

**On the Advancement of Optimal
Experimental Design with Applications
to Infectious Diseases**

David Price

*Thesis submitted for the degree of
Doctor of Philosophy*

in

Applied Mathematics and Statistics

at

The University of Adelaide

Faculty of Engineering, Computer and Mathematical Sciences

School of Mathematical Sciences



August, 2015

Contents

Signed Statement	iv
List of Tables	v
List of Figures	vii
Acknowledgements	xii
Abstract	xiv
1 Introduction	1
2 Background	7
2.1 Markov Chains	7
2.2 Markovian Epidemic Models	12
2.3 Inference for Markov Chains	15
2.3.1 Frequentist Framework	15
2.3.2 Bayesian Framework	18
2.4 Optimal Experimental Design	44
2.4.1 Frequentist Optimal Experimental Design	46
2.4.2 Bayesian Optimal Experimental Design	51
3 On Exploiting the Locally Optimal Design of Exponentially-distributed Life Testing	61
3.1 Introduction	61

3.1.1	Technical Background	63
3.2	Single Observation Time	66
3.3	Multiple Observation Times	80
3.4	Conclusion	84
4	Optimal Experimental Design for an Epidemic Model	86
4.1	Markovian SIS Epidemic Model	89
4.2	Optimal Bayesian Experimental Designs via MCMC	92
4.2.1	Optimal Bayesian Experimental Design with an ABC Posterior Distribution	94
4.3	Optimal Bayesian Experimental Design for the Markovian SIS Epidemic Model	96
4.4	Conclusion	103
5	On the use of Approximate Bayesian Computation to obtain Optimal Designs Efficiently	105
5.1	ABCdE Algorithm	107
5.2	Examples	111
5.3	Results	118
5.3.1	Death Model	119
5.3.2	SI Model	124
5.3.3	SIS Model	127
5.4	Discussion	130
6	Optimal Experimental Design of Group Dose-response Challenge Experiments	137
6.1	Group Dose-response Challenge Experiments	137
6.2	Modelling Group Dose-response Challenge Experiments	140
6.2.1	Simple Model	141
6.2.2	Transmission Model	142

6.2.3	Latent Model	143
6.2.4	Complete Model	146
6.2.5	Specification of Model and Design Parameters for Bayesian Optimal Experimental Design	148
6.3	Bayesian Optimal Experimental Design for the Simple, Transmission and Latent Models	154
6.3.1	Optimal Design for the Simple Model	155
6.3.2	Sensitivity Analysis for the Simple Model	158
6.3.3	Optimal Design for the Transmission Model	162
6.3.4	Sensitivity Analysis for the Transmission Model	164
6.3.5	Optimal Design for the Latent Model	164
6.3.6	Sensitivity Analysis for the Latent Model	168
6.4	Bayesian Optimal Experimental Design for the Complete Model . . .	174
6.4.1	Sensitivity Analysis for the Complete Model	177
6.4.2	Estimation of Dose-response and Transmission Parameters Si- multaneously	184
6.5	Discussion	186
7	Optimal Experimental Design for Group Dose-response Challenge Experiments: An Alternative Utility Function	189
7.1	Mean Absolute Percentage Error	189
7.2	Optimal Experimental Designs	191
7.2.1	Latent Model	191
7.2.2	Complete Model	193
7.3	Discussion	194
8	Conclusion	197
8.1	Summary	197
8.2	Future Research	199
	Bibliography	202

Signed Statement

I, David Price, certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signature: *Date:*

List of Tables

4.3.1	Illustration of run-times for Drovandi and Pettitt [2013] algorithm, when evaluating one, two and three optimal observations of the Markovian SIS epidemic model.	97
4.3.2	Candidate optimal experimental designs under the different ‘smoothing factors’, for the SIS epidemic model when performing one, two or three observations.	98
4.3.3	Optimal observation times for one, two and three observations of the SIS epidemic model, using the algorithm of Drovandi and Pettitt [2013].	99
5.3.1	Comparison of the optimal observation times for the death process, from Cook <i>et al.</i> [2008], Drovandi and Pettitt [2013], ABCdE and a naïve approach. $ t $ is the pre-determined number of observation times, and i is the i^{th} time.	120
5.3.2	Illustration of run-times for Drovandi & Pettitt [2013] algorithm compared to ABCdE.	122
5.3.3	Comparison of the optimal observation times for the SI process, from Cook <i>et al.</i> [2008], Drovandi and Pettitt [2013], ABCdE, and a naïve approach. $ t $ is the pre-determined number of observation times, and i is the i^{th} time	124

5.3.4	Optimal observation times for the SIS process utilising the approach of Drovandi and Pettitt [2013], ABCdE, and a naïve approach. $ t $ is the pre-determined number of observation times, and i is the i^{th} time	127
6.2.1	Typical values of design parameters considered when determining the optimal experimental designs for the various models.	149
6.3.1	The top five designs, and the worst design, for the dose-response model, with percentage of information compared to the optimal design.	155
6.3.2	The top five designs, and the worst design, for the dose-response model with $N = 20$ chickens.	159
6.3.3	The top five designs, and the worst design, for the dose-response model, with an alternative prior distribution for β	161
6.3.4	The top five designs, and the worst design, for the transmission model, with percentage of information compared to the optimal. . .	163
6.3.5	The top five designs, and the worst design, for the latent model, with percentage of information compared to the optimal design. . .	165
6.3.6	The top five designs, and the worst design, for the latent model, when evaluated on a coarse grid for dose (step size of 0.5 across dose).168	
6.3.7	The top five designs, and the worst design, for the latent model, with an alternative number of available chickens.	169
6.3.8	Optimal experimental design for the latent model, with an alternative prior distribution for β	170
6.3.9	Erlang distributions used throughout the sensitivity analysis of the latent period. The Erlang(2,1) distribution denoted *, is the latent period distribution used in the original analysis.	172
6.3.10	Optimal experimental designs for the latent model, with a range of latent period distributions.	173

6.4.1	The top five designs, and the worst design, for the complete model, with percentage of information compared to the optimal.	176
6.4.2	The top five designs, and the worst design, for the complete model with $N = 20$ chickens.	177
6.4.3	The top five designs, and the worst design, for the complete model with an alternative prior distribution for β	178
6.4.4	Erlang distributions used throughout the sensitivity analysis of the latent period. The Erlang(2,1) distribution denoted *, is the latent period distribution used in the original analysis.	180
6.4.5	Optimal experimental designs for the complete model, with a range of latent period distributions.	180
6.4.6	The top five designs, and the worst design, for the complete model with $\beta_T = 4$	182
6.4.7	The top five designs, and the worst design, for the complete model with $\beta_T = 1$	183
6.4.8	The top five designs, and the worst design, for the complete model when estimating α, β and β_T simultaneously, with percentage of information compared to the optimal.	185
7.2.1	The top five designs, and the worst design, for the latent model with percentage of utility compared to the optimal.	192
7.2.2	The top five designs, and the worst design, for the complete model with percentage of utility compared to the optimal.	193

List of Figures

- 2.3.1 Trace plots of three Metropolis-Hastings chains from their initial point, showing burn-in phase. 29
- 2.3.2 Trace plots of three Metropolis-Hastings chains from their initial point, for all 50,000 iterations. 31
- 2.3.3 Diagram of Metropolis-Hastings chain (thin black line) to estimate target density (thick black ellipse), saving only consecutive points (red dots). 33
- 2.3.4 Diagram of Metropolis-Hastings chain (thin black line) to estimate target density (thick black ellipse), saving a thinned sample (red dots). 34
- 2.3.5 Diagram of Metropolis-Hastings chain (thin black line) to estimate target density (thick black ellipse), saving all points (red dots). . . . 34
- 2.3.6 Univariate and bivariate posterior density estimates for ρ and α . . . 35
- 2.3.7 Comparison of full (left) and sampled (right) bivariate posterior density estimates for ρ and α from Metropolis-Hastings. 42
- 2.3.8 Bivariate ABC posterior density estimates using a range of tolerances. 43
- 2.3.9 Comparison of Metropolis-Hastings (left) and ABC (right) bivariate posterior density estimates for ρ and α 44
- 2.4.1 Illustrative diagram of A-optimality, with $p = 2$ 47
- 2.4.2 Illustrative diagram of E-optimality, with $p = 2$ 48
- 2.4.3 Illustrative diagram of D-optimality, with $p = 2$ 49

3.2.1	Fisher information for Weibull distribution with $\theta = n = 2$, as a function of time.	71
3.2.2	Comparison of the exponential CDF (green with rate parameter 4.1827), the Gamma CDF (blue with shape parameter 1.5 and scale parameter 0.1594), and the Priceless CDF (red with $\theta' = 1.0503$), all with the same mean 0.2391.	76
3.2.3	Comparison of the exponential PDF (green with rate parameter 4.1827), the Gamma PDF (blue with shape parameter 1.5 and scale parameter 0.1594), and the Priceless PDF (red with $\theta' = 1.0503$), all with the same mean 0.2391.	77
3.2.4	Fisher information for non-separable $H(t, \theta)$ with $\theta' \approx 1.0503$, as a function of time, with maximum at $t^* \approx 0.3859$	79
4.1.1	Prior distributions for α (red) and ρ (blue).	90
4.1.2	A typical realisation of the stochastic SIS epidemic model with population size $N = 50$, five initially infectious individuals, $\beta = 4$ and $\mu = 1$	91
4.3.1	The sampled utility surface for the SIS epidemic model when making one observation, using the algorithm of Drovandi and Pettitt [2013]. The black dashed lines correspond to the candidate designs, listed in Table 4.3.2. The solid red line corresponds to the optimal design listed in Table 4.3.3.	100
4.3.2	The sampled utility surface for the SIS epidemic model when making two observations, using the algorithm of Drovandi and Pettitt [2013]. The black crosses correspond to the candidate designs, listed in Table 4.3.2. The red star corresponds to the optimal design listed in Table 4.3.3.	101

4.3.3	The sampled utility surface for the SIS epidemic model when making three observations, using the algorithm of Drovandi and Pettitt [2013]. The black crosses correspond to the candidate designs, listed in Table 4.3.2. The red star corresponds to the optimal design listed in Table 4.3.3.	102
5.2.1	A typical realisation of the stochastic death process with $b_1 = 1$ and $N = 50$	113
5.2.2	A typical realisation of the stochastic SI epidemic model with $b_1 = 0.0275$, $b_2 = 0.01125$, and $N = 50$	115
5.2.3	A typical realisation of the stochastic SIS epidemic model with $N = 50$, five initially infectious individuals, $\beta = 4$ and $\mu = 1$	117
5.3.1	Bias (a) and variance (b) of posterior distributions for b_1 in the death model, when one, two or three observations are permitted. . .	121
5.3.2	Distribution of Kullback-Leibler divergence (KLD) for 100 simulations of the Markovian death model, observed at the optimal design for one, two, three and four observations.	123
5.3.3	Bias of estimates of b_1 (a) and b_2 (b), log-variance of b_1 (c) and b_2 (d), and covariance of b_1 and b_2 (e), of the joint posterior distribution of (b_1, b_2) for the SI model.	126
5.3.4	Bias of estimates of α (a), and ρ (b), variance of α (c) and ρ (d), and covariance between estimates of α and ρ (e), of the joint posterior distribution of (α, ρ) for the SIS model.	129
6.2.1	Diagram illustrating the progression of chickens through the dose-response model.	142
6.2.2	Diagram illustrating the progression of chickens through the transmission model.	143
6.2.3	Diagram illustrating the progression of chickens through the latent model.	144

6.2.4	Comparison of the exponential ($\lambda = 1/2$), and Erlang ($k = 4, \gamma = 2$) distributions, both with mean ($\mu = 2$).	145
6.2.5	Diagram illustrating the progression of chickens through the complete model.	147
6.2.6	Prior distributions for dose-response model parameters, α and β . . .	150
6.2.7	Possible dose-response curves based on the prior distribution end points for α , and the 95% highest density interval for β . The blue-shaded region corresponds to viable dose-response curves under different combinations of α and β	151
6.2.8	The cumulative distribution function (CDF) for the Erlang(2,1) distribution, used to model the time taken to pass through the latent period.	153
6.2.9	Prior distribution for transmission parameter β_T , in the CTMC SE_kI model.	154
6.3.1	Difference in possible dose-response curves based on the prior distribution end points for α , and the 95% highest density interval end points for β	157
6.3.2	Viable dose-response curves based on the end-points of the prior distributions for α and β	161
6.3.3	Comparison of dose-response curves when observing the process at time five (blue) and time three (red), with black markers representing doses 3, 3.5, 4, 4.5, 5 and 5.5 \log_{10} CFU.	167
6.3.4	The cumulative distribution functions for each of the latent period distributions with shape k , and rate γ	172
6.4.1	The cumulative distribution function for each of the latent period distributions with shape k , and rate γ	179

Acknowledgements

First of all, I would like to thank my supervisors, Dr Jonathan Tuke, Associate Professor Joshua Ross, and Professor Nigel Bean. From encouraging me to even consider further study in the first place, to offering me a project that perfectly tied in the elements of probability and statistics that I enjoyed, to offering me boundless help and advice – with my project or otherwise – whenever needed throughout the past few years, thank you. Thank you not only for your guidance on this project, but also in making the overall experience a fun and memorable one. I might have learned some things about optimal design and maths, but I have gained so much more knowledge from each of you regarding general research practice, teaching, and overall life skills. I can honestly say I would not have gotten to this point without the support that you have each given me. I'd also like to give a brief thank you to Dr Daniel Pagendam, for his correspondence, code and discussion regarding some of his work in optimal design, which helped get me started.

Next, I would like to sincerely thank a number of people that have made my time at University memorable. First, to the office staff at the School of Mathematical Sciences. Thank you for your amazingly friendly demeanour, and help – whether it be with organising teaching, or just getting some new stationery. Thank you to Jess, Kale, Kate, Kyle, Max and Sophie for all the great times at the beginning of my time at university. I'd like to pass on a special thanks to Ben, Mingmei, Nic, and Vincent. Each of you has made what would otherwise be the day-to-day dullness of studying, not only bearable, but a barrel of laughs. Mingmei, thank you for your

every day banter, talk about football or cricket, and your help with my (often) simple questions about code. Nic, thank you for all the memorable quotes and good times. Vincent, thanks for being a *different* pure mathematician, and providing plenty of laughs and distractions. Finally, to Ben. We have, in your words, “dominated stats teaching” together for a while now. With all of the banter and jokes, I can honestly say... it would not have been the same without you.

To my family, and in particular my parents. Thank you for your support in every form, throughout not only my time at university, but the whole 24 years of my life. Finally, thank you to my amazing girlfriend, Claire. For putting up with me when I’m stressed, grumpy, tired, busy, or all of the above; you have been amazing. Thank you for all of your love and support.

Abstract

In this thesis, we investigate the optimal experimental design of some common biological experiments. The theory of optimal experimental design is a statistical tool that allows us to determine the optimal experimental protocol to gain the most information about a particular process, given constraints on resources. We focus on determining the optimal design for experiments where the underlying model is a Markov chain — a particularly useful stochastic model.

Markov chains are commonly used to represent a range of biological systems, for example: the evolution and spread of populations and disease, competition between species, and evolutionary genetics. There has been little research into the optimal experimental design of systems where the underlying process is modelled as a Markov chain, which is surprising given their suitability for representing the random behaviour of many natural processes. While the first paper to consider the optimal experimental design of a system where the underlying process was modelled as a Markov chain was published in the mid 1980's, this research area has only recently started to receive significant attention.

Current methods of evaluating the optimal experimental design within a Bayesian framework can be computationally inefficient, or infeasible. This is due to the need for many evaluations of the posterior distribution, and thus, the model likelihood — which is computationally intensive for most non-linear stochastic processes. We implement an existing method for determining the optimal Bayesian experimental design to a common epidemic model, which has not been considered in a Bayesian

framework previously. This method avoids computationally costly likelihood evaluations by implementing a likelihood-free approach to obtain the posterior distribution, known as Approximate Bayesian Computation (ABC). ABC is a class of methods which uses model simulations to estimate the posterior distribution. While this approach to optimal Bayesian experimental design has some advantages, we also note some disadvantages in its implementation.

Having noted some drawbacks associated with the current approach to optimal Bayesian experimental design, we propose a new method – called ABCdE – which is more efficient, and easier to implement. ABCdE uses ABC methods to calculate the utility of all designs in a specified region of the design space. For problems with a low-dimensional design space, it evaluates the optimal design in significantly less computation time than the existing methods. We apply ABCdE to some common epidemic models, and compare the optimal Bayesian experimental designs to those published in the literature using existing methods. We present a comparison of how well the designs – obtained from each of the different methods – performs when used for statistical inference. In each case, the optimal designs obtained via ABCdE are similar to those obtained via existing methods, and the statistical performance is indistinguishable.

The main applications we consider are concerned with group dose-response challenge experiments. A group dose-response challenge experiment is an experiment in which we expose subjects to a range of doses of an infectious agent or bacteria (or drug), and measure the number that are infected (or, the response) at each dose. These experiments are routinely used to quantify the infectivity or harmful (or safe) levels of an infectious agent or bacteria (e.g., minimum dose required to infect 50% of the population), or the efficacy of a drug. We focus particularly on the introduction of the bacteria *Campylobacter jejuni* to chickens. *C. jejuni* can be spread from animals to humans, and is the species most commonly associated with enteric (intestinal) disease in humans. By quantifying the dose-response relationship of the bacteria in chickens – via group dose-response challenge experiments – we can determine the

safe levels of bacteria in chickens with the aim to minimise, or eradicate, the risk of transmission amongst the flock, and thus, to humans. Thus, accurate estimates of the dose-response relationship are crucial – and can be obtained efficiently by considering the optimal experimental design. However, the statistical analysis of most dose-response experiments assume that the subjects are independent. Chickens engage in coprophagic activity (oral ingestion of faecal matter), and are social animals meaning they must be housed in groups. Thus, oral-faecal transmission of the bacteria may be present in these experiments, violating the independence assumption and altering the measured dose-response relationship. We use a Markov chain model to represent the dynamics of these experiments, accounting for the latency period of the bacteria, and the transmission between chickens. We determine the optimal experimental design for a range of models, and describe the relationship between different model aspects and the resulting designs.