}$ is either $\hat{\phi}_f^{(i)}$ or $\hat{\phi}_g^{(i)}$ and λ is a monotonic and differentiable link function applied element-wise to $\hat{\phi}^{(i)}$. The link function is applied so that the elements of $\mathbf{Z}$ are in $\mathbb{R}$, e.g. a log link if the auxiliary parameters are positive. Under the MGP, we assume that

$$\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{A} \sim \mathrm{MN}\left(\boldsymbol{\beta}\mathbf{1}_M, \mathbf{A}, \boldsymbol{\Sigma}\right) \tag{15}$$

where $\mathrm{MN}\left(\boldsymbol{\beta}\mathbf{1}_M, \mathbf{A}, \boldsymbol{\Sigma}\right)$ denotes the matrix-normal distribution with $v \times M$ mean matrix $\boldsymbol{\beta}\mathbf{1}_M$, $v \times v$ unstructured row covariance matrix $\boldsymbol{\Sigma}$ and $M \times M$ column correlation matrix $\mathbf{A}$. In (15), $\mathbf{1}_M$ is an $1 \times M$ matrix of ones, and $\boldsymbol{\beta}$ is an $v \times 1$ matrix of coefficients. The ijth element of $\mathbf{A}$ is given by

$$A_{ij} = \kappa\left(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\rho}\right) + \eta I(i = j),$$

where $\mathbf{x}_i = (\boldsymbol{\theta}^{(i)}, \mathbf{d}^{(i)})$ (in the case of the conditional auxiliary model) or $\mathbf{x}_i = \mathbf{d}^{(i)}$ (in the case of the marginal auxiliary model), $\kappa(\cdot, \cdot; \boldsymbol{\rho})$ is a valid correlation function depending on parameters $\boldsymbol{\rho}$, and $\eta > 0$ is referred to as a nugget. Note that $\boldsymbol{\rho}$ is an $s \times 1$ vector where $s = p + w$ (in the case of the conditional auxiliary model) or $s = w$ (in the case of the marginal auxiliary model).

Let $\mathbf{x} = (\boldsymbol{\theta}, \mathbf{d})$ (in the case of the conditional auxiliary model) or $\mathbf{x} = \mathbf{d}$ (in the case of the marginal auxiliary model). Suppose we wish to predict the value of $\mathbf{z} = \lambda(\boldsymbol{\phi}(\mathbf{x}))$ for any value of $\mathbf{x}$. Under (15), the predictive distribution for $\mathbf{z}$ is given by

$$\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\rho}, \eta \sim \mathrm{N}\left(\boldsymbol{\beta} + \left(\mathbf{Z} - \boldsymbol{\beta}\mathbf{1}_M\right)\mathbf{A}^{-1}\mathbf{a}, \left(1 + \eta - \mathbf{a}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{a}\right)\boldsymbol{\Sigma}\right), \tag{16}$$

where $\mathbf{a}$ is an $M \times 1$ vector with ith element

$$a_i = \kappa\left(\mathbf{x}_i, \mathbf{x}; \boldsymbol{\rho}\right).$$

For simplicity, we specify that the function $\hat{\boldsymbol{\phi}}(\mathbf{x})$ is given by the inverse link of the predictive mean, i.e. the mean of (16). This depends on the parameters $\boldsymbol{\beta}$, $\boldsymbol{\rho}$ and η. We replace these parameters by their maximum likelihood estimates (MLEs) where the likelihood is given by (15). Thus

$$\hat{\boldsymbol{\phi}}(\mathbf{x}) = \lambda^{-1}\left(\hat{\boldsymbol{\beta}} + \left(\mathbf{Z} - \hat{\boldsymbol{\beta}}\mathbf{1}_M\right)\hat{\mathbf{A}}^{-1}\hat{\mathbf{a}}\right). \tag{17}$$

In (17), $\hat{\mathbf{a}}$ and $\hat{\mathbf{A}}$ are $\mathbf{a}$ and $\mathbf{A}$, respectively, with $\boldsymbol{\rho}$ and η replaced by their maximum likelihood estimates, $\hat{\boldsymbol{\rho}}$ and $\hat{\eta}$, respectively.

2.3 Assessing adequacy of auxiliary models

Before applying the auxiliary models described in Section 2.2 to approximate the expected utility, their adequacy should be assessed for plausibility, i.e. do they provide a reasonable approximation to the assumed model. The approach proposed is based on posterior predictive assessments (see, for example, Gelman et al. 2014, Chapter 6). Here M_0 test samples are generated from the assumed and auxiliary models, and sample statistics from each compared. We propose to separately assess a) the conditional and marginal auxiliary models and; b) the coupled auxiliary model.

2.3.1 Assessing the conditional and marginal auxiliary models

We generate M_0 samples of size N from the assumed and auxiliary models using the following two steps.

1. Generate samples, $\left\{\boldsymbol{\theta}^{(i)}\right\}_{i=1}^{M_0}$ and $\left\{\mathbf{d}^{(i)}\right\}_{i=1}^{M_0}$, of size M_0 from the prior distribution of $\boldsymbol{\theta}$ and uniformly over $\mathcal{D}$, respectively.

2. For $i = 1, \ldots, M_0$ and $j = 1, \ldots, N$ generate

$$
\begin{aligned}
y_f^{(ij)} &\sim \mathcal{F}\left(\boldsymbol{\theta}^{(i)}, \mathbf{d}^{(i)}\right) & y_{fX}^{(ij)} &\sim \mathcal{H}_X\left(\hat{\boldsymbol{\phi}}_f\left(\boldsymbol{\theta}^{(i)}, \mathbf{d}^{(i)}\right)\right), \\
y_g^{(ij)} &\sim \mathcal{G}\left(\mathbf{d}^{(i)}\right) & y_{gX}^{(ij)} &\sim \mathcal{H}_X\left(\hat{\boldsymbol{\phi}}_g\left(\mathbf{d}^{(i)}\right)\right).
\end{aligned}
$$

Let

$$
\begin{aligned}
\mathbf{y}_f^{(i)} &= \left(y_f^{(i1)}, \ldots, y_f^{(iN)}\right) & \mathbf{y}_{fX}^{(i)} &= \left(y_{fX}^{(i1)}, \ldots, y_{fX}^{(iN)}\right), \\
\mathbf{y}_g^{(i)} &= \left(y_g^{(i1)}, \ldots, y_g^{(iN)}\right) & \mathbf{y}_{gX}^{(i)} &= \left(y_{gX}^{(i1)}, \ldots, y_{gX}^{(iN)}\right).
\end{aligned}
$$

We propose two diagnostics. First, plot sample statistics of the $\mathbf{y}_f^{(i)}$'s ($\mathbf{y}_g^{(i)}$'s) against $\mathbf{y}_{fX}^{(i)}$'s ($\mathbf{y}_{gX}^{(i)}$'s), where suggested sample statistics are mean, variance, median, etc. If the conditional and marginal auxiliary models are adequate then the points should approximately lie on a straight line through the origin with unit slope. Second, single number summaries of conditional and marginal auxiliary model adequacy are given by the Bayesian posterior predictive p-values (Gelman et al., 2014, page 146) given by

$$
\text{p-value}_f = \frac{1}{M_0}\sum_{i=1}^{M_0} I\left(\sum_{j=1}^{N} \log h_X\left(y_f^{(ij)}|\hat{\boldsymbol{\phi}}_f(\boldsymbol{\theta}^{(i)}, \mathbf{d}^{(i)})\right) < \sum_{j=1}^{N} \log h_X\left(y_{fX}^{(ij)}|\hat{\boldsymbol{\phi}}_f(\boldsymbol{\theta}^{(i)}, \mathbf{d}^{(i)})\right)\right),
$$

$$
\text{p-value}_g = \frac{1}{M_0}\sum_{i=1}^{M_0} I\left(\sum_{j=1}^{N} \log h_X\left(y_g^{(ij)}|\hat{\boldsymbol{\phi}}_g(\mathbf{d}^{(i)})\right) < \sum_{j=1}^{N} \log h_X\left(y_{gX}^{(ij)}|\hat{\boldsymbol{\phi}}_g(\mathbf{d}^{(i)})\right)\right),
$$

respectively. Posterior predictive p-values close to zero or one indicate that the auxiliary model is inadequate.

2.3.2   Assessing the coupled auxiliary model

To assess the adequacy of the coupled auxiliary model, we generate M_0 samples of size n from both the assumed model and the coupled auxiliary model. To do this does require estimation of the copula parameters for every sample generated.

1. Generate a sample $\left\{\boldsymbol{\theta}^{(i)}\right\}_{i=1}^{M_0}$ of size M_0 from the prior distribution of $\boldsymbol{\theta}$. Generate a sample $\left\{\mathbf{D}^{(i)}\right\}_{i=1}^{M_0}$ of size M_0 uniformly from Δ where $\mathbf{D}^{(i)} = \left(\mathbf{d}^{(i1)}, \ldots, \mathbf{d}^{(in)}\right)$ with $\mathbf{d}^{(ik)}$ uniformly generated from $\mathcal{D}$, for $k = 1, \ldots, n$.

2. For $i = 1, \ldots, M_0$ and $k = 1, \ldots, n$, generate

$$
y^{(ik)} \sim \mathcal{F}(\boldsymbol{\theta}^{(i)}, \mathbf{d}^{(ij)}),
$$

and let $\mathbf{y}^{(i)} = \left(y^{(i1)}, \ldots, y^{(in)}\right)$.

3. For $i = 1, \ldots, M_0$, complete the following steps

   (a) Generate $\bar{\boldsymbol{\theta}}^{(i)}$ from the prior distribution of $\boldsymbol{\theta}$.

   (b) For $l = 1, \ldots, L$ and $k = 1, \ldots, n$, generate

$$
\bar{y}^{(lk)} \sim \mathcal{F}\left(\bar{\boldsymbol{\theta}}^{(i)}, \mathbf{d}^{(ik)}\right).
$$

(c) Calculate

$$\hat{\boldsymbol{\gamma}}^{(i)} = \arg\max_{\boldsymbol{\gamma}} \prod_{l=1}^{L} c_X \left(G_X(\bar{y}^{(l1)}|\mathbf{d}^{i1}), \ldots, G_X(\bar{y}^{(ln)}|\mathbf{d}^{in})|\boldsymbol{\gamma} \right).$$

4. For $i = 1, \ldots, M_0$ generate $\mathbf{u}^{(i)}$ from the copula, $\mathcal{C}_X$ with parameters $\hat{\boldsymbol{\gamma}}^{(i)}$. In the case of the normal copula this is achieved by generating $\mathbf{v}^{(i)} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{R}(\hat{\boldsymbol{\gamma}}^{(i)})\right)$ and setting

$$\mathbf{u}^{(i)} = \left(\Phi(v^{(i1)}), \ldots, \Phi(v^{(in)}) \right),$$

where $v^{(ik)}$ is the kth element of $\mathbf{v}^{(i)}$. Now

$$y_X^{(ik)} = G_X^{-1}\left(u^{(ik)}|\mathbf{d}^{(ik)} \right),$$

where G_X^{-1} is the inverse distribution function of $\mathcal{G}_X$ and $u^{(ik)}$ is the kth element of $\mathbf{u}^{(i)}$. Set $\mathbf{y}_X^{(i)} = \left(y_X^{(i1)}, \ldots, y_X^{(in)} \right)$.

To compare the samples generated from both the assumed model and the coupled auxiliary model we use a posterior predictive p-value given by

$$\text{p-value} = \frac{1}{M_0} \sum_{i=1}^{M_0} I \left(\log \pi_X \left(\mathbf{y}^{(i)}|\mathbf{D}^{(i)} \right) > \log \pi_X \left(\mathbf{y}_X^{(i)}|\mathbf{D}^{(i)} \right) \right).$$

Similar to Section 2.3.1, a posterior predictive p-value close to zero or one suggests an inadequate coupled auxiliary model.

3   Examples

We apply the proposed methodology on a series of examples. In Section 3.2, we apply the proposed methodology on two illustrative examples to demonstrate the methodology and to assess its efficacy. In Sections 3.3 and 3.4, we apply the methodology on more challenging examples. However, first, in Section 3.1 we describe some implementation details common to all examples.

3.1   Implementation details

3.1.1   ACE algorithm

We specify the controllable quantities in the ACE algorithm (see Appendix A). We set $B = 1000$ and $B = 20000$ for fitting the GP model and the independent Bayesian hypothesis step, respectively. We also set the GP training sample to be $Q = 20$. These are the default values in the acebayes package. Additionally we found that twenty iterations of the ACE algorithm was sufficient to achieve approximate convergence in all examples. Finally, for each example, we restart the ACE algorithm from twenty different starting designs as proposed by Overstall and Woods (2017). This is to mitigate against convergence to local optima.

3.1.2   Correlation function

For all of the examples in this paper, for the MGP emulator, we use the squared exponential correlation function, i.e.

$$\kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\rho}) = \exp\left(-\sum_{l=1}^{s} \rho_l \left(x_l - x_l' \right)^2 \right),$$

where x_l and x_l' are the lth elements of $\mathbf{x}$ and $\mathbf{x}'$, respectively, for $l = 1, \ldots, s$. This is a commonly-used correlation function for the MGP (e.g. Overstall and Woods, 2016).

3.1.3 Sample sizes

For the training samples, we set $M = 500$ (number of marginal and conditional auxiliary model training samples), $N = 10000$ (size of training sample size) and $L = 500$ (number of copula training samples). These were found to be sufficient in all examples to provide adequate auxiliary models. To assess adequacy, we used $M_0 = 100$ test samples for the illustrative examples in Section 3.2 and $M_0 = 2000$ for the the more challenging examples in Sections 3.3 and 3.4.

3.2 Illustrative examples

3.2.1 Compartmental model

In this section we apply the proposed methods to find a Bayesian design under the SIG utility for a compartmental model. For this model, the likelihood is available in closed form so the aim of this example is to assess the efficacy of the approach. Compartmental models simulate how materials flow through an organism. The standard design problem is to specify the n sampling times $\mathbf{D} = (t_1, \ldots, t_n)$ (in hours) at which to measure the concentration of a drug in an individual, following the individual being administered the drug at time $t = 0$. The basis of this example comes from Ryan et al. (2014) and Overstall and Woods (2017). The concentration at time t_k is denoted by y_k where it is assumed that

$$y_k \sim \mathrm{N}\left(\mu(\boldsymbol{\theta}; t_k), \nu(\boldsymbol{\theta}; t_k)\right),$$

with $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ being the unknown parameters, and

$$\mu(\boldsymbol{\theta}; t_k) = \frac{400\theta_2}{\theta_3 (\theta_2 - \theta_1)} \left(\exp\left(-\theta_1 t_k\right) - \exp\left(-\theta_2 t_k\right)\right),$$

$$\nu(\boldsymbol{\theta}; t_k) = 1 + \frac{\mu(\boldsymbol{\theta}; t_k)^2}{10},$$

and $n = 15$. Independent prior distributions are assumed for the elements of $\boldsymbol{\theta}$, where, on the log scale, the common variance is 0.05 and the expectations are $\log(0.1)$, $\log(1)$ and $\log(20)$, respectively. Additionally, a constraint is imposed on the design whereby sampling times must be at least 15 minutes apart. Overstall and Woods (2017) describe how such constraints can be incorporated into the ACE algorithm in a straightforward manner.

For the family of distributions $\mathcal{H}_X$, we use the skewed t-distribution (see, for example, Azzalini and Capatanio 2003). Non-skewed normal and t-distributions were investigated but it was found that the resulting auxiliary models were not adequate using the methodology of Section 2.3. The skewed t-distribution is indexed by $v = 4$ auxiliary parameters, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_4)$ controlling location, scale, skew and kurtosis. The parameters controlling scale and kurtosis are positive so the λ link function is chosen to be the log function for these elements of $\boldsymbol{\phi}$. The posterior predictive p-values associated with the conditional and marginal auxiliary models are p-value$_f = 0.62$ and p-value$_g = 0.42$. The posterior predictive p-value associated with the coupled auxiliary model is 0.47. Figure 1 shows plots of sample statistics (mean and variance) of the $\mathbf{y}_f^{(i)}$'s (the $\mathbf{y}_g^{(i)}$'s) against the $\mathbf{y}_{fX}^{(i)}$'s (the $\mathbf{y}_{gX}^{(i)}$'s). The p-values and Figure 1 show that the auxiliary models appear adequate.

We find Bayesian designs under the SIG utility using ACE under four different approaches. In the first, we use the auxiliary Monte Carlo approximation to the expected utility, as described in Section 2, where we approximate both the likelihood and marginal likelihood using the auxiliary models. In the second approach, we only approximate the marginal likelihood, using the coupled auxiliary model, since the likelihood for the compartmental model is available in closed form. We then repeat these two approaches under an independence copula, i.e. $c_X(\mathbf{u}) = 1$, to investigate the benefit (or not) of using a copula to model the dependence between the elements of $\mathbf{y}$ in the marginal model. We compare these four designs against a) the SIG design found by using nested Monte Carlo (under the exact likelihood) to approximate the expected utility; and b) the design given by equally-spaced sampling times. The former was found by

Figure 1: Plots of sample mean of (a) $\mathbf{y}_f^{(i)}$ against $\mathbf{y}_{fX}^{(i)}$ and (b) $\mathbf{y}_g^{(i)}$ against $\mathbf{y}_{gX}^{(i)}$, and plots of sample variances of (c) $\mathbf{y}_f^{(i)}$ against $\mathbf{y}_{fX}^{(i)}$ and (d) $\mathbf{y}_g^{(i)}$ against $\mathbf{y}_{gX}^{(i)}$ for the auxiliary models found for the compartmental model in Section 3.2.1. In each plot a straight line of unit slope through the origin as been included as a reference.
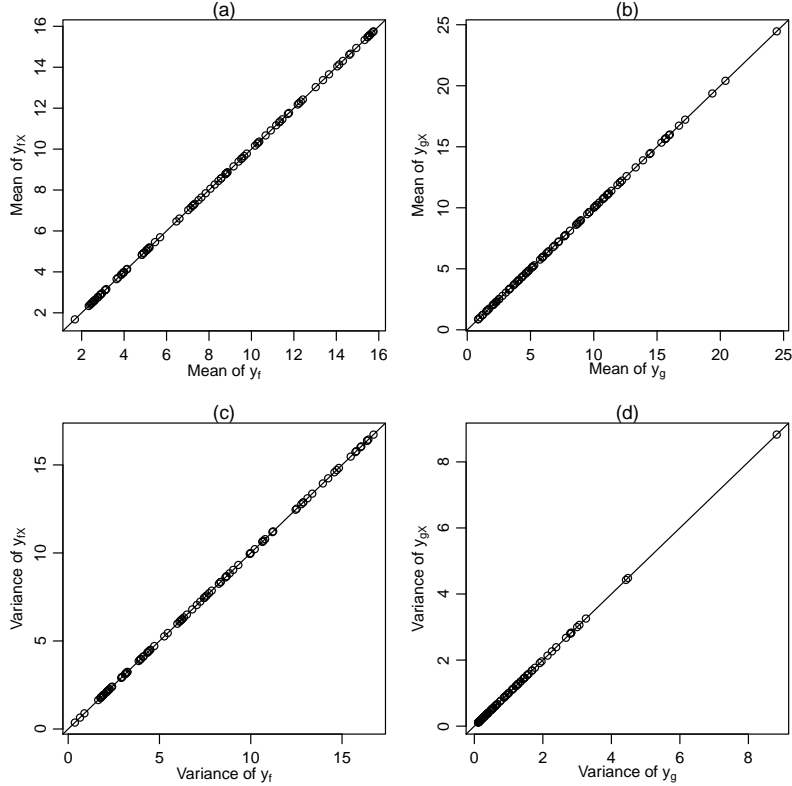

Table 1: Mean (standard error) nested Monte Carlo approximation (under the exact likelihood) to the expected SIG utility for the compartmental model under the six different designs under comparison for the compartmental model in Section 3.2.1.

Design	Mean nested Monte Carlo approximation
Auxiliary Monte Carlo	4.14 (0.003)
Auxiliary Monte Carlo (exact likelihood)	4.41 (0.003)
Auxiliary Monte Carlo (independence copula)	2.93 (0.002)
Auxiliary Monte Carlo (exact likelihood) (independence copula)	2.90 (0.002)
Nested Monte Carlo (exact likelihood)	4.51 (0.003)
Equally-spaced design	3.70 (0.003)

Overstall and Woods (2017) using the ACE algorithm. Table 1 shows the mean and standard error of twenty nested Monte Carlo approximations (under the exact likelihood) to the expected utility under each of the six designs. We can add extra light to these results when we take computing time into account. It took an average of 8.9 hours to find the nested Monte Carlo design, whereas this decreases to 1.7 and 2.7 hours for the auxiliary modelling designs, under the exact and approximate likelihoods, respectively. Clearly, not taking account of the dependency with the auxiliary copula leads to a poor final design. The design found under the methodology proposed in this paper for intractable likelihood models (i.e. first row of Table 1) performs reasonably. Obviously being able to evaluate the exact likelihood (second row of Table 1) improves this design to a point that it has performance close to the design found under nested Monte Carlo. We conclude that the methodology is competitive.

3.2.2   Lotka-Volterra model

We now return to the LV model described in Section 1.3. Similar to the compartmental model, the design problem is to specify the n sampling times $\mathbf{D} = (t_1, \ldots, t_n)$ (in days) at which to observe the number of prey. We assume that the initial prey and predator population sizes are $R(0) = 5$ and $S(0) = 10$, and that the prior distribution for $\boldsymbol{\theta}$ is multivariate log-normal with, on log scale, mean vector of $(\log(1), \log(0.005), \log(0.6))$ and variance matrix

$$
\begin{pmatrix}
5 \times 10^{-2} & 0 & 0 \\
0 & 1.25 \times 10^{-4} & 1 \times 10^{-3} \\
0 & 1 \times 10^{-3} & 1.5 \times 10^{-2}
\end{pmatrix}.
$$

The covariance between θ_2 and θ_3 corresponds to a correlation of 0.75. It seems reasonable that these parameters, controlling the rates of predators reproducing and dying, are strongly correlated. The support for the responses are the non-negative integers. A natural choice for the family of distribution $\mathcal{H}_X$ is the Poisson distribution. This distribution is indexed by $v = 1$ positive auxiliary parameter giving both the mean and variance. The λ link function is chosen to be the log function. Figure 2 shows plots of sample statistics (mean and log variance) of the $\mathbf{y}_f^{(i)}$'s (the $\mathbf{y}_g^{(i)}$'s) against the $\mathbf{y}_{fX}^{(i)}$'s (the $\mathbf{y}_{gX}^{(i)}$'s) as grey plotting characters. We chose to plot the variances on the log scale to make this plot clearer. From Figures 2(a) and (b), there appears to be a good correspondence between the true mean and the mean of the conditional and marginal auxiliary models. However, from Figures 2(c) and (d), we see that both the conditional and marginal auxiliary models are significantly underestimating the variance. To rectify this we consider the negative binomial distribution which is commonly used to model over-dispersion in models for count responses. The negative binomial is indexed by $v = 2$ positive parameters so again the log link function is chosen for both elements of $\boldsymbol{\phi}$. The posterior predictive p-values associated with the conditional and marginal and auxiliary models are p-value$_f = 0.31$ and p-value$_g = 0.36$, respectively. We consider a range of values for the experimental sample size of $n = 1, \ldots, 5$ and posterior predictive p-values range from 0.5 to 0.66 over these different values of n. Figure 2 shows plots of sample statistics (mean and log variance) of the $\mathbf{y}_f^{(i)}$'s (the $\mathbf{y}_g^{(i)}$'s) against the $\mathbf{y}_{fX}^{(i)}$'s (the $\mathbf{y}_{gX}^{(i)}$'s) as black plotting characters. Clearly, Figure 2 shows that the use of the negative binomial distribution has significantly improved the conditional and marginal auxiliary models. Figure 2(d) does show that the marginal auxiliary model does slightly underestimate the variance and this discrepancy increases as the variance increases. Additionally, it appears that for small variances, both the conditional and marginal auxiliary models overestimate the variance. We suspect that this is due to the assumed model exhibiting zero-inflation. We could account for this using a zero-inflated negative binomial model but for the purposes of this exercise we conclude that the auxiliary models are adequate.

For comparison, we find designs under the SIG utility using ACE for two different approaches: auxiliary Monte Carlo and nested Monte Carlo (under the auxiliary likelihood). For comparison, we also consider a design where the sampling times are equally-spaced. Table 2 shows the mean and standard error of twenty nested Monte Carlo approximations (under the auxiliary likelihood) to the expected utility under each of the three designs, for each value of n. There is a small performance advantage for the design found under nested Monte Carlo. However when we consider the average computing times (see Table 3), the designs found under auxiliary Monte Carlo take a fraction of the time to compute.

Figure 2: Plots of sample mean of (a) $\mathbf{y}_f^{(i)}$ against $\mathbf{y}_{fX}^{(i)}$ and (b) $\mathbf{y}_g^{(i)}$ against $\mathbf{y}_{gX}^{(i)}$, and plots of log sample variances of (c) $\mathbf{y}_f^{(i)}$ against $\mathbf{y}_{fX}^{(i)}$ and (d) $\mathbf{y}_g^{(i)}$ against $\mathbf{y}_{gX}^{(i)}$ for both the Poisson (grey) and negative binomial (black) auxiliary models for the Lotka-Volterra model in Section 3.2.2. In each plot a straight line of slope one through the origin as been included as a reference.
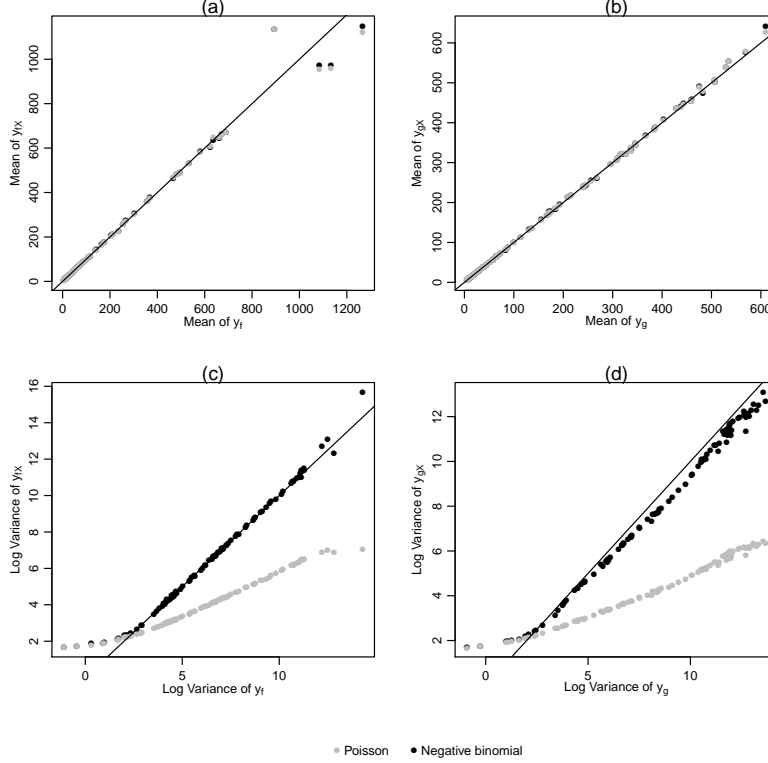

Table 2: Mean (standard error) nested Monte Carlo approximation (under auxiliary likelihood) to the expected SIG utility for the LV model under the three different designs found for the Lotka-Volterra model in Section 3.2.2.

Number of runs, n	Auxiliary Monte Carlo	Nested Monte Carlo	Equally-spaced
1	0.66 (0.001)	0.69 (0.002)	0.46 (0.001)
2	1.05 (0.001)	1.06 (0.001)	0.74 (0.001)
3	1.27 (0.002)	1.28 (0.002)	0.90 (0.001)
4	1.43 (0.002)	1.45 (0.003)	1.06 (0.001)
5	1.54 (0.003)	1.58 (0.003)	1.17 (0.002)

Table 3: Average computing time (in hours) for designs found under a) auxiliary; and b) nested Monte Carlo for the Lotka-Volterra model in Section 3.2.2.

Number of runs, n	Auxiliary Monte Carlo	Nested Monte Carlo
1	0.1	0.6
2	0.2	2.1
3	0.3	3.7
4	0.4	5.6
5	0.5	8.2

We can further investigate the computational expense of the nested Monte Carlo approximation by considering the decomposition given in (10). Under the negative binomial auxiliary model, the functions α, β and γ are given by

$$
\begin{aligned}
\alpha(y, \mathbf{d}) &= -y! \\
\beta(\boldsymbol{\theta}, \mathbf{d}) &= \hat{\phi}_{f2} \log \hat{\phi}_{f2} - \log \Gamma\left(\log \hat{\phi}_{f2}\right) \\
\gamma(y, \boldsymbol{\theta}, \mathbf{d}) &= \log \Gamma(\hat{\phi}_{f2} + y) + y\hat{\phi}_{f1} - (y + \hat{\phi}_{f2}) \log(\hat{\phi}_{f1} + \hat{\phi}_{f2}),
\end{aligned}
$$

where $\left(\hat{\phi}_{f1}, \hat{\phi}_{f2}\right) = \hat{\boldsymbol{\phi}}_f(\boldsymbol{\theta}, \mathbf{d})$ are the $v = 2$ estimated auxiliary parameters under the conditional auxiliary model. Note that the function $\gamma(y, \boldsymbol{\theta}, \mathbf{d})$ cannot be written in the form of (11) due to the appearance of the Gamma function. The same is true even when applying the Stirling approximation to the Gamma function (e.g. Abramowitz and Stegun, 2002, page 257).

3.3   Parasite model

We now consider a parasite model example modified from Drovandi and Pettitt (2013) and Ryan et al. (2016a) which in turn was based on experiments ran by Denham et al. (1977). In the experiment, the kth host cat is injected with $d_{k1} \in [100, 200]$ *Brugia pahangi* larvae at time $t = 0$, for $k = 1, \ldots, n$. After time $d_{k2} \in (30, 300)$ (in days), the kth host cat is sacrificed and the number of mature parasites, y_k, are counted at autopsy. Riley et al. (2003) proposed a Markov process to simulate the population of parasites within the host cat. At time t, let $J(t)$ and $M(t)$ denote the number of juvenile and mature parasites, respectively. Furthermore, let $I(t)$ be a discrete representation of the host immunity. The dynamics of the model are as follows

$$
\begin{aligned}
&\mathrm{P}\left(J(t + \delta t) = J(t) - 1, M(t + \delta t) = M(t) + 1, I(t + \delta t) = I(t)\right) = \theta_1 J(t)\delta t + o(\delta t), \\
&\mathrm{P}\left(J(t + \delta t) = J(t) - 1, M(t + \delta t) = M(t), I(t + \delta t) = I(t)\right) \\
&\hspace{6cm} = \left(\theta_4 + \theta_5 I(t)\right) J(t)\delta t + o(\delta t), \\
&\mathrm{P}\left(J(t + \delta t) = J(t), M(t + \delta t) = M(t) - 1, I(t + \delta t) = I(t)\right) = \theta_2 M(t)\delta t + o(\delta t), \\
&\mathrm{P}\left(J(t + \delta t) = J(t), M(t + \delta t) = M(t), I(t + \delta t) = I(t) + 1\right) = \theta_3 J(t)\delta t + o(\delta t), \\
&\mathrm{P}\left(J(t + \delta t) = J(t), M(t + \delta t) = M(t), I(t + \delta t) = I(t) - 1\right) = \theta_6 I(t)\delta t + o(\delta t),
\end{aligned}
$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)$ are unknown parameters. Note that for the kth run, $J(0) = d_{k1}$, $M(0) = 0$, $I(0) = 0$ and $M(d_{k2}) = y_k$. Both Drovandi and Pettitt (2013) and Ryan et al. (2016a) fixed all parameter values except θ_3 and θ_4 and considered a design space with a maximum dimensionality of four. This was either by setting $n = 4$ and fixing $d_{k1} = 200$ or setting $n = 2$. We consider all elements of $\boldsymbol{\theta}$ to be unknown and number of runs $n = 2, 4, 6, 8, 10, 20, 30, 40$, thus considering a design space with a maximum dimensionality of 80. The prior distributions for $\boldsymbol{\theta}$ follow from the analysis of responses from a previous experiment with $n = 212$ (Denham et al., 1977). Following Drovandi and Pettitt (2013), the prior distribution for θ_3 and θ_4 is given by

$$
\begin{pmatrix} \sqrt{\theta_3} \\ \sqrt{\theta_4} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0.0361 \\ 0.0854 \end{pmatrix}, \begin{pmatrix} 2.03 \times 10^{-5} & -1.07 \times 10^{-4} \\ -1.07 \times 10^{-4} & 1.17 \times 10^{-3} \end{pmatrix} \right).
$$

The remaining parameters are given Gamma prior distributions with mean and variances

$$
\begin{aligned}
\mathrm{E}\left(\theta_1\right) &= 0.04 & \mathrm{var}\left(\theta_1\right) &= 4.00 \times 10^{-4} \\
\mathrm{E}\left(\theta_2\right) &= 0.00147 & \mathrm{var}\left(\theta_2\right) &= 2.56 \times 10^{-7} \\
\mathrm{E}\left(\theta_5\right) &= 1.10 & \mathrm{var}\left(\theta_5\right) &= 0.21 \\
\mathrm{E}\left(\theta_5\right) &= 0.31 & \mathrm{var}\left(\theta_5\right) &= 0.18.
\end{aligned}
$$

The prior means are given by the assumed fixed values of Drovandi and Pettitt (2013), with prior standard deviations given by the corresponding standard errors found by Riley et al. (2003) inferred from 95% confidence intervals.

Figure 3: Plots of sample mean of (a) $\mathbf{y}_f^{(i)}$ against $\mathbf{y}_{fX}^{(i)}$ and (b) $\mathbf{y}_g^{(i)}$ against $\mathbf{y}_{gX}^{(i)}$, and plots of log sample variances of (c) $\mathbf{y}_f^{(i)}$ against $\mathbf{y}_{fX}^{(i)}$ and (d) $\mathbf{y}_g^{(i)}$ against $\mathbf{y}_{gX}^{(i)}$ for the beta binomial auxiliary models for the parasite model in Section 3.3. In each plot a straight line of slope one through the origin as been included as a reference.
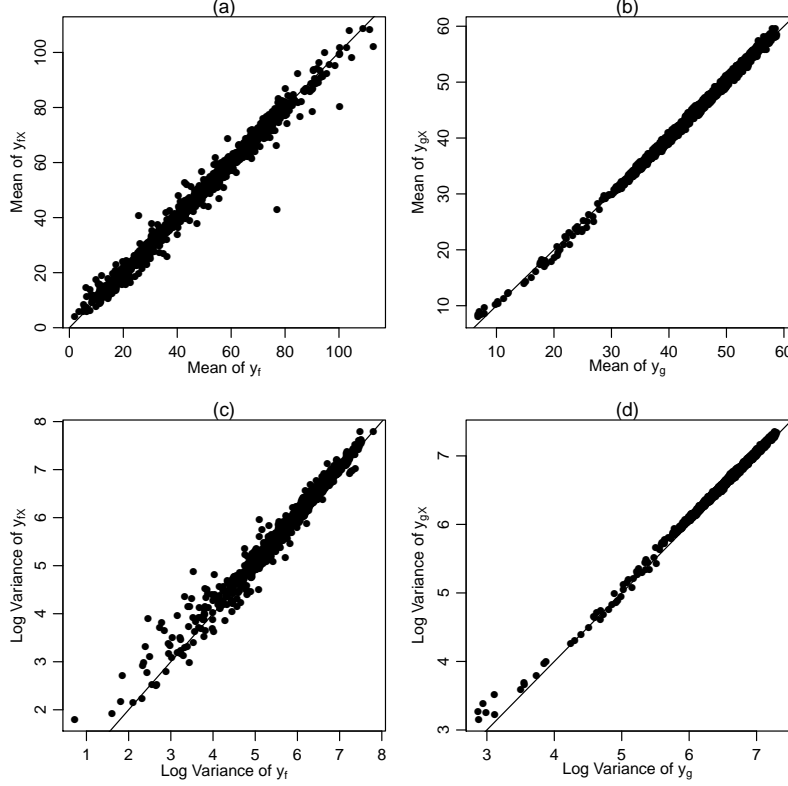
A mature parasite can only materialise from a juvenile thus d_{k1} is an upper bound on y_k with a lower bound of zero. We choose $\mathcal{H}_X$ to be the beta-binomial distribution. This distribution was also chosen by Drovandi and Pettitt (2013) and Ryan et al. (2016a) and reflects the fact that the binomial distribution is under-dispersed, when compared to the assumed model. The posterior predictive p-values for the conditional and marginal auxiliary models were p-value$_f$ = 0.20 and p-value$_g$ = 0.11, respectively. The posterior predictive p-values for the coupled auxiliary models ranged from 0.42 to 0.70 (over the different values of n considered). Figure 3 shows plots of sample statistics (mean and log variance) of the $\mathbf{y}_f^{(i)}$'s (the $\mathbf{y}_g^{(i)}$'s) against the $\mathbf{y}_{fX}^{(i)}$'s (the $\mathbf{y}_{gX}^{(i)}$'s). The posterior predictive p-values and Figure 3 indicate that the auxiliary models are adequate.

We consider finding Bayesian designs under the SIG utility and an alternative likelihood-based utility that we term *likelihood ratio* (LR). The likelihood ratio utility is given by

$$u_{LR}(\boldsymbol{\theta}, \mathbf{y}, \mathbf{d}) = 1 - \frac{\pi(\mathbf{y}|\mathbf{d})^{\frac{1}{2}}}{\pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d})^{\frac{1}{2}}}.$$

The design that maximises the expected LR utility, equivalently maximises the expected Hellinger distance between the prior and posterior distributions. The advantage of the expected Hellinger distance, over the expected Kullback-Liebler divergence, as a measure of discrepancy between the prior and posterior distributions is that the expected Hellinger distance is bounded on the $[0, 1]$ interval. This allows some level

Table 4: Mean (standard error) nested Monte Carlo approximation (under the auxiliary likelihood) to the expected SIG and LR utilities for the parasite model in Section 3.3 for the three different designs and each value of n.

	SIG		
Number of runs, n	Auxiliary Monte Carlo	Maximin Latin hypercube	Nested Monte Carlo
2	0.63 (0.001)	0.43 (0.002)	0.65 (0.001)
4	0.90 (0.004)	0.70 (0.005)	-
6	1.09 (0.006)	0.87 (0.008)	-
8	1.25 (0.008)	1.07 (0.010)	-
10	1.33 (0.011)	1.21 (0.012)	-
20	1.74 (0.021)	1.67 (0.029)	-
30	1.90 (0.040)	1.88 (0.038)	-
40	2.07 (0.061)	2.07 (0.027)	-
	LR		
Number of runs, n	Auxiliary Monte Carlo	Maximin Latin hypercube	Nested Monte Carlo
2	0.30 (0.0003)	0.21 (0.0003)	0.30 (0.0004)
4	0.36 (0.0003)	0.30 (0.0004)	-
6	0.40 (0.0004)	0.36 (0.0003)	-
8	0.44 (0.0004)	0.41 (0.0004)	-
10	0.46 (0.0003)	0.44 (0.0003)	-
20	0.57 (0.0003)	0.55 (0.0005)	-
30	0.61 (0.0003)	0.61 (0.0005)	-
40	0.65 (0.0004)	0.64 (0.0004)	-

Table 5: Average computing time (in hours) for designs found under auxiliary and nested Monte Carlo for the SIG utility for the parasite model in Section 3.3.

Number of runs, n	2	4	6	8	10	20	30	40
Coupled auxiliary model	0.2	0.6	1.1	1.8	2.6	10.0	22.9	44.6
Monte Carlo	6.9	-	-	-	-	-	-	-

of absolute assessment of design performance. We find designs using ACE under auxiliary Monte Carlo for each value of n. For $n = 2$ we also use ACE to find a design using the nested Monte Carlo approximation (under the auxiliary likelihood). We only considered $n = 2$ here due to the computational expense of finding designs under nested Monte Carlo for $n > 2$. As a further comparison we also find maximin Latin hypercube designs for each value of n. Table 4 shows the mean and standard error of twenty nested Monte Carlo approximations (under the auxiliary likelihood) to the expected utility for the three designs and each value of n. Additionally, Table 5 shows the average computing times for the designs found under auxiliary and nested Monte Carlo for the SIG utility. The corresponding times for the LR utility are similar.

For $n = 2$ the design found under the nested and auxiliary Monte Carlo approximations, for both utilities, have very similar performance. However note that the design found under auxiliary Monte Carlo only took about 3% of the time to find. The Latin hypercube design is relatively poor for small n but becomes indistinguishable in performance from the auxiliary Monte Carlo design as n approaches 40. Since finding such a design has a minimal computational overhead, this design would be favoured. However, to know that this design had essentially the same performance as the SIG or LR design first requires us to find the SIG or LR design.

3.4 Model comparison in epidemiological dynamics

We now consider a modified version of the model comparison example considered by Dehideniya et al. (2018). Here the interest is in designing experiments which are best able to distinguish between competing epidemiological models for the spread of a disease in a given population. Suppose the population size is assumed known and is $K = 100$. The experiment involves observing y_k the number of infected individuals in the population at time t_k, for $k = 1, \ldots, n$. Thus the design problem is to choose $\mathbf{D} = (t_1, \ldots, t_n)$.

Dehideniya et al. (2018) consider four competing models. First, let $S(t)$, $E(t)$, and $I(t)$ denote the number of susceptible, exposed and infected individuals, respectively, at time t, with the constraint that $S(t) + E(t) + I(t) = K$. We assume that at time $t = 0$, $S(0) = K$ and $I(0) = E(0) = 0$. Each model described below is identified by a model indicator given by m.

1. **Death model** $(m = 1)$

   In the death model, individuals transition from susceptible to infected directly, i.e. they do not become exposed as an intermediate step. The rate of transition is proportional to the number of susceptible individuals left in the population. These dynamics are given by

   $$P\left(S(t + \delta t) = S(t) - 1, E(t + \delta t) = E(t), I(t + \delta t) = I(t) + 1\right) = \theta_{11} S(t)\delta t + o(\delta t),$$

   where $\boldsymbol{\theta}_1 = (\theta_{11})$ is an unknown parameter.

2. **Susceptible-Infected (SI) model** $m = 2$

   The SI model modifies the death model so that the rate of transition from susceptible to infected is proportional to the rate at which susceptible and infected individuals meet. These dynamics are given by

   $$P\left(S(t + \delta t) = S(t) - 1, E(t + \delta t) = E(t), I(t + \delta t) = I(t) + 1\right)$$
   $$= (\theta_{21} + \theta_{22} I(t))\, S(t)\delta t + o(\delta t),$$

   where $\boldsymbol{\theta}_2 = (\theta_{21}, \theta_{22})$ are unknown parameters.

3. **Susceptible-Exposed-Infected (SEI) model** $(m = 3)$

   In the SEI model, the status transitions from susceptible to exposed to infected. The rate of these two transitions are proportional to the number of susceptible and exposed individuals, respectively. These dynamics are given by

   $$P\left(S(t + \delta t) = S(t) - 1, E(t + \delta t) = E(t) + 1, I(t + \delta t) = I(t)\right)$$
   $$= \theta_{31} S(t)\delta t + o(\delta t),$$
   $$P\left(S(t + \delta t) = S(t), E(t + \delta t) = E(t) - 1, I(t + \delta t) = I(t) + 1\right)$$
   $$= \theta_{32} E(t)\delta t + o(\delta t),$$

   where $\boldsymbol{\theta}_3 = (\theta_{31}, \theta_{32})$ are unknown parameters.

4. **Susceptible-Exposed-Infected-II (SEI-II) model** $(m = 4)$

   The SEI-II model is a modification of the SEI model such that the rate of transition from susceptible to exposed is proportional to the rate at which susceptible and infected individuals meet. These dynamics are given by

   $$P\left(S(t + \delta t) = S(t) - 1, E(t + \delta t) = E(t) + 1, I(t + \delta t) = I(t)\right)$$
   $$= (\theta_{41} + \theta_{42} I(t))\, S(t)\delta t + o(\delta t),$$
   $$P\left(S(t + \delta t) = S(t), E(t + \delta t) = E(t) - 1, I(t + \delta t) = I(t) + 1\right)$$
   $$= \theta_{43} E(t)\delta t + o(\delta t),$$

   where $\boldsymbol{\theta}_4 = (\theta_{41}, \theta_{42}, \theta_{43})$ are unknown parameters.

Table 6: Posterior predictive p-values for the marginal auxiliary model and minimum and maximum posterior predictive p-values for the coupled auxiliary model for the four competing models in the epidemiological dynamics example in Section 3.4.

Model, m	Marginal auxiliary model	Coupled auxiliary model Min	Coupled auxiliary model Max
1	0.52	0.56	0.62
2	0.43	0.48	0.51
3	0.44	0.53	0.63
4	0.49	0.24	0.45

An experimental aim involving model comparison can be represented by a utility function denoted by $u(m, \mathbf{y}, \mathbf{d})$. The utility used in this example is known as the 0-1 utility (Overstall et al., 2018) and is given by

$$u_{01}(m, \mathbf{y}, \mathbf{d}) = I(m = \tilde{m}). \tag{18}$$

In (18), $\tilde{m}$ is the posterior modal model given by

$$\tilde{m} = \arg \max_{m \in \mathcal{M}} \pi(m|\mathbf{y}, \mathbf{d}),$$

where $\mathcal{M}$ is a set of competing models and $\pi(m|\mathbf{y}, \mathbf{d})$ is the posterior model probability of model m. The posterior model probability is given by

$$\pi(m|\mathbf{y}, \mathbf{d}) = \frac{\pi(\mathbf{y}|\mathbf{d}, m)\pi(m)}{\sum_{m \in \mathcal{M}} \pi(\mathbf{y}|\mathbf{d}, m)\pi(m)},$$

where

$$\pi(\mathbf{y}|\mathbf{d}, m) = \int_{\Theta_m} \pi(\mathbf{y}|\boldsymbol{\theta}_m, \mathbf{d}, m)\pi(\boldsymbol{\theta}_m|m)\mathrm{d}\boldsymbol{\theta}_m \tag{19}$$

is the marginal likelihood and $\pi(m)$ is the prior model probability, respectively, for model m. In (19), $\pi(\mathbf{y}|\boldsymbol{\theta}_m, \mathbf{d}, m)$ is the likelihood for model m with parameters $\boldsymbol{\theta}_m$ having prior distribution with pdf $\pi(\boldsymbol{\theta}_m|m)$.

The 0-1 utility is one if the true model is $\tilde{m}$ and zero otherwise. The design that maximises the 0-1 utility (over the joint distribution of m and $\mathbf{y}$) equivalently maximises the expected posterior model probability of the posterior modal model. The 0-1 utility is likelihood-based only depending on the marginal likelihoods of all models under consideration.

We consider $n = 1, 2, 4, 6, \ldots, 20$. We include $n = 1$ due to the observed large increase in the expected 0-1 utility as n increases from $n = 1$ to $n = 2$. To apply the methodology proposed in this paper, we find a coupled auxiliary model for each of the four models. Table 6 shows the posterior predictive p-values for the marginal auxiliary model and the minimum and maximum posterior predictive p-values for the coupled auxiliary models over the different values of n considered. Furthermore, Figure 4 shows plots of sample mean and variance of $\mathbf{y}_g^{(i)}$ against $\mathbf{y}_{gX}^{(i)}$ for each model. There appears no reason to suspect an inadequate coupled auxiliary model for any of the models.

We find Bayesian designs for each value of n under the 0-1 utility. Table 7 shows the nested Monte Carlo approximation (under the auxiliary likelihood) to the expected 0-1 utility for each value of n. There is a sharp increase in expected 0-1 utility as the number of time points goes from $n = 1$ to $n = 2$. However, the approximate expected utility decreases as n increases. This is technically not possible for optimal designs for each value of n. This information loss is due to the use of an approximate model instead of the assumed model in the calculation of the approximate expected utility. To investigate this, we perform an experiment where we approximate the expected utility using the ABC marginal likelihood with $\epsilon = 0$. The corresponding approximations are shown in Table 7. Due to the computational expense of computing the ABC approximations, we are only able to feasibly do so for $n \leq 10$. For example, just one evaluation of the ABC approximation to the expected 0-1 utility for $n = 10$ took over 42 hours in computing time. However, from the results for $n \leq 10$, increasing n does lead to a small increase in the expected 0-1 utility for the designs found under the auxiliary Monte Carlo approach.

Figure 4: Plots of sample mean (left column) and variance (right column) of $\mathbf{y}_g^{(i)}$ against $\mathbf{y}_{gX}^{(i)}$ for model m on mth row for the epidemiological dynamics example in Section 3.4. In each plot a straight line of slope one through the origin as been included as a reference.
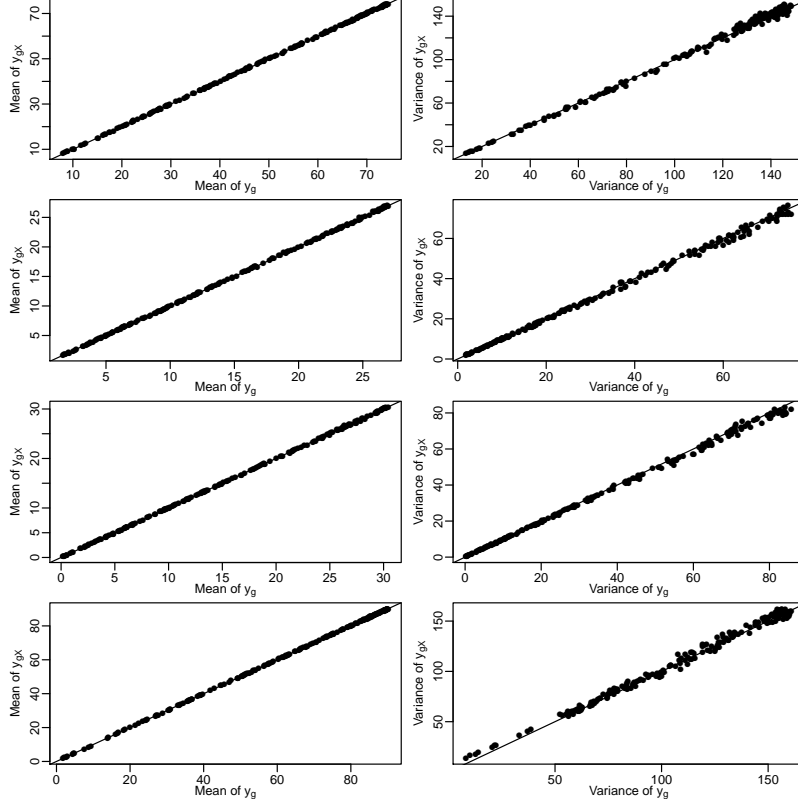

Table 7: Mean (standard error) nested and ABC approximations to the expected 0-1 utility for the designs found using the auxiliary Monte Carlo approximation for each value of n for the epidemiological dynamics example in Section 3.4.

Number of of runs, n	Mean (standard error) Monte Carlo approximation	ABC approximation
1	0.676 (0.076)	0.673
2	0.902 (0.075)	0.901
4	0.900 (0.075)	0.908
6	0.898 (0.076)	0.919
8	0.895 (0.075)	0.928
10	0.889 (0.075)	0.938
12	0.884 (0.075)	-
14	0.880 (0.076)	-
16	0.875 (0.076)	-
18	0.871 (0.076)	-
20	0.864 (0.076)	-

Table 8: Average computing time (in hours) for designs found under auxiliary Monte Carlo for each value of n for the epidemiological dynamics example in Section 3.4.

Number of runs, n	1	2	4	6	8	10
Average computing time	0.04	0.10	0.22	0.40	0.64	0.95
Number of runs, n	12	14	16	18	20	
Average computing time	1.33	1.78	2.29	3.03	3.62	

4   Discussion

In this paper we have introduced a general-purpose approach for finding Bayesian designs under intractable likelihood models. This approach can be used for all likelihood-based utility functions, for realistic-sized experiments and for experimental aims of parameter estimation and model comparison.

The methodology in this paper can only be applied to likelihood-based utility functions. This is where the utility is a functional of the likelihood and/or marginal likelihood. Some experimental aims may not be well represented by such a utility function. An example in the case of parameter estimation, and, in particular, point estimation, is the negative squared error loss (NSEL, e.g. Overstall and Woods 2017) given by

$$u_{NSEL}(\boldsymbol{\theta}, \mathbf{y}, \mathbf{d}) = -\sum_{l=1}^{p} \left(\theta_l - \mathrm{E}\left(\theta_l | \mathbf{y}, \mathbf{d} \right) \right)^2 .$$

The design that maximises the expected NSEL utility, equivalently minimises the trace of the expected posterior variance matrix of $\boldsymbol{\theta}$. The NSEL utility is not a direct functional of the likelihood or marginal likelihood, and, as such, the coupled auxiliary model methodology presented in this paper would need to be extended to account for this type of utility. The conditional auxiliary model could be used to approximate the posterior mean using importance sampling. However, as we saw in Section 3, this does carry a large computational burden. We actually argue that likelihood-based utility functions are more appropriate for the types of posterior distribution that may be encountered in intractable likelihood models. For example, the NSEL utility only seems appropriate when the posterior is close to a symmetric and unimodal distribution. Otherwise, the posterior mean may be in a region of the parameter space with low posterior density.

We have used the ACE algorithm to maximise the proposed approximation to the expected utility. This is because it is the current state of the art of algorithms of finding Bayesian designs. However the methodology proposed here is applicable with any optimisation algorithm that only requires a Monte Carlo approximation to the expected utility.

Some issues for future investigation are now discussed. Ryan et al. (2016a) suggested the combination of auxiliary modelling and using normal-based approximations to posterior quantities (e.g. Long et al., 2013; Overstall et al., 2018). However, we believe that the type of posterior distribution encountered in intractable likelihood models may not be well approximated by a normal distribution. Instead the use of other deterministic approximations, such as expectation propagation (e.g. Gelman et al., 2014, pages 338-343) could be more suitable.

Finally, the auxiliary Monte Carlo approximation to the expected utility is applicable to all models from which we can generate samples, including tractable likelihood models. Indeed, for the compartmental model example in Section 3.2.1 the use of the auxiliary approximation allowed a Bayesian design to be found in a fraction of the time required to find a design under the standard nested Monte Carlo approximation, albeit with lower expected performance. The trade-off between accuracy and computing time for tractable likelihood models is an avenue for future research.

A Details on the approximate coordinate exchange algorithm

A.1 ACE algorithm

1. Choose an initial design $\mathbf{D}^0 = (D_1^0, \ldots, D_q^0)$ and set the current design to be $\mathbf{D}^C = (D_1^C, \ldots, D_q^C) = \mathbf{D}^0$.

2. For $i = 1, \ldots, q$ complete the following steps

   (a) Let $U^i(D) = L(D_1^C, \ldots, D_{i-1}^C, D, D_{i+1}^C, \ldots, D_q^C)$ be the function given by the expected utility which only varies over the design space, $\mathcal{D}^i$, for the ith element.

   (b) For $j = 1, \ldots, Q$, evaluate the Monte Carlo approximation to the expected utility given by

   $$z_j = \hat{U}^i(D_j),$$

   for $\{D_1, \ldots, D_Q\} \in \mathcal{D}^i$. Fit a Gaussian process emulator to $\{z_j, d_j\}_{j=1}^Q$ and set $\tilde{U}^i(D)$ to be the resulting predictive mean.

   (c) Find
   $$D_i^* = \operatorname{argmin}_{D^i \in \mathcal{D}} \tilde{U}^i(D),$$
   and let $\mathbf{D}^* = (D_1^C, \ldots, D_{i-1}^C, D^*, D_{i+1}^C, \ldots \ldots, D_q^C)$ be the proposed design.

   (d) Set $\mathbf{D}^C = \mathbf{D}^*$ with probability p^*.

3. Return to step 2.

In step 2d, we accept the proposed design, $\mathbf{D}^*$ with probability p^*. The proposed design originates from from the Gaussian process emulator. Similar to all statistical models, Gaussian process emulators can fit inadequately. To mitigate the effects of an inadequate emulator, Overstall and Woods (2017) proposed a comparison between the proposed design $\mathbf{D}^*$ and the current design $\mathbf{D}^C$ which is independent of the current Gaussian process emulator. Note that the proposed design $\mathbf{D}^*$ should be accepted if

$$U(\mathbf{D}^*) > U(\mathbf{D}^C). \tag{20}$$

For $b = 1, \ldots, B$ we generate samples $\{u_*^b\}_{b=1}^B$ and $\{u_C^b\}_{b=1}^B$ as follows

$$
\begin{aligned}
u_*^b &= u(\mathbf{y}^{*b}, \boldsymbol{\theta}^{*b}, \mathbf{D}^*), \\
u_C^b &= u(\mathbf{y}^b, \boldsymbol{\theta}^b, \mathbf{D}^C),
\end{aligned}
$$

where $\{\boldsymbol{\theta}^{*b}, \mathbf{y}^{*b}\}_{b=1}^B$ and $\{\boldsymbol{\theta}^b, \mathbf{y}^b\}_{b=1}^B$ are samples from the joint distribution of $\boldsymbol{\theta}$ and $\mathbf{y}$ conditional on $\mathbf{D}^*$ and $\mathbf{D}^C$, respectively. We use these samples to perform a Bayesian hypothesis test of (20). The form of the Bayesian hypothesis test, as described in Overstall and Woods (2017), assumes that the u_*^b's and u_C^b's are continuous and their distribution reasonably assumed normal. In this case, the probability of accepting the proposed design is

$$p^* = 1 - F\left(-\frac{B\bar{u}_* - B\bar{u}_C}{\sqrt{2B\hat{v}}}\right),$$

where $F(\cdot)$ is the distribution function of the t-distribution with $2B - 2$ degrees of freedom,

$$\hat{v} = \frac{\sum_{b=1}^B (u_C^b - \bar{u}_C)^2 + \sum_{b=1}^B (u_*^b - \bar{u}_*)^2}{2B - 2},$$

and $\bar{u}_C$ and $\bar{u}_*$ are the sample means of the u_C^b's and u_*^b's, respectively.

The assumption of normality will clearly be violated for the 0-1 loss function for model discrimination, described in Section 3.4, where the u_*^b's and u_C^b's will be binary in the set $\{0,1\}$. For such loss functions, Overstall et al. (2018) introduced a modification where

$$p^* = 1 - \frac{1}{B} \sum_{b=1}^{B} F\left(\rho_C^b; 1 + B\bar{u}_*, 1 + B - B\bar{u}_*\right),$$

where $F\left(\cdot; a, b\right)$ denotes the distribution function of the Beta(a, b) and $\left\{\rho_C^b\right\}_{b=1}^{B}$ is a sample from Beta$\left(1 + B\bar{u}_C, 1 + B - B\bar{u}_C\right)$.

References

Abramowitz, M. and Stegun, I. A. (2002) Handbook of mathematical functions. Available at www.math.sfu.ca/$\sim$cbm/aands/toc.htm.

Amzal, B., Bois, F. Y., Parent, E. and Robert, C. P. (2006) Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical Association* **101**, 773–785.

Arbenz, P. (2013) Bayesian copulae distributions, with application to operational risk management-some comments. *Methodology and Computing in Applied Probability* **15**, 103–108.

Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007) *Optimum experimental designs, with SAS*. Oxford University Press Inc., New York.

Azzalini, A. and Capatanio, A. (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *Journal of the Royal Statistical Society Series B* **65**, 367–389.

Beaumont, M. A., Zhang, W. and Balding, D. J. (2002) Approximate Bayesian computation in population genetics. *Genetics* **162(4)**, 2025–2035.

Chaloner, K. and Verdinelli, I. (1995) Bayesian experimental design: A review. *Statistical Science* **10**, 273–304.

Conti, S. and O'Hagan, A. (2010) Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference* **140**, 640–651.

Dean, A., Morris, M., Stufken, J. and Bingham, D. (eds.) (2015) *Handbook of Design and Analysis of Experiments*. Boca Raton: CRC Press.

Dehideniya, M., Drovandi, C. C. and McGree, J. M. (2018) Optimal Bayesian design for discriminating between models with intractable likelihoods in epidemiology. *Computational Statistics and Data Analysis* **In press**.

Denham, D., Ponnudurai, T., Nelson, G., Guy, F. and Rogers, R. (1977) Studies with brugia pahangi. i. parasitological observations on primary infections of cats (felis catus). *Journal for Parasitology* **2**, 239–247.

Drovandi, C. C. and Pettitt, A. N. (2013) Bayesian experimental design for models with intractable likelihoods. *Biometrics* **69(4)**, 937–948.

Drovandi, C. C., Pettitt, A. N. and Faddy, M. J. (2011) Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **60(3)**, 503–524.

Drovandi, C. C., Pettitt, A. N. and Lee, A. (2015) Bayesian indirect inference using a parametric auxiliary model. *Statistical Science* **30**, 72–95.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2014) *Bayesian Data Analysis*. Chapman and Hall 3rd edn.

Gillespie, D. (1977) Exact stochastic inference of coupled chemical reactions. *Journal of Physical Chemistry* **81**.

Gourieroux, C., Monfort, A. and Renault, E. (1993) Indirect inference. *Journal of Applied Econometrics* **8**, 85–118.

Hainy, M., Müller, W. and Wagner, H. (2013) Likelihood-free simulation-based optimal design: An introduction. In *Topics in Statistical Simulation* (eds. V. Melas, S. Mignani, P. Monari and L. Salmaso) 271–278.

Heggland, K. and Frigessi, A. (2004) Estimating functions in indirect inference. *Journal of the Royal Statistical Society Series B* **104**, 117–131.

Huan, X. and Marzouk, Y. M. (2013) Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics* **232(1)**, 288–317.

Jones, M., Goldstein, M., Jonathan, P. and Randell, D. (2016) Bayes linear analysis for Bayesian optimal experimental design. *Journal of Statistical Planning and Inference* **171**, 115–129.

Lange, K. (2013) *Optimization*. Springer 2nd edn.

Lindley, D. (1956) On a measure of the information provided by an experiment. *Annals of Mathematical Statistics* **27**, 986–1005.

Long, Q., Scavino, M., Tempone, R. and Wang, S. (2013) Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Computer Methods in Applied Mechanics and Engineering* **259**, 24–39.

Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **100(26)**, 15324–15328.

Meyer, R. and Nachtsheim, C. (1995) The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometerics* **37(1)**, 60–69.

Müller, P. (1999) Simulation-based optimal design. *Bayesian Statistics* **6**, 459–474.

Müller, P. and Parmigiani, G. (1995) Optimal design via curve fitting of Monte Carlo experiments. *Journal of the American Statistical Association* **90(432)**, 1322–1330.

Müller, P., Sanso, B. and De Iorio, M. (2004) Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association* **99**, 788–798.

Nelson, R. (1998) *An Introduction to Copulas*. Springer Verlag New York.

Overstall, A. M., McGree, J. M. and Drovandi, C. C. (2018) An approach for finding fully Bayesian optimal designs using normal-based approximations to loss functions. *Statistics and Computing* **28**, 343–358.

Overstall, A. M. and Woods, D. C. (2016) Multivariate emulation of computer simulators: model selection and diagnostics with application to a humanitarian relief model. *Journal of the Royal Statistical Society, Series C* **65**, 483–505.

— (2017) Bayesian design of experiments using approximate coordinate exchange. *Technometrics* **59**, 458–470.

Overstall, A. M., Woods, D. C. and Adamou, M. (2017) *acebayes: Optimal Bayesian Experimental Design using the ACE Algorithm*. URL `https://cran.r-project.org/web/packages/acebayes/index.html`. R package version 1.3.4.

Pagendam, D. and Pollett, P. (2013) Optimal design of experimental epidemics. *Journal of Statistical Planning and Inference* **143**, 563–572.

Panagiotelis, A., Czado, C. and Joe, H. (2012) Pair copula constructions for multivariate discrete data. *Journal of American Statistical Association* **107**, 1063–1072.

Parker, B., Gilmour, S., Schormans, J. and Maruri-Aguilar, H. (2015) Optimal design of measurements on queueing systems. *Queueing Systems* **79**, 365–390.

Price, D., Bean, N., Ross, J. and Tuke, J. (2016) On the efficient determination of optimal Bayesian experimental designs using ABC: A case study in optimal observation of epidemics. *Journal of Statistical Planning and Inference* **172**, 1–15.

Riley, S., Donnelly, C. and Ferguson, N. (2003) Robust parameter estimation techniques for stochastic within-host macroparasite models. *Journal of Theoretical Biology* 419–430.

Ryan, C., Drovandi, C. and Pettitt, A. N. (2016a) Optimal Bayesian experimental design for models with intractable likelihoods using indirect inference applied to biological process models. *Bayesian Analysis* **11**, 857–883.

Ryan, E. G., Drovandi, C. C., McGree, J. M. and Pettitt, A. N. (2016b) A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review* **84**, 128–154.

Ryan, E. G., Drovandi, C. C., Thompson, M. H. and Pettitt, A. N. (2014) Towards Bayesian experimental design for nonlinear models that require a large number of sampling times. *Computational Statistics and Data Analysis* **70**, 45–60.

Ryan, K. (2003) Estimating expected information gains for experimental designs with application to the random fatigue-limit model. *Journal of Computational and Graphical Statistics* **12**, 585–603.

Santner, T., Williams, B. and Notz, W. (2003) *The Design and Analysis of Computer Experiments*. Springer, New York.

Sisson, S. A., Fan, Y. and Tanaka, M. (2007) Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **104(6)**, 1760–1765.

Tavaré, S., Balding, D. J., Griffiths, R. and Donnelly, P. (1997) Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518.

Tran, M., Nott, D. J. and Kohn, R. (2017) Variational Bayes with intractable likelihood. *Journal of Computational and Graphical Statistics* To appear.

Weaver, B., Williams, B., Anderson-Cook, C. and Higdon, D. (2016) Computational enhancements to Bayesian design of experiments using Gaussian processes. *Bayesian Analysis* **11**, 191–213.

Wood, S. (2010) Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466**, 1102–1104.

Woods, D., Overstall, A., Adamou, M. and Waite, T. (2017) Bayesian design of experiments for generalised linear models and dimensional analysis with industrial and scientific application. *Quality Engineering* **29**, 91–103.