Subsetting Big Data and Parameter Estimation for Complex Models

Dr Christopher Drovandi

School of Mathematical Sciences
ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)
Queensland University of Technology

23 February 2018

Big Data

- Big Data: Volume, Velocity, Variety
- Standard statistical approaches often don't work

Potential Solution:

- Avoid Big Data analysis by taking 'informative' subset of data
- Data subset may still be sufficient to address analysis aims

Here we subset data using ideas from optimal experimental design. (Drovandi et al 2017 Statistical Science)

Optimal Experimental Design

In optimal design want to choose the values of controllable variables d so that we expect to achieve analysis aims.

Example: Logistic regression
$y_i \sim Binary(p_i)$ where

$$\text{logit}(p_i) = \beta_0 + \beta_1 d_i,$$

where d_i is a covariate for which we can control what values it takes.

Example optimal design problem: selecting values of $d_1, \ldots, d_n$ to learn most about β_0, β_1.

Subsetting with Optimal Design

Assume we have proposed a Model M parameterised by θ with covariates d:

- Take training sampling from big data to get initial information about θ.
- Solve optimal design problem to get d^*.
- Take observation in big data closest to d^*.
- Update information about θ
- Rinse and repeat until happy.

Example - Accelerometer Data

- Assess four different methods for classifying activities that are based on accelerometer output (e.g. walking, basketball etc). Response is whether or not the method classifies correctly.
- 12 different activities.
- 222 participants aged between 5 and 18 years.
- Each participant observed roughly 4 times (1 year apart).

Not big data but big enough...

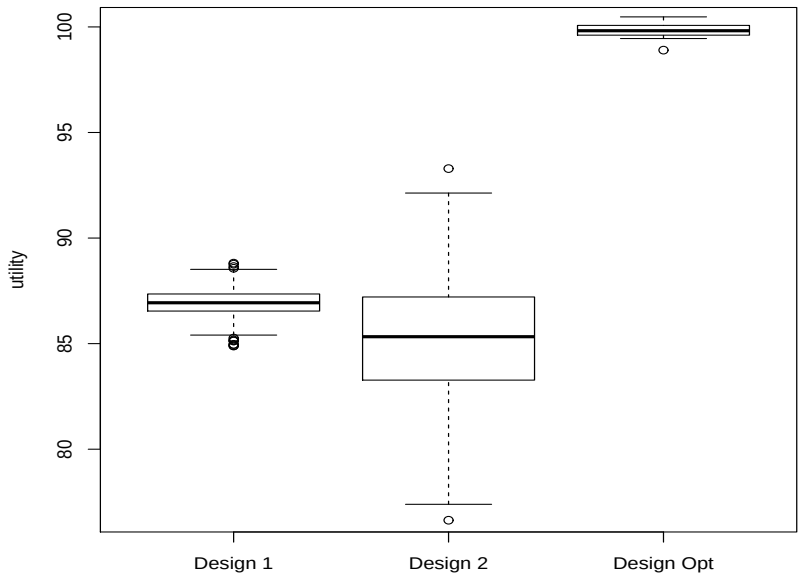
Example

The Model

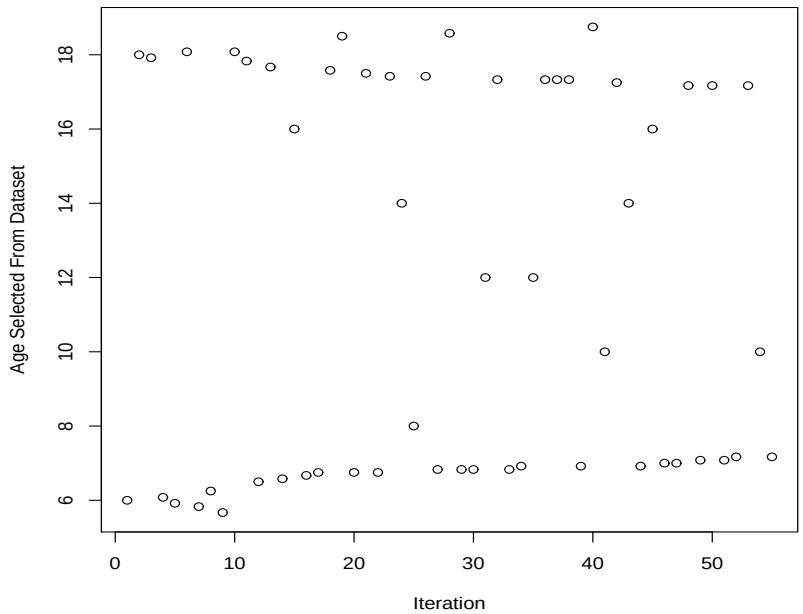
$$\text{logit}(\pi_{ti}) = \beta_0 + b_t + \beta_{\text{age}}\text{age}_{ti} + \sum_{j=1}^{3}\beta^j_{\text{cut}}\text{cut}^j_{ti} + \sum_{j=1}^{11}\beta^j_{\text{trial}}\text{trial}^j_{ti} +$$
$$\sum_{j=1}^{33}\beta^j_{\text{cut,trial}}\text{cut}^j_{ti} \times \text{trial}^j_{ti} + \sum_{j=1}^{3}\beta^j_{\text{age,cut}}\text{age}^j_{ti} \times \text{cut}^j_{ti} +$$
$$\sum_{j=1}^{11}\beta^j_{\text{age,trial}}\text{age}^j_{ti} \times \text{trial}^j_{ti},$$

where $b_t \overset{iid}{\sim} \mathcal{N}(0, \phi)$ for $t = 1, \ldots, 212$

Results

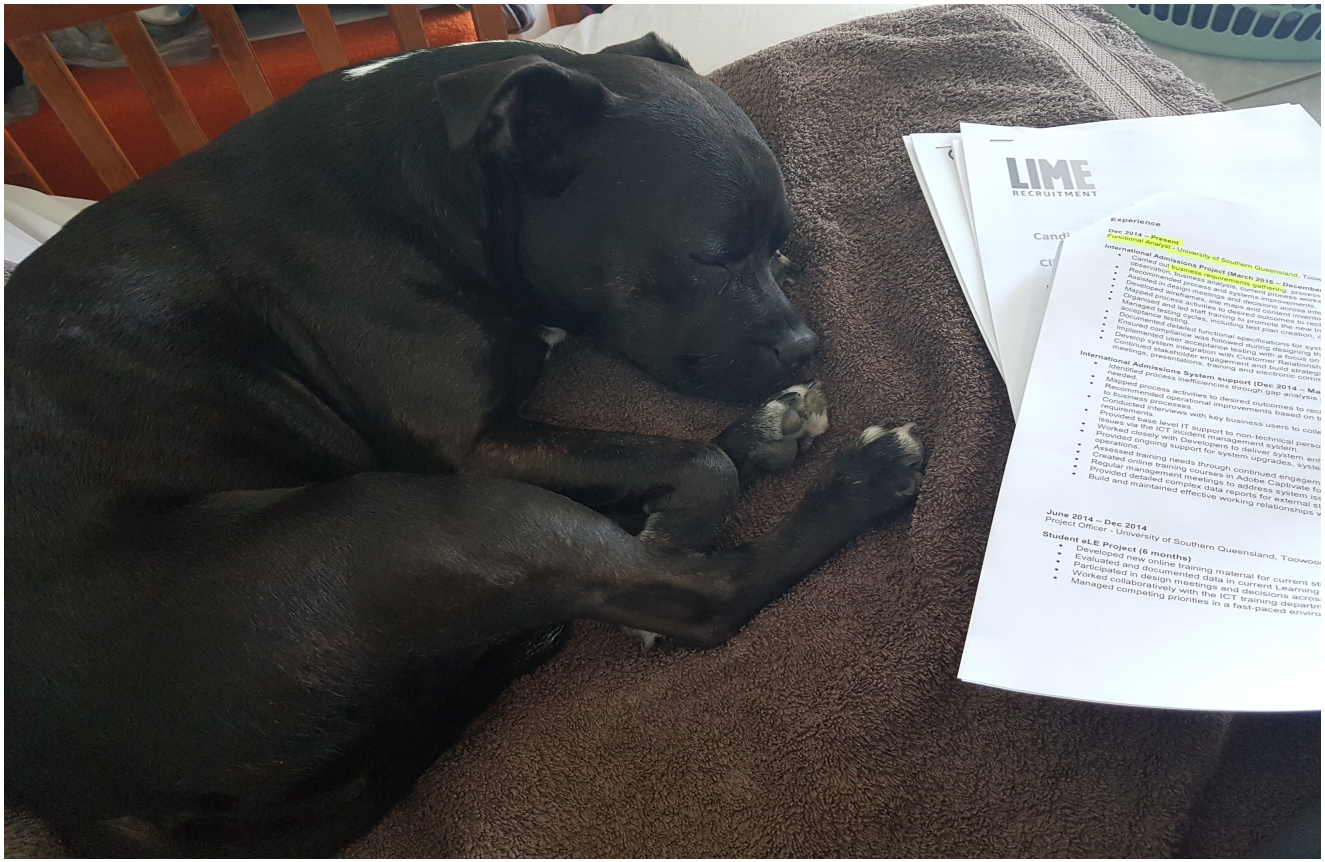


(a) distribution of observed utilities



(b) selected ages

Questions?



Simulation-based Parameter Estimation

Bayesian statistics is based on the posterior distribution $p(\theta|y)$

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

where $p(y|\theta)$ is the likelihood and $p(\theta)$ is the prior.

Typically require the likelihood $p(y|\theta)$ to be tractable.

But many models have intractable likelihoods.

However, it is often still feasible to simulate the model for a given θ.

This motivates simulation-based approaches to parameter estimation.

Approximate Bayesian Computation (ABC)

ABC is a popular simulation-based method.

Simplest implementation (ABC rejection):

- Draw θ from prior
- Simulate data x from model based on θ
- If x 'close' to y then keep θ as a sample
- Repeat until sufficient number of samples kept

Produces only 'approximate' posterior.

Cell Biology Example

Cell motility and proliferation are important parts of many biological processes (e.g. skin cancer growth, wound healing).

One way to investigate this is through a scratch assay. A ‘scratch’ is made which separates the cells. Images of the cells are taken at regular time intervals until the cells are once again in contact.

Images taken every 5 minutes for 12 hours (145 images)

Approximately map cells onto a rectangular lattice (binary matrix where a 1 indicates presence of a cell a particular location).

Cell Biology Example

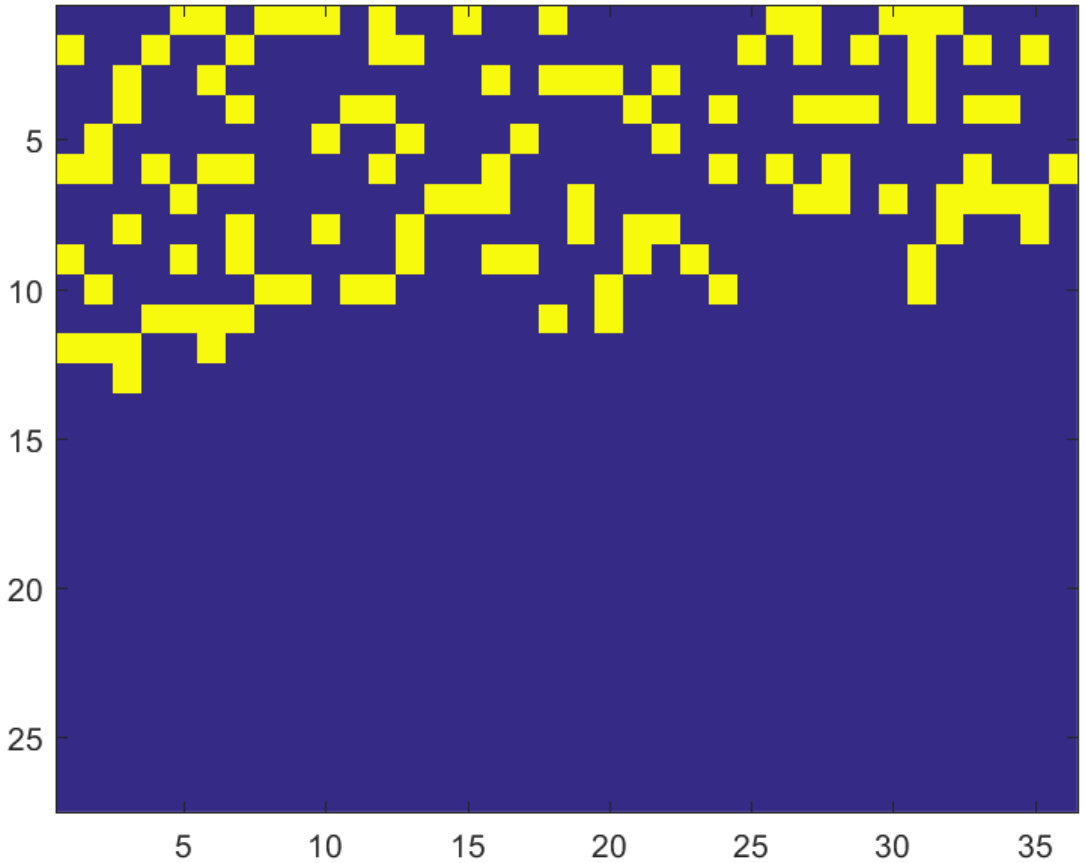
Stochastic Model (see Johnston et al 2014)

In time step τ cells given chance to move to neighbouring location with probability P_m.

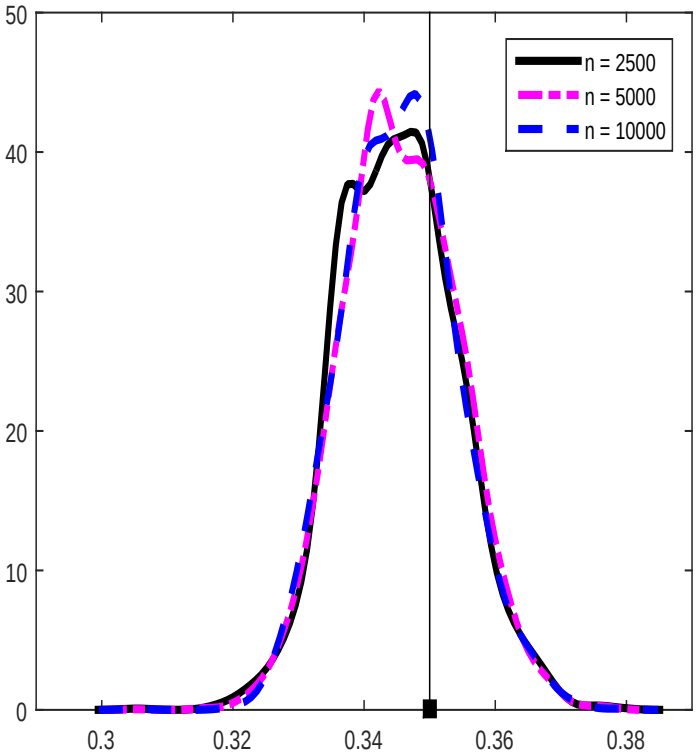
During time step cells can give 'birth' with probability P_p and place new cell at neighbouring location.

Cell Biology Example

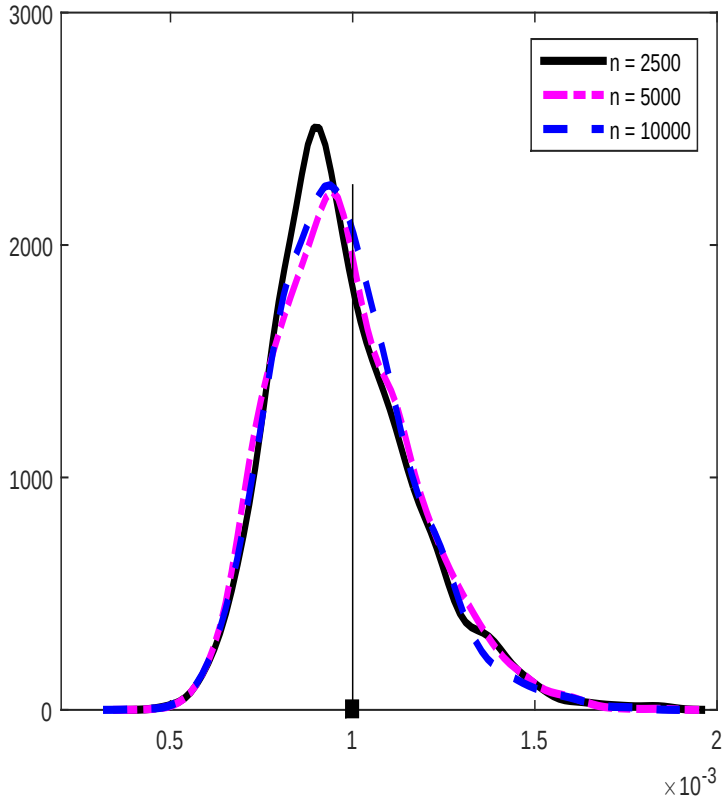
Simulated data from model with $P_m = 0.35$ and $P_p = 0.001$



Results



(c) P_m



(d) P_p

Contact Details

Email: c.drovandi@qut.edu.au

Website:
https://chrisdrovandi.weebly.com/

Twitter: @chris_drovandi