Bayesian Experimental Design for Models with Intractable Likelihoods Using Indirect Inference

Dr Chris Drovandi
Queensland University of Technology
c.drovandi@qut.edu.au

Collaborators: Tony Pettitt, Caitriona Ryan
Acknowledgements: Australian Research Council Discovery Grant and organisers of DEMA

December 14, 2015
Bayesian Experimental Design

- Set-up: Have prior distribution $p(\theta)$ for parameter of statistical model, with likelihood $p(y|\theta, d)$.
- Wish to design for next $n$ observations. Design variable: $d = (d_1, \ldots, d_n)$.
- Define (general) utility function $u(d, y, \theta)$. $y$ is future data.

$$u(d) = E_{\theta,y}[u(d, y, \theta)] = \int_y \int_\theta u(d, y, \theta)p(y|d, \theta)p(\theta)d\theta dy,$$

- Objective

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

Too difficult to do directly
Bayesian utility functions

Parameter Estimation Design

- **Kullback-Leibler divergence** \( u(d, y) = KL(p(\theta|y, d)||p(\theta)) \). Mutual info between \( \theta \) and \( y \).
- **Concentration of Posterior distribution**
  - Set \( u(d, y) \) as entropy of \( p(\theta|y, d) \)
  - Posterior precision \( u(d, y) = 1/det(\text{var}(\theta|y, d)) \)

These utility functions assume that the likelihood function is easily computable!!!
To estimate $u(d, y)$ need to obtain posterior distribution

$$p(\theta|y, d) \propto p(y|\theta, d)p(\theta).$$

But what to do if $p(y|\theta, d)$ is intractable?

- Approximate Bayesian Computation (ABC)
- Bayesian Indirect Inference (II)
Approximate Bayesian Computation

- simulation based method that does not involve likelihood evaluations
- involves a joint ‘approximate’ posterior distribution

\[ p(\theta, x|y, \epsilon) \propto g(y|x, \epsilon)p(x|\theta)p(\theta) \]

where \( g(y|x, \epsilon) \) is a weighting function

- popular choice \( g(y|x, \epsilon) = 1(\rho(y, x) \leq \epsilon) \)
- how to choose \( \rho(y, x) \)?
  - usually based on small set of summary statistics
- choice of \( \epsilon \) trade-off between accuracy and efficiency
Rejection ABC

1. generate $\theta^i \sim p(\theta)$ for $i = 1, \ldots, N$
2. simulate $x^i \sim p(y|\theta^i, d)$ for $i = 1, \ldots, N$
3. compute discrepancies $\rho^i = \rho(y^i, x^i)$ for $i = 1, \ldots, N$, creating particles $\{\theta^i, \rho^i\}_{i=1}^N$
4. sort particle set via discrepancy $\rho$
5. calculate $\epsilon = \rho^{\lfloor \alpha N \rfloor}$ (where $\lfloor \cdot \rfloor$ denotes the floor function)

ABC posterior samples = $\{\theta^i | \rho^i \leq \epsilon\}_{i=1}^{\alpha N}$

Steps 1 and 2 are independent of data and $\{\theta^i, x^i\}_{i=1}^N$ can be stored.

Discrepancy function

$$\rho(y, x) = \sum_{i=1}^{D} \frac{|y_i - x_i|}{\text{std}_{p(\theta)}(x_i)}$$

(1)
ABC Rejection in Design (Drovandi and Pettitt 2013)

- Discretise design space (e.g. one (time) dimension $t_{min}, t_{max}, t_{inc}$)
- Store the prior simulations at each design point in the design space
- Search over discretised design space for optimal design
- For each proposed design $d^*$ can use the relevant stored prior simulations to sample from $p(\theta | y, d^*, \epsilon)$

Drawbacks:

- Must search over discretised design space
- **Storage intensive** (does not scale to increase in design dimension)
- Choosing $\epsilon$ (here we end up having different $\epsilon$ for each posterior approximation)
- Require larger $\epsilon$ for larger number of observations

We use Bayesian indirect inference to overcome these drawbacks
Assume that there exists an auxiliary model with a fully tractable likelihood \( p_a(y|\phi, d) \), so that approximately

\[
p(y|\theta, d) \propto p_a(y|\phi = g(\theta), d).
\]

Unfortunately the mapping function \( g(\theta) \) is unknown but can be estimated via simulation, \( \hat{g}(\theta) \). Then we obtain the following approximate posterior:

\[
p_a(\theta|y, d) \propto p_a(y|\phi = \hat{g}(\theta), d)p(\theta).
\]

See Smith (1993) for a classical version of this approach.
An estimate of the auxiliary parameter $\phi$ is found by

$$\hat{\phi} = \hat{g}(\theta) = \arg\max_{\phi \in \Phi} p_a(x|\phi, d_T), \quad x \sim p(x|\theta, d_T),$$

- Free to choose the training design $d_T$ (technically also a design problem!)
- To bring down variance of $\hat{\phi} = \hat{g}(\theta)$ can consider simulating $m$ independent datasets at $\theta$, $x_{1:m} = (x_1, \ldots, x_m)$.
- Do this for a collection of $n$ samples across the prior space to obtain the estimated mapping $\{\theta^i, \hat{\phi}^i\}_{i=1}^n$. Store the mapping.
Consider a dataset $y$. Can obtain an importance sampling approximation of the II posterior with weights

$$W^i \propto p_a(y|\hat{\phi}^i, d)$$

for $i = 1, \ldots, n$.

$$\{W^i, \theta^i\}_{i=1}^n$$ can be used to estimate $u(d, y)$. 
Advantages:

- Free to optimise over continuous design space
- Requires storage of the mapping \( \{ \theta^i, \hat{\phi}^i \}_{i=1}^n \) only. Should scale better with design dimension.
- No need to choose \( \epsilon \).
- Scale better with number of observations (see later)

Drawbacks:

- User must select auxiliary model. Not an obvious choice in all applications.
Once a strategy has been devised to estimate $u(d, y)$ any optimisation approach can be used to solve $\arg\max u(d)$ where $u(d) = \int_Y u(d, y)p(y|d)\,dy$.

Here we used the Muller (1999) algorithm. Involves MCMC simulation from augmented probability model with marginal $u(d)^J$ then calculation of multivariate mode (very hard!)
A large value of $J$ is required to substantially sharpen the utility surface (computationally intensive).

Consider the modified utility

$$\tilde{u}(d, y) = \max\{(u(d, y) - su_p), 0\}, \quad (2)$$

where $u_p$ is the utility calculated from the prior distribution;

$$u_p = 1/\det(\text{Var}(\theta)). \quad (3)$$

Scaling factor $s$ helps to ensure utility is positive.

Intuitively: how much utility do we obtain above and beyond the prior.

$\tilde{u}(d)$ has same mode as $u(d)$ but $\tilde{u}$ has greater curvature.
Sharpening the Utility Surface with $\tilde{u}$
Motivating Example - Macroparasite Immunity

- estimate parameters of Markov process model explaining macroparasite population development with host immunity
- 212 hosts (cats) $i = 1, \ldots, 212$. Each cat injected with $l_i$ juvenile *Brugia pahangi* larvae
- at time $t_i$ host is sacrificed and number of mature parasites recorded
- three variable problem
  - $M(t)$ matures, $L(t)$ juveniles, $I(t)$ immunity
- only $L(0)$ and $M(t_i)$ observed for each host

[Durham et al (1972)]

- Mature Parasites
  - $M(t)$
  - $\mu_M M(t)$ (Natural death)
  - $\nu L(t)$ (Gain of immunity)
- Juvenile Parasites
  - $L(t)$
  - $\gamma L(t)$ (Maturation)
  - $\nu L(t)$ (Natural death)
  - $\beta (t) L(t)$ (Death due to immunity)
- Immunity
  - $I(t)$
  - $\mu_I I(t)$ (Loss of immunity)
  - $\nu L(t)$ (Gain of immunity)

\[ \begin{align*}
M(t) &\rightarrow \text{Maturation} \rightarrow L(t) \\
M(t) &\rightarrow \text{Natural death} \rightarrow I(t) \\
L(t) &\rightarrow \text{Natural death} \rightarrow I(t) \\
L(t) &\rightarrow \text{Death due to immunity} \\
I(t) &\rightarrow \text{Gain of immunity} \\
I(t) &\rightarrow \text{Loss of immunity}
\end{align*} \]
Design Problem

How to choose the sacrifice times and initial juvenile injections to gain most information about parameters ($\nu$ and $\mu_L$)?
The data show too much variation for Binomial

A Beta-Binomial model has an extra parameter to capture dispersion

\[
p(m_i|\alpha_i, \beta_i) = \binom{l_i}{m_i} \frac{B(m_i + \alpha_i, l_i - m_i + \beta_i)}{B(\alpha_i, \beta_i)},
\]

Useful reparameterisation \( p_i = \frac{\alpha_i}{\alpha_i + \beta_i} \) and \( \xi_i = \frac{1}{\alpha_i + \beta_i} \)

Relate the proportion and over dispersion parameters to time, \( t_i \), and initial larvae, \( l_i \), covariates

\[
\logit(p_i) = f_p(t_i, l_i) = \beta_0 + \beta_1 \log(t_i) + \beta_2 \log(t_i)^2,
\]

\[
\log(\xi_i) = \eta.
\]

Four auxiliary parameters \( \phi = (\beta_0, \beta_1, \beta_2, \eta) \)
II design choices (Design just for sacrifice times)

For $d_T$ we consider 1000 randomly selected sacrifice times in (30, 300) days where the initial number of larvae is set to 100 for every host in the training design. $m = 1$ and $n = 10,000$
## Results - Sacrifice Times Only

<table>
<thead>
<tr>
<th>method</th>
<th>design (days)</th>
<th>II utility</th>
<th>ABC utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>II design</td>
<td>77.4</td>
<td>5.00</td>
<td>5.95</td>
</tr>
<tr>
<td>ABC design</td>
<td>99</td>
<td>4.97</td>
<td>6.08</td>
</tr>
<tr>
<td>II design</td>
<td>(69.7, 97.5)</td>
<td>5.38</td>
<td>7.20</td>
</tr>
<tr>
<td>ABC design</td>
<td>(71, 127)</td>
<td>5.35</td>
<td>7.18</td>
</tr>
<tr>
<td>II design</td>
<td>(56.5, 82.8, 120.3)</td>
<td>5.73</td>
<td>7.39</td>
</tr>
<tr>
<td>ABC design</td>
<td>(95, 105, 231)</td>
<td>5.54</td>
<td>7.28</td>
</tr>
<tr>
<td>II design</td>
<td>(50.5, 75.2, 98.2, 143.8)</td>
<td>6.10</td>
<td>7.28</td>
</tr>
<tr>
<td>ABC design</td>
<td>(79, 121, 231, 273)</td>
<td>5.61</td>
<td>7.07</td>
</tr>
</tbody>
</table>
Extend the auxiliary model (only one extra parameter)

\[
\begin{align*}
\text{logit}(p_i) &= f_p(t_i, l_i) = \beta_0 + \beta_1 \log(t_i) + \beta_2 \log(t_i)^2, \\
\log(\xi_i) &= f_\xi(t_i, l_i) = \eta_0 + \eta_1 l_i,
\end{align*}
\]

\(d_T\): 1000 hosts in the training design, where the sacrifice times are sampled randomly in (30, 300) days and the initial larvae counts are randomly taken from the set of integers between 100 and 200 inclusive.
Results with Initial Juveniles

One host: 69.7 days, 111 juveniles

Two hosts: 1st host 60.9 days, 108 juveniles. 2nd host 135.3 days and 105 juveniles

Validation: consider 1 host 70 days, estimated utility with ABC for 100, 150, 200 juveniles
The expected utilities for 100, 150 and 200 initial larvae are roughly $6.05 \times 10^{11}$, $5.32 \times 10^{11}$ and $4.93 \times 10^{11}$. Indeed 100 juveniles appears optimal.
Future Work

- Obtain training design $d_T$ optimally.
- Smooth out the estimated mapping to obtain more precise mapping.
- Develop a Laplace approximation of II posterior (allow for high-dimensional design for which importance sampling is not suitable).
- Other motivating applications???
Key References


- Drovandi and Pettitt (2013). Bayesian experimental design for models with intractable likelihoods. Biometrics


Publication page:
http://eprints.qut.edu.au/view/person/Drovandi,_Christopher