Experimental design for models with intractable likelihoods

Mitchell O'Sullivan

Queensland University of Technology
Dr Chris Drovandi

1 Introduction

Optimal designs of experiments are useful when it is of interest to minimise the use of man-power or costs involved in performing experiments. In this paper the methodology we use to find the optimal designs for models with intractable likelihoods but where it is straightforward to simulate from the model is summarised.

1.1 Motivating Example

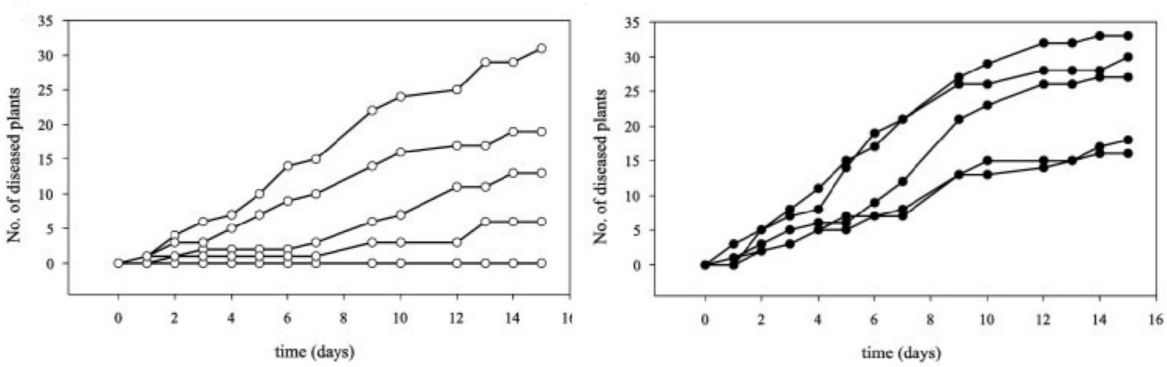


Figure 1.1: Disease progress curves for epidemics of damping-off of Radish caused by *R. solani* in five replicate microcosms of radish in the presence (left) and absence (right) of the biological control agent *T. viride*. Source: Gibson *et al.* (2004)

Gibson *et al.* (2004) use 2 stochastic compartmental models to describe the data (shown in Figure 1.1), accounting for both primary (from inoculum in the soil, $\mathbf{r_p}$) and secondary (plant-to-plant transmission, $\mathbf{r_s}$). The models also incorporate the time decaying susceptibility of the plants characteristic to infection by *R. Solani*. The first is a simple SI model where the transition from the susceptible (S) class to the infected (I) class is governed by:

$$\Pr(I(t+\mathrm{d}t) = I(t)+1) = (\mathbf{r_p} + \mathbf{r_s} I(t)) s(t) S(t) \mathrm{d}t,$$

where $s(t) = \mathrm{e}^{-\mathbf{a}t}$, and the parameters of the model are bolded. This forms a Markov model with a likelihood function which is computationally too intensive to be of use when designing for the model. In the second model we introduce an exposed class (E) where infected plants initially transition to the E class by the same probabilistic rule,

$$\Pr(E(t+\mathrm{d}t) = E(t)+1) = (\mathbf{r_p} + \mathbf{r_s} I(t)) s(t) S(t) \mathrm{d}t,$$

where $s(t)$ is defined as before. In this model the plants move to the Infectious class after a period of time known as the latent period, denoted $\boldsymbol{\mu}$, also a parameter of the model. In this Non-Markovian model the exposed class are visually indistinguishable from the susceptible plants, meaning that only the infectious plants are observed. There is no analytic form for the likelihood of this model.

In their experiment, Gibson *et al.* (2004) had 5 independent replicates of 50 susceptible plants and observed the system at various times. An observation consists of the number of infected plants, identified by their symptoms. Sampling was done 13 different times to create the curves seen in figure 1.1. Our task was to design (i.e. obtain optimal sampling times) for both models and both treatment cases such that if the experiment were to be repeated, it could be done so in the most efficient manner.

2 Methods

2.1 Static Experimental Design

Of fundamental importance to the Bayesian statistician is the relationship between the 'posterior', 'prior', and the likelihood function. That is,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

In the presence of observed data, $\boldsymbol{y}$, we 'update' our initial beliefs of the distribution of the parameters encoded in the prior distribution, to arrivate at the posterior distribution. In this paper we account only for the parameter uncertainty in the model, and not the uncertainty in the model choice. The information available is contained within the prior distribution, $p(\boldsymbol{\theta})$, which includes data from the experiment in Figure 1.1.

In experimental design we first specify a utility function, $u(\boldsymbol{d}, \boldsymbol{y}, \boldsymbol{\theta})$, where $\boldsymbol{d}$ is the design applied, and $\boldsymbol{y}$ is data that is observed under the model parameters $\boldsymbol{\theta}$. The quantity of interest then is the expected utility of applying the design $\boldsymbol{d}$, where the expectation is taken over the prior distribution of the parameters, as well as the data that is yet to be observed,

$$u(\boldsymbol{d}) = \mathrm{E}_{\boldsymbol{\theta},\boldsymbol{y}}\left[u(\boldsymbol{d}, \boldsymbol{y}, \boldsymbol{\theta})\right] = \int_{\boldsymbol{y}} \int_{\boldsymbol{\theta}} u(\boldsymbol{d}, \boldsymbol{y}, \boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{d}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\boldsymbol{y},$$

where $p(\boldsymbol{y}|\boldsymbol{d}, \boldsymbol{\theta})$ is the likehood function of the future data given that the design $\boldsymbol{d}$ is applied. The optimal design, $\boldsymbol{d}^*$, maximises the expected utility of the design space, $\mathcal{D}$,

$$\boldsymbol{d}^* = \arg\max_{\boldsymbol{d} \in \mathcal{D}} u(\boldsymbol{d}).$$

Unfortunately, the expected utility function, $u(\boldsymbol{d})$, cannot be derived analytically and is computationally too intensive to optimise directly, so we must consider other approaches to find the optimal design, $\boldsymbol{d}^*$. Muller (1999) proposes to sample from the joint distribution

$$h(\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{y}) \propto u(\boldsymbol{d}, \boldsymbol{y}, \boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{d}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{d}) \tag{1}$$

which admits the marginal $h(\boldsymbol{d}) \propto u(\boldsymbol{d})p(\boldsymbol{d})$. We introduce the prior of the design, $p(\boldsymbol{d})$ in Eq$^{\mathrm{n}}$ 1, to include any constraints of the design. For example, sampling times are aften ordered. An estimate of the optimal design, $\boldsymbol{d}^*$, can then be found as the mode of this distribution. It is sometimes useful to sample from

$$h(\boldsymbol{d}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_J}, \boldsymbol{y_1}, \ldots, \boldsymbol{y_J}) \propto p(\boldsymbol{d}) \prod_{j=1}^{J} u(\boldsymbol{d}, \boldsymbol{y_j}, \boldsymbol{\theta_j}) p(\boldsymbol{y_j}|\boldsymbol{d}, \boldsymbol{\theta_j}) p(\boldsymbol{\theta})$$

which makes mode identification easier as it tightens the distribution around the modes since the marginal in $\boldsymbol{d}$ is $h(\boldsymbol{d}) \propto u(\boldsymbol{d})^J p(\boldsymbol{d})$. Muller proposes MCMC sampling from this joint distribution. The approach of Muller (1999) is given in Algorithm 1, where $q(\boldsymbol{d}|\boldsymbol{d^{i-1}})$ is an arbitrarily chosen transition function to explore the design space. We also note that the utility $u(\boldsymbol{d}, \boldsymbol{y}, \boldsymbol{\theta})$ must be approximated, which is discussed further below.

Algorithm 1 MCMC Algorithm for experimental design by Muller (1999).

Initialise: Draw $\boldsymbol{d^0} \sim p(\boldsymbol{d})$, draw $\boldsymbol{\theta^0} \sim p(\boldsymbol{\theta})$, and generate $\boldsymbol{y^0} \sim p(\boldsymbol{y}|\boldsymbol{\theta^0}, \boldsymbol{d^0})$, compute $u^0 = u(\boldsymbol{d^0}, \boldsymbol{y^0}, \boldsymbol{\theta^0})$

1: **for** $i = 1$ to M **do**
2: Propose $\boldsymbol{d^*} \sim q(\boldsymbol{d}|\boldsymbol{d^{i-1}}), \boldsymbol{\theta^*} \sim p(\boldsymbol{\theta}), \boldsymbol{y^*} \sim p(\boldsymbol{y}|\boldsymbol{\theta^*}, \boldsymbol{d^*})$
3: Compute $u^* = u(\boldsymbol{d^*}, \boldsymbol{y^*}, \boldsymbol{\theta^*})$
4: Compute $\alpha = \min\left(1, \frac{u^* p(\boldsymbol{d^*}) q(\boldsymbol{d^{i-1}}|\boldsymbol{d^*})}{u^{i-1} p(\boldsymbol{d^{i-1}}) q(\boldsymbol{d^*}|\boldsymbol{d^{i-1}})}\right)$
5: **if** $\alpha >$ unif(0, 1) **then**
6: Set $u^i = u^*, \boldsymbol{d^i} = \boldsymbol{d^*}, \boldsymbol{\theta^i} = \boldsymbol{\theta^*}, \boldsymbol{y^i} = \boldsymbol{y^*}$
7: **else**
8: Set $u^i = u^{i-1}, \boldsymbol{d^i} = \boldsymbol{d^{i-1}}, \boldsymbol{\theta^i} = \boldsymbol{\theta^{i-1}}, \boldsymbol{y^i} = \boldsymbol{y^{i-1}}$
9: **end if**
10: **end for**

By simulating data from the model in Algorithm 1 we see that the likelihood terms cancel in the Metropolis-Hastings ratio (line 4). We have a limited choice for the utility function since the likelihood

is not available for the models being considered. It seems that a Bayesian utility involving the posterior distribution is most suitable, and can be accomplished by approximating the true posterior distribution. We use a utility based on the concentration of the posterior distribution, which we approximate via ABC, described in the next section.

$$u(\boldsymbol{d}, \boldsymbol{y}) = 1/\det(\mathrm{Var}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d})).$$

2.2 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) is a simulation based method that builds up samples from the posterior by comparing simulated and observed data, and accepting parameter values from the prior which generate data close enough to the observed. The 'observed data' in the context of Algorithm 1, is the data, $\boldsymbol{y}$, that is simulated from the model at each iteration of the algorithm. The ABC posterior distribution is given by

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \epsilon) = \int_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) 1(\rho(\boldsymbol{y}, \boldsymbol{x}) \leq \epsilon) \, \mathrm{d}\boldsymbol{x},$$

where $\boldsymbol{x}$ is simulated data, $\rho(\cdot, \cdot)$ is a discrepancy function that compares the observed and simulated data, and ϵ is the tolerance which determines how close the data must be to be accepted. Clearly, for any $\epsilon > 0$, error is introduced to the approximation, but it is a tradeoff between accuracy and efficiency.

Typically, this discrepancy function compares a low dimensional set of summary statistics. For our purposes, since we only consider low dimensional designs we can directly compare the simulated and observed data,

$$\rho(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{D} \frac{|y_i - x_i|}{\mathrm{std}(x_i)}, \tag{2}$$

where $\mathrm{std}(\cdot)$ is the standard deviation and D is the number of design points. How we find $\mathrm{std}(x_i)$ is discussed later. The evaluate the utility when the likelihood is intractable we replace the true posterior with the ABC posterior. There are numerous way to sample from the ABC posterior, but because we are required to obtain an ABC posterior for each simulated dataset, $\boldsymbol{y}$, at each iteration of Algorithm 1, it was important to choose a suitable method. We chose ABC rejection (Beaumont et al., 2002) because the parameter draws from the prior, and their associated simulations (lines 1, 2 of Algorithm 2) can be stored and reused, saving on computational effort.

Algorithm 2 ABC Rejection

1: Generate $\boldsymbol{\theta^i} \sim p(\boldsymbol{\theta})$ for $i = 1, \ldots, N$
2: Simulate $\boldsymbol{x^i} \sim p(\boldsymbol{y}|\boldsymbol{\theta^i}, \boldsymbol{d})$ for $i = 1, \ldots, N$
3: Compute discrepancies $\rho^i = \rho(\boldsymbol{y}, \boldsymbol{x^i})$ for $i = 1, \ldots, N$, creating particles $\{\boldsymbol{\theta^i}, \rho^i\}_{i=1}^N$
4: Sort the particle set via the discrepancy ρ
5: Accept K of the particles with the lowest discrepency. Effectively $\epsilon = \rho_{(K)}$

In experimental design, the data, $\boldsymbol{y}$, is dependent on the design, $\boldsymbol{d}$. In this application, the design variable is the time at which to take samples. In order to pre-compute the ABC parameter draws and their data simulations, we discretise the design space to a regular grid with a minimum of $t_{\min}$, a maximum of $t_{\max}$, and an increment of t_{inc}. The stochastic process is simulated N times with the state of the process being recorded at each discrete time point. The Metropolis-Hastings algorithm samples over the discrete design space. Note for the models considered here that $t_{\min} = t_{\mathrm{inc}} = 0.5$, and $t_{\max} = 16$.
By pre-computing the prior simulations, the standard deviation of the prior-predictive distributions at each unique time point can be calculated. This is important so that the standard deviation, $\mathrm{std}(x_i)$, can be incorporated into our ABC discrepancy function (Equation 2). If we don't account for the variability in the data then sampling times with higher variability are unfairly penalised.

After running algorithm 1 we used a non-parametric estimate of the multivariate density of the design space to locate modes on the utility surface. It is important to use this non-parametric technique under a variety of settings to ensure we find the correct mode, so after obtaining candidate design points we calculated their utility to find the optimal design.

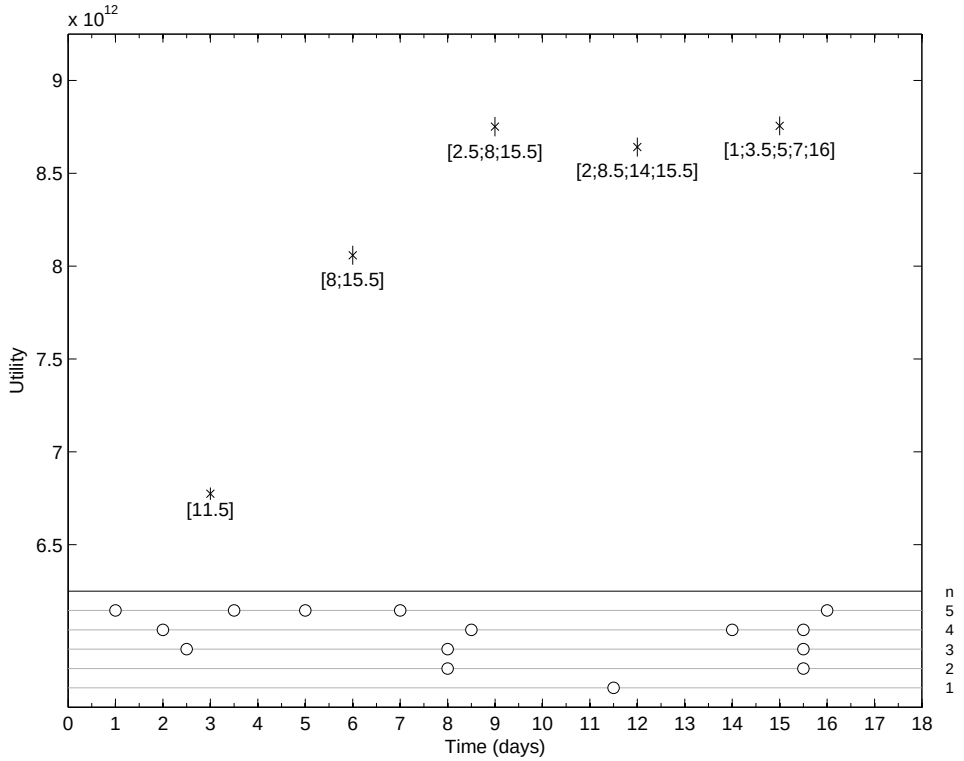
3 Results

In figure 3.1a we see the expected utility rise as we increase the number of samples taken up until $n = 3$. There is an apparent loss of information for 4 observations, and it appears that there is little increase in information from 3 observations to 5. We conjecture that this is an artefact of applying ABC in this application as well as our inability to accurately identify the modes in the higher dimensional design spaces. It is difficult to achieve low ABC tolerances as the number of observations to match on increases. It can be seen that, similar to figure 3.1b, the sampling times spread almost evenly across the design space for n up to 3. However, this is not the case when we look at the sampling schedule for $n > 3$. We believe this is due to the non-parametric technique employed when searching for the optimum design, which does not perform as well for higher dimensional designs. These two factors are believed to be responsible for the apparent lack of information obtained when taking more samples.

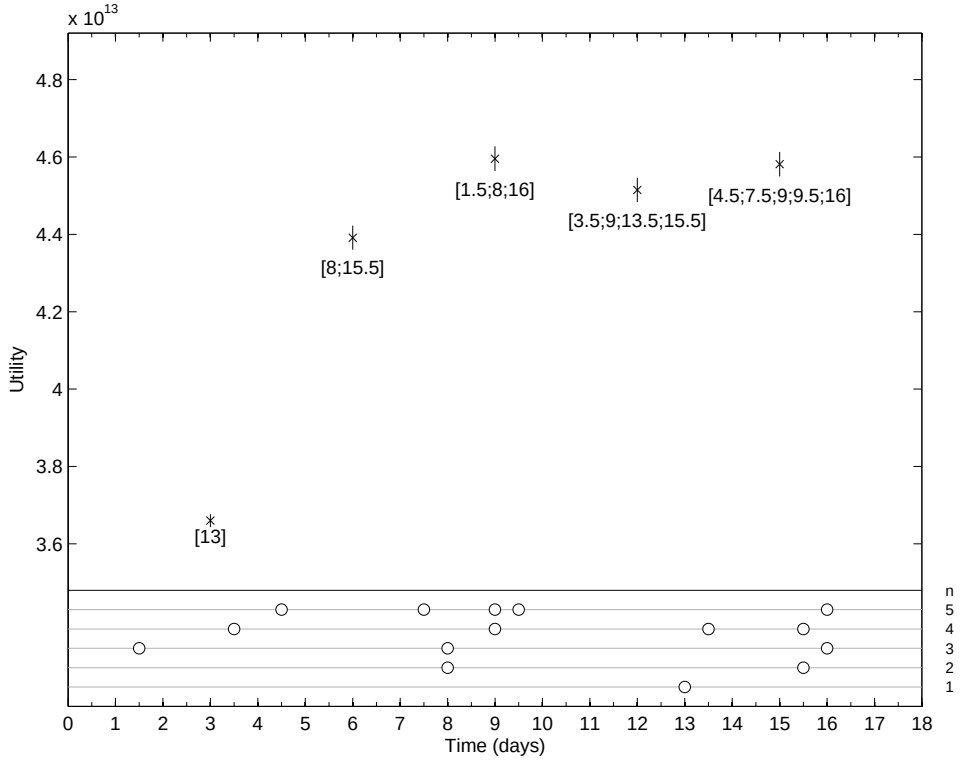
In figure 3.2a there is an increase in the information gained when taking $n = 1, \ldots, 4$ samples, although there is little improvement from 3 observations to 4, but no improvement from 3 to 5 observations. Figure 3.2b shows a similar pattern, but with a small upward trend for $n \geq 3$. Almost all of the optimal designs found for the Non-Markov models require sampling at $t = 0.5$, which is intuitive as we know that parameter $\boldsymbol{\mu} < 1$ (This is indicated by the nonzero number of infected plants at $t = 1$ in figure 1.1).

To further investigate the effect of the latent period on the optimal design we also looked at the Non-Markov no treatment case but with an altered utility function. Figure 3.3a shows the results when μ is excluded from the posterior precision calculation (removed μ from all posterior samples first). The expected utility behaves similarly to that in figure 3.2b, and the optimal designs found are similar to the markov model for $n \leq 3$. In figure 3.3b we see that when only concerned with the precision of the parameter μ there is little information to gain from sampling at times in addition to $t = 0.5$, which is due to μ being less than 1.

In figure 3.4, we use a finer discretisation of the design space, $t_{\min} = t_{\text{inc}} = 0.1$, and $t_{\max} = 3$, to investigate the optimal sampling times for μ. We see there is much more information to be gained for the parameter μ when we can sample this finer design space. Most observations times fall between 0 and 1 as expected.

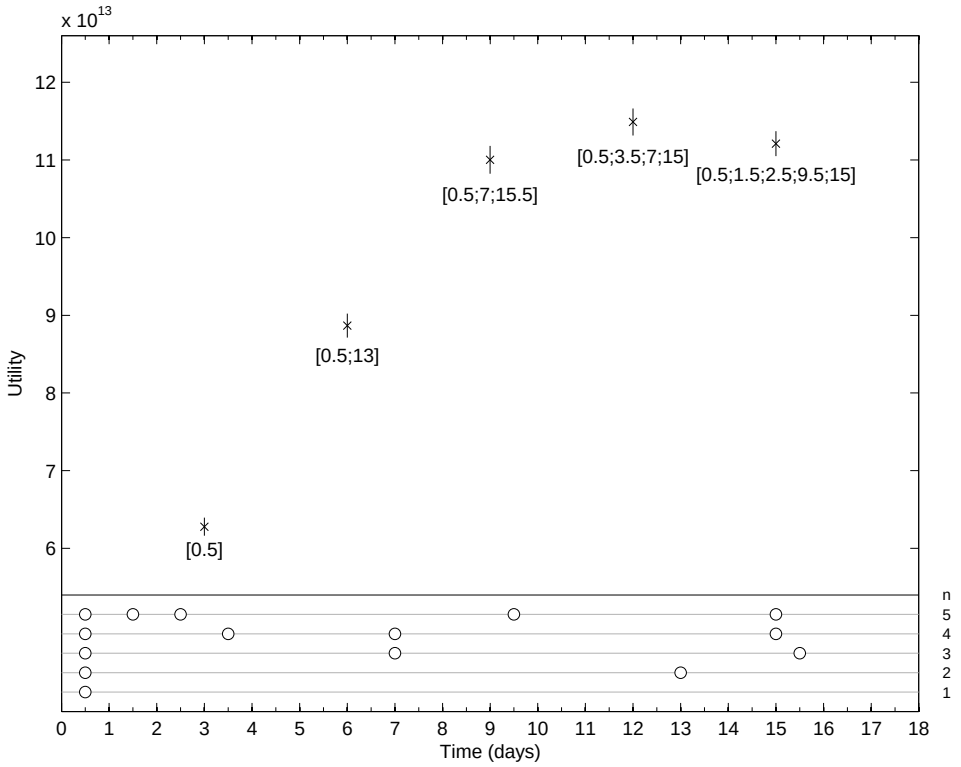


(a) No Treatment

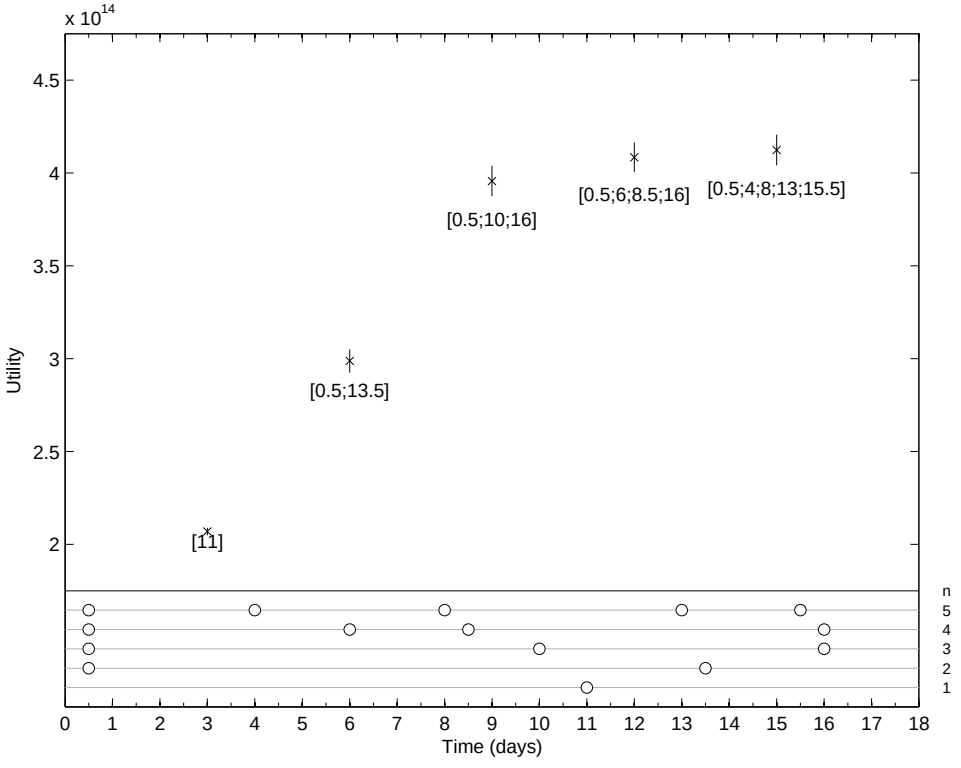


(b) Treatment

Figure 3.1: Optimal design points for $\boldsymbol{d^*} = (d_1, \ldots, d_n)$ and the expected utility of each design with error bars for the Markov model in the presence (Treatment) and absence (No Treatment) of *T.viride*.

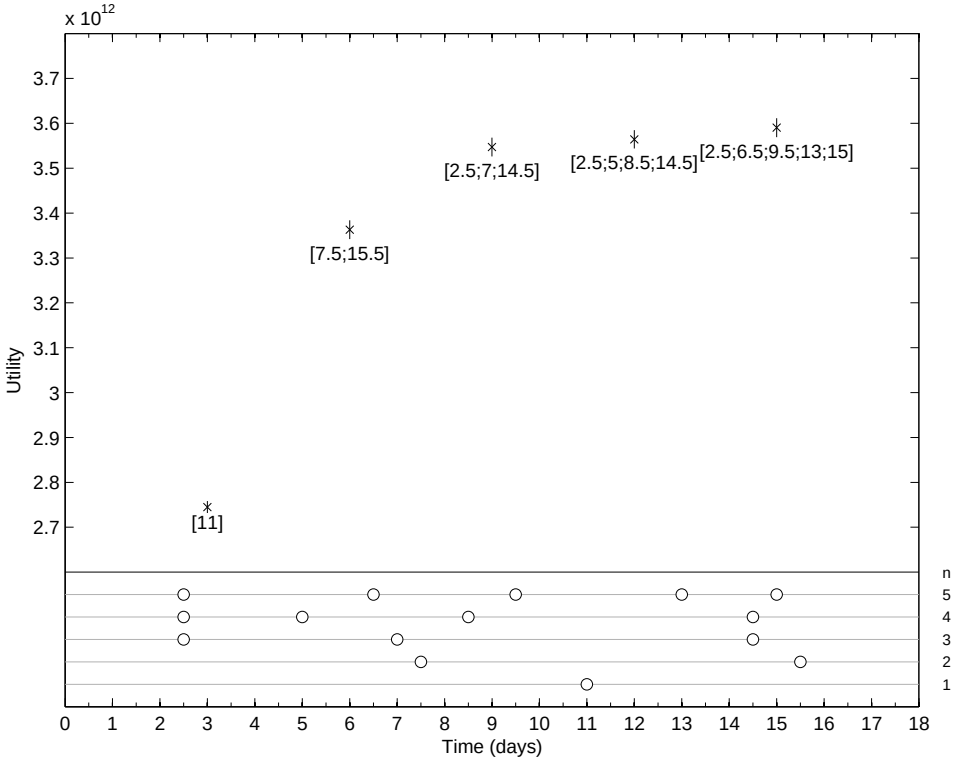


(a) No Treatment

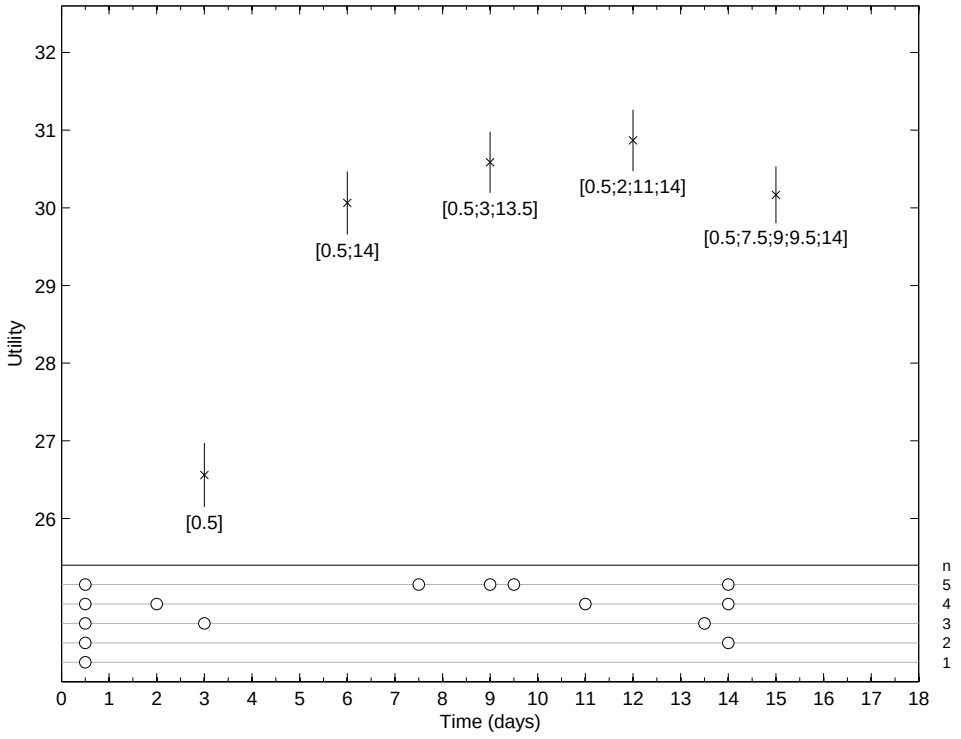


(b) Treatment

Figure 3.2: Optimal design points for $\boldsymbol{d}^* = (d_1, \ldots, d_n)$ and the expected utility of each design with error bars for the Non-Markov model in the presence (Treatment) and absence (No Treatment) of *T.viride*.



(a) Excluding μ



(b) Just μ

Figure 3.3: Optimal design points for $\boldsymbol{d^*} = (d_1, \ldots, d_n)$ and the expected utility of each design with error bars for the Non-Markov model without treatment where the utility is modified to either exclude the parameter μ (3.3a) or only include μ (3.3b).

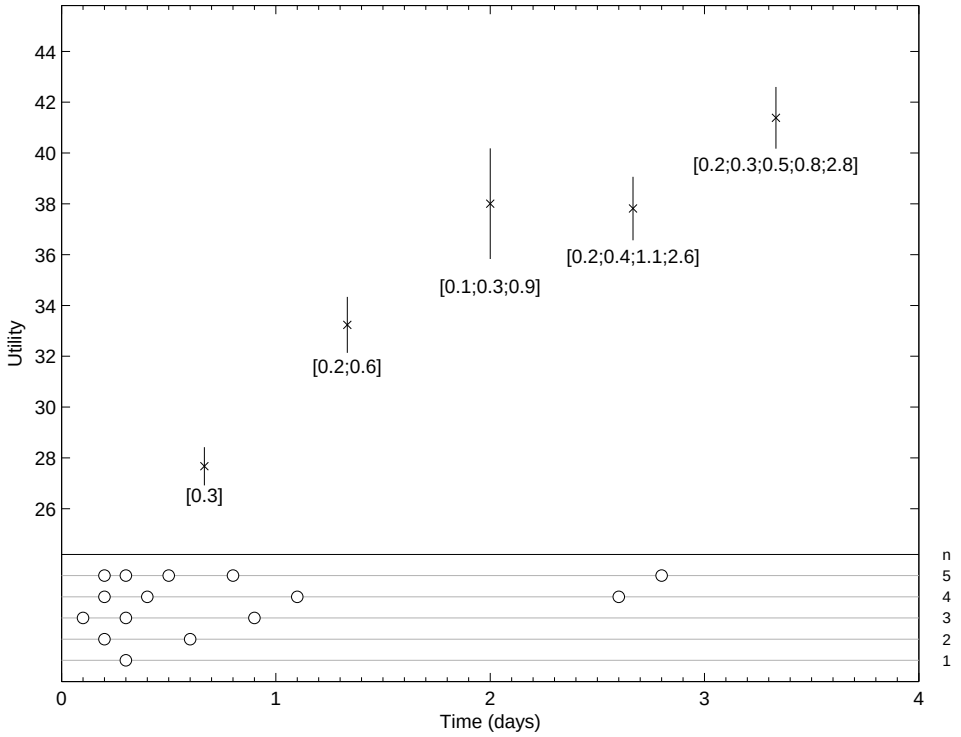


Figure 3.4: Optimal design points for $\boldsymbol{d}^* = (d_1, \ldots, d_n)$ and the expected utility of each design with error bars for the Non-Markov model without treatment where the utility is modified to only include μ, and the design space is discretised to a finer extent.

4 Limitations and Extensions

The methodology used here is best suited for low dimensional designs, and we need an informative prior to reduce ABC tolerances. We are also limited by the non-parametric estimator that suffers from the curse of dimensionality which makes identification of higher dimensional modes difficult. Future work includes extending to higher dimensional designs, designing for experiments with sampling from independent replicates, and incorporating the model uncertainty into a robust experimental design.

References

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

[2] Cook, A. R., Gibson, G. J., and Gilligan, C. A. (2008). Optimal observation times in experimental epidemic processes. *Biometrics*, 64(3):860-868.

[3] Drovandi, C. C., and Pettitt, A. N. (2012). Bayesian experimental design for models with intractable likelihoods. `http://eprints.qut.edu.au/53924/`

[4] Gibson, G. J., Kleczkowski, A., and Gilligan, C. A. (2004). Bayesian analysis of botanical epidemics using stochastic compartmental models. Proc Natl Acad Sci U S A. 2004 August 17; 101(33): 1212012124.

[5] Muller, P. (1999). Simulation-based optimal design. In *Bayesian statistics 6: proceedings of the Sixth Valencia International Metting, June 6-10, 1998*, volume 6, page 459. Oxford University Press, USA.