Presentation Overview

- Introduction
- Intuition
- Proposed methodology
- Example application
- Comparison to Beaumont at al.
Part I

Introduction
Notation

- Data $X_{\text{obs}}$ observed.
- Model with parameter vector $\lambda$.
- Have prior distribution for $\lambda$, density $\pi(\lambda)$.
- Aim: infer (approximate) posterior distribution $\lambda|X_{\text{obs}}$. 
Example ABC algorithm (Rejection Sampling)

1. Propose parameters $\lambda$ from prior density $\pi(\lambda)$
2. Simulate data $X_{\text{sim}}$ given $\lambda$
3. If $d(X_{\text{sim}}, X_{\text{obs}}) < \epsilon$ accept the proposal
4. Repeat

- $d(\cdot, \cdot)$ is a distance metric (e.g. Euclidean distance).
- $\epsilon > 0$ is a tuning parameter – trades approximation error against computational error.
- Accepted proposals have distribution $\lambda|d(X_{\text{sim}}, X_{\text{obs}}) < \epsilon$
- Usual to define $d$ in terms of lower-dimensional summaries $S(X)$; so $d(X_{\text{sim}}, X_{\text{obs}}) = d(S(X_{\text{sim}}), S(X_{\text{obs}}))$.
- Ideal is $S(\cdot)$ a set of sufficient statistics.
Aim of Talk

Rejection sampling ABC method is inefficient. Various alternatives have been proposed (e.g. MCMC or SMC). We instead focus on:

- How to choose summary statistics?
- How to choose distance metric $d$?
- Is the “cut-off” acceptance rule the best choice?

Implementation uses MCMC – but ideas apply to any ABC method.
Part II

Intuition
Mixture Representation of Posterio

Assume we accept simulated data $x_{\text{sim}}$ with probability $\alpha(x_{\text{sim}}, x_{\text{obs}})$. The resulting ABC posterior is

$$
\pi_{\text{ABC}}(\lambda) \propto \int \pi(\lambda)p(x_{\text{sim}}|\lambda)\alpha(x_{\text{sim}}, x_{\text{obs}})dx_{\text{sim}}
$$

$$
= \int \pi(\lambda|x_{\text{sim}})\beta(x_{\text{sim}})dx_{\text{sim}}
$$

where

$$
\beta(x_{\text{sim}}) = \frac{\pi(x_{\text{sim}})\alpha(x_{\text{sim}}, x_{\text{obs}})}{\int \pi(x_{\text{sim}})\alpha(x_{\text{sim}}, x_{\text{obs}})dx_{\text{sim}}}.
$$

So ABC posterior is a continuous mixture of true posteriors. $\beta$ is likely to be dominated by $\alpha$ (for small acceptance probabilities).
Minimising Posterior Variance

Standard result gives ABC posterior variance as

$$\text{Var}_{ABC}(\lambda) = \mathbb{E}(\text{Var}(\lambda|X_{\text{sim}})) + \text{Var}(\mathbb{E}(\lambda|X_{\text{sim}})).$$

where on the RHS mean and variance is with respect to $\beta(x_{\text{sim}})$.

This suggests it is natural to choose $\alpha(x_{\text{sim}}, x_{\text{obs}})$ to “minimise”:

$$\text{Var}(\mathbb{E}(\lambda|X_{\text{sim}})),$$

subject to some average acceptance probability.
It seems reasonable to focus on acceptance probabilities that are symmetric.

Consider the case where overall acceptance probability is small. This will correspond to accepted data being close to the observed data.

Look at “minimising” $\text{Var}(E(\lambda|X_{\text{sim}}))$ for fixed acceptance probability of rejection sampling method.
For some $p \times n$ matrix $D$, 

$$E(\lambda|X) \approx E(\lambda|X_{\text{obs}}) + D[X - X_{\text{obs}}],$$

and thus

$$\text{Var}(E(\lambda|X_{\text{sim}})) \approx E(\beta([X - X_{\text{obs}}]DD^T[X - X_{\text{obs}}]^T)).$$

To minimise the sum of the individual variances: acceptance based on

$$[X - X_{\text{obs}}]^TD^TD[X - X_{\text{obs}}] < \epsilon$$

Let $S(X) = DX$, then this is equivalent to

$$[S(X) - S(X_{\text{obs}})]^T[S(X) - S(X_{\text{obs}})] < \epsilon,$$
Intuition from Result

Ideally parameters should be uncorrelated and have similar scales. Then:

- Should have one summary statistic per parameter.
- Calculation of summary statistic requires calculation of $D$: which is output of (local?)-linear regression.
- Cut-off acceptance rule appears best.
- Diagnostics for this approach would be to test the validity of the linear model approximation.

[Note there is a big hole in any formal proof from this argument.]
Part III

Proposed Methodology / Example Application
Proposed Methodology

Overview

- Have a preliminary run of ABC to obtain a region of high posterior probability.
- Simulate parameters from this region, and data for each parameter value. Use this Linear Regression on this simulated data to generate summary statistics (one per parameter).
- Consider adding extra data – such as powers – to give a better linear fit.
- Illustrate with an example.

We have other theoretical results which support these suggestions.
Allingham et al investigate ‘quantile distributions’

Distributions defined by their inverse cdf: \( F^{-1}(x) \)

A quantile distribution may not have easily available likelihood, but can be simulated by inversion

- Simulate \( u \sim U(0, 1) \)
- Calculate \( F^{-1}(u) \)

ABC is a natural method of inference
A particular quantile distribution is the $g$-and-$k$ distribution:

$$F^{-1}(x) = A + B \left(1 + c \frac{1 - \exp(-gz(x))}{1 + \exp(-gz(x))}\right) (1 + z(x)^2)^k z(x)$$

- $z(x)$ is the $x$th quantile of the $N(0, 1)$ distribution
- $A$ and $B$ are location and scale parameters
- $g$ and $k$ parameters control skewness and kurtosis
- Final parameter $c$ is usually fixed as 0.8
- Allingham et al propose this as a flexible distribution with small number of parameters
Allingham et al applied ABC analysis to the following problem

- Illustration rather than real application
- Sample of 10,000 independent $g$-and-$k$ draws made
  - Parameters $A = 3$, $B = 1$, $g = 2$, $k = 0.5$ and $c = 0.8$
- This used as observed data
- Parameter $c$ taken as known
- Others to be estimated
- Uniform prior on region $[0, 10]^4$
Each simulated data set has 10,000 simulated values.

Allingham et al calculated the order statistics and used these as the summary statistics.

- i.e. 10,000 summary statistics (all of the data)

Analysis used:
- ABC-MCMC algorithm
- Euclidean distance metric
- Cut-off acceptance rule

We replicated this analysis and used it as a preliminary run.
Output of the preliminary run gives an approximate posterior
Used to create a training distribution for parameter values
Large set of training parameter values sampled from this
For each training parameter value
  - Data simulated from the $g$-and-$k$ distribution
  - Order statistics calculated
Regressions performed for each parameter
  - Training parameter values as responses
  - Simulated order statistics as covariates
$g$-and-$k$ Example
Construction of summary statistics - issues

- 10,000 covariates caused computational difficulties
- Pick a subset of covariates – percentiles – and do regression for these
- Also included powers of these percentiles.
- Parameters related to higher moments, so using powers as covariates is natural.
Transition density was Normal with variance matrix based on preliminary run output variance

\( \epsilon \) chosen to give acceptance rate roughly 1%

10,000 iterations performed in each run

Output thinned to reduce autocorrelation

Results based on 500 output points for each method
**g-and-k Example**

**Results**

- **Density estimates of marginal ABC output (after thinning)**
- **n.b. g poorly identified by original method**

- **Black = regression (percentile)**
- **Green = regression (full)**
- **Pink = original method**
- **Vertical lines = true parameter values**
### Regression summary statistics perform better

- Using powers of data improves performance
- Percentile case has speed improvement

| Method                      | Regression (percentile) | Regression (full) | Allingham et al |
|-----------------------------|-------------------------|-------------------|-----------------|
| Time for ABC run(s)         | 46.9                    | 5193.7            | 4807.1          |
| $\epsilon$ used            | 0.11                    | 0.14              | 13.3            |
| Acceptance rate             | 1.01%                   | 0.98%             | 0.88%           |
| A std dev                   | 0.049                   | 0.061             | 0.083           |
| B std dev                   | 0.053                   | 0.079             | 0.068           |
| g std dev                   | 0.094                   | 0.439             | 2.560           |
| k std dev                   | 0.058                   | 0.124             | 0.043           |
Part IV

Comparison with Beaumont et al.
Comparison with Beaumont et al.

- There are links with the method of Beaumont et al.
- We use linear-regression on the complete data to choose summary statistics. These then used within ABC.
- Beaumont et al. use ABC, then apply a linear regression correction to get parameter estimate.
- In applications, they assume a small number of summary statistics have been chosen. Results in Blum (2009) suggest the method performs poorly as the number of summary statistics increases.
Empirical Comparison: Toy example

- We have iid normal data $X_1, \ldots, X_p$ where $X_1, X_2, X_3$ have mean log $\lambda$; and the other data values are uninformative.
- Can calculate the true posterior analytically.
- For a range of values of $p$ we simulate 100 data sets, implement each ABC method, and calculate the increase in mean square error.
- Compare ABC, ABC with Beaumont et al. correction, and our approach.
- Implementation such that CPU cost of all methods were the same. [So higher acceptance probability in our approach.]
Results
Part V

Conclusion
Summary

- “Theoretical” results motivate a semi-automatic way of deriving summary statistics, which uses linear regression.
- Approach supported empirically for the application of Allingham et al.
- Important link with work of Beaumont et al.

Other Results
- Looked at other ways of constructing summary statistics – none work better than Linear Regression.
- Similar improvement over published work on a genetics example.
D. Allingham, R. King, and K. Mengersen. Bayesian estimation of quantile distributions *Statistics and Computing*, 19(2), 2009

Paul Marjoram, Vincent Plagnol, and Simon Tavare. Markov chain Monte Carlo without likelihoods *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003