Approximate Bayesian Computation

Dennis Prangle

Newcastle University

18th June 2018

Intractability

- Statistical inference:
  - Given some data
    (e.g. forensic evidence at a crime scene)
  - Set up random process – model – that could have produced it
  - Infer unknowns – parameters – in the model
    (e.g. identity of perpetrator)
- Standard methods (e.g. maximum likelihood, Bayes) based on probability calculations under the model
- Can be **intractable**: impossible or impractically time consuming!
- Especially for complex modern models

Intractability

- Statistical inference:
  - Given some data
    (e.g. forensic evidence at a crime scene)
  - Set up random process – model – that could have produced it
  - Infer unknowns – parameters – in the model
    (e.g. identity of perpetrator)
- Standard methods (e.g. maximum likelihood, Bayes) based on probability calculations under the model
- Can be **intractable**: impossible or impractically time consuming!
- Especially for complex modern models

ABC idea

- Often models are **generative**
- i.e. can simulate data from model given parameters
- Can be used for inference without probability calculations!
- Simulate data under many parameter values
- Accept parameters giving data “close” to observations
- Gives **approximation** to exact inference
- Main idea of approximate Bayesian computation (ABC)
- One of several **likelihood-free** methods

ABC idea

- Often models are **generative**
- i.e. can simulate data from model given parameters
- Can be used for inference without probability calculations!
- Simulate data under many parameter values
- Accept parameters giving data “close” to observations
- Gives **approximation** to exact inference
- Main idea of approximate Bayesian computation (ABC)
- One of several **likelihood-free** methods

ABC idea

- Often models are **generative**
- i.e. can simulate data from model given parameters
- Can be used for inference without probability calculations!
- Simulate data under many parameter values
- Accept parameters giving data “close” to observations
- Gives **approximation** to exact inference
- Main idea of approximate Bayesian computation (ABC)
- One of several **likelihood-free** methods

Example applications

- Population genetics
- Infectious disease epidemiology
- Systems biology / molecular dynamics
- Ecology
- Astrophysics / high energy physics
- Finance
- Agent based models
- Weather / climate

Overview of talk

- Recap of Bayesian inference
- Example of intractable likelihoods
- Introduction to ABC
- Summary statistics I
- Example analysis
- Summary statistics II
- Post-processing
- Efficient ABC algorithms
- Software
- Other likelihood-free methods
- More methodology
- Pros and cons
- References

Bayesian inference

Likelihood

- Observed data y_{obs}
- Model proposed with density $\pi(y|\theta)$
- We wish to infer parameters θ
- The **likelihood function** is $L(\theta) = \pi(y_{\text{obs}}|\theta)$
- Maximum likelihood finds θ maximising $L(\theta)$

- n.b. for discrete data use probabilities instead of densities

Likelihood

- Observed data y_{obs}
- Model proposed with density $\pi(y|\theta)$
- We wish to infer parameters θ
- The **likelihood function** is $L(\theta) = \pi(y_{\text{obs}}|\theta)$
- Maximum likelihood finds θ maximising $L(\theta)$

- n.b. for discrete data use probabilities instead of densities

Likelihood

- Observed data y_{obs}
- Model proposed with density $\pi(y|\theta)$
- We wish to infer parameters θ
- The **likelihood function** is $L(\theta) = \pi(y_{\text{obs}}|\theta)$
- Maximum likelihood finds θ maximising $L(\theta)$

- n.b. for discrete data use probabilities instead of densities

Bayesian inference

- We must specify a **prior distribution** $\pi(\theta)$
  - Beliefs about parameters before data observed
- We're interested in the **posterior distribution** $\pi(\theta|y_{\text{obs}})$
  - Beliefs updated to take data into account
- Posterior depends on prior and likelihood through **Bayes theorem**:

$$\pi(\theta|y_{\text{obs}}) = \pi(\theta)\pi(y_{\text{obs}}|\theta)/Z$$

i.e. posterior $\propto$ prior $\times$ likelihood

- where $Z = \int \pi(\theta)\pi(y_{\text{obs}}|\theta)d\theta$ (normalising constant)

Bayesian inference

- We must specify a **prior distribution** $\pi(\theta)$
  - Beliefs about parameters before data observed
- We're interested in the **posterior distribution** $\pi(\theta|y_{\text{obs}})$
  - Beliefs updated to take data into account
- Posterior depends on prior and likelihood through **Bayes theorem**:

$$\pi(\theta|y_{\text{obs}}) = \pi(\theta)\pi(y_{\text{obs}}|\theta)/Z$$

i.e. posterior $\propto$ prior $\times$ likelihood

- where $Z = \int \pi(\theta)\pi(y_{\text{obs}}|\theta)d\theta$ (normalising constant)

Bayesian inference

- We must specify a **prior distribution** $\pi(\theta)$
  - Beliefs about parameters before data observed
- We're interested in the **posterior distribution** $\pi(\theta|y_{\text{obs}})$
  - Beliefs updated to take data into account
- Posterior depends on prior and likelihood through **Bayes theorem**:

$$\pi(\theta|y_{\text{obs}}) = \pi(\theta)\pi(y_{\text{obs}}|\theta)/Z$$

$$\text{i.e. posterior} \propto \text{prior} \times \text{likelihood}$$

- where $Z = \int \pi(\theta)\pi(y_{\text{obs}}|\theta)d\theta$ (normalising constant)

Bayesian inference

- We must specify a **prior distribution** $\pi(\theta)$
  - Beliefs about parameters before data observed
- We're interested in the **posterior distribution** $\pi(\theta|y_{\text{obs}})$
  - Beliefs updated to take data into account
- Posterior depends on prior and likelihood through **Bayes theorem**:

$$\pi(\theta|y_{\text{obs}}) = \pi(\theta)\pi(y_{\text{obs}}|\theta)/Z$$
$$\text{i.e. posterior} \propto \text{prior} \times \text{likelihood}$$

- where $Z = \int \pi(\theta)\pi(y_{\text{obs}}|\theta)d\theta$ (normalising constant)

Monte Carlo

- Direct calculation of posterior generally infeasible
- Common alternative approach is **Monte Carlo**
- Monte Carlo aims to produce a sample $\theta_1, \theta_2, \ldots$ from the posterior distribution
- Can then estimate posterior quantities
  (point estimates, interval estimates, quantiles etc)
- Or produce density estimates
  (histograms, contour plots etc)

Monte Carlo methods

- Many standard Monte Carlo algorithms:
  - Rejection sampling
  - Importance sampling
  - Markov chain Monte Carlo (MCMC)
  - Sequential Monte Carlo (SMC)
- All require many evaluations of the likelihood function
- Not feasible for **intractable likelihoods** - evaluation not possible or very slow

- n.b. some Monte Carlo methods only require **unbiased estimates** of the likelihood function
- Helps with some cases of intractable likelihood

Monte Carlo methods

- Many standard Monte Carlo algorithms:
  - Rejection sampling
  - Importance sampling
  - Markov chain Monte Carlo (MCMC)
  - Sequential Monte Carlo (SMC)
- All require many evaluations of the likelihood function
- Not feasible for **intractable likelihoods** - evaluation not possible or very slow

- n.b. some Monte Carlo methods only require **unbiased estimates** of the likelihood function
- Helps with some cases of intractable likelihood

Examples of intractable likelihoods

Computer models

- Some models exist as computer simulation programs
- Equation for likelihood not available (and would be extremely complicated)
- Example: **agent based models**
- Each agent obeys simple rules, interact to form a complex system
- Applications include
  - ecology (e.g. agents represent animals)
  - systems biology (e.g. agents represent cells)
  - economics (e.g. agents represent firms)

Partial observation

- Suppose we have a tractable probability model $\pi(x, y|\theta)$ for **complete data** (x, y)
- However we only observe that $y = y_{\text{obs}}$ (i.e. **partial data**)
- So x is an unobserved latent variable
- Likelihood is

$$L(\theta) = \pi(y_{\text{obs}}|\theta) = \int \pi(x, y_{\text{obs}}|\theta) dx$$

- Integral typically intractable, especially if x high dimensional

Partial observation: examples

- Epidemic models
  - x is times of all infections/recoveries, y is final number affected
- Biochemical networks
  - x is all reaction times, y is partial measurements of one species
- Population genetics
  - x is coalescent/mutation/recombination history, y is observed sequences

ABC

Likelihood-free inference

- General idea:
    - Simulate data y from various parameter values θ
    - Consider closest matches of y to y_{obs}
    - Use corresponding parameters for inference
- Can be implemented in many different ways
- Many approaches suggested in various fields over last 40+ years
- ABC puts this idea into a Bayesian framework

Likelihood-free inference timeline

1970s Various applications *Hoel and Mitchell, Ross etc*

1984 **Inference for implicit models** *Diggle and Gratton*

1984 Bayesian inference by simulating data *Rubin*

1989 **Simulated method of moments** *McFadden* (Econometrics)

1992 **GLUE** *Beven and Binley* (Hydrology)

1993 **Indirect inference** *Gourieroux et al* (Econometrics)

1997 **ABC** *Tavaré et al/Fu and Li* (Population genetics)

2005 **Convolution filter** *Rossi and Vila*

2006 **Iterated filtering** *Ionides et al*

2010 **Synthetic likelihood** *Wood*

. . . and many more!

ABC algorithm - simplest version

Input: observed data y_{obs}, **threshold** $h \geq 0$

```
For $i = 1, 2, \ldots, N$:
    1  Sample parameter vector $\theta_i$ from prior $\pi(\theta)$
    2  Simulate data from $\pi(y|\theta_i)$
    3  If $d(y, y_{\text{obs}}) \leq h$ accept $\theta_i$
```

where $d(y, y_{\text{obs}})$ is a distance function e.g. Euclidean

Output: accepted θ_i values

This is a **rejection sampling** algorithm

ABC algorithm - simplest version

Input: observed data y_{obs}, **threshold** $h \geq 0$

For $i = 1, 2, \ldots, N$:

1. Sample parameter vector θ_i from prior $\pi(\theta)$
2. Simulate data from $\pi(y|\theta_i)$
3. If $d(y, y_{\text{obs}}) \leq h$ accept θ_i

where $d(y, y_{\text{obs}})$ is a distance function e.g. Euclidean

Output: accepted θ_i values

This is a **rejection sampling** algorithm

ABC target distribution

- Consider a proposed (θ, y) pair
- Sampled from $\pi(\theta)\pi(y|\theta) = \pi(\theta, y)$

- Acceptance is conditional on $y \approx y_{\text{obs}}$
- So accepted pairs drawn from $\pi(\theta, y|y \approx y_{\text{obs}})$
- And θ from $\pi(\theta|y \approx y_{\text{obs}})$
- An approx to exact posterior $\pi(\theta|y = y_{\text{obs}})$

- Taking $h = 0$ only accepts when $y = y_{\text{obs}}$
- Samples from exact posterior, but typically not practical (acceptances impossible/rare)

ABC target distribution

- Consider a proposed (θ, y) pair
- Sampled from $\pi(\theta)\pi(y|\theta) = \pi(\theta, y)$

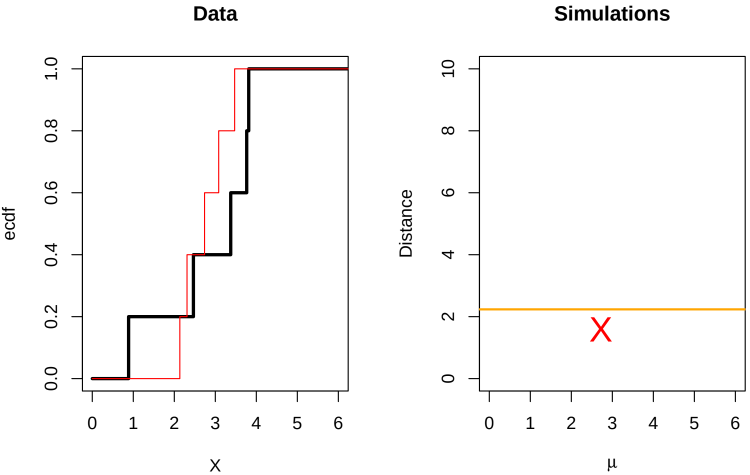
- Acceptance is conditional on $y \approx y_{\text{obs}}$
- So accepted pairs drawn from $\pi(\theta, y|y \approx y_{\text{obs}})$
- And θ from $\pi(\theta|y \approx y_{\text{obs}})$
- An approx to exact posterior $\pi(\theta|y = y_{\text{obs}})$

- Taking $h = 0$ only accepts when $y = y_{\text{obs}}$
- Samples from exact posterior, but typically not practical (acceptances impossible/rare)

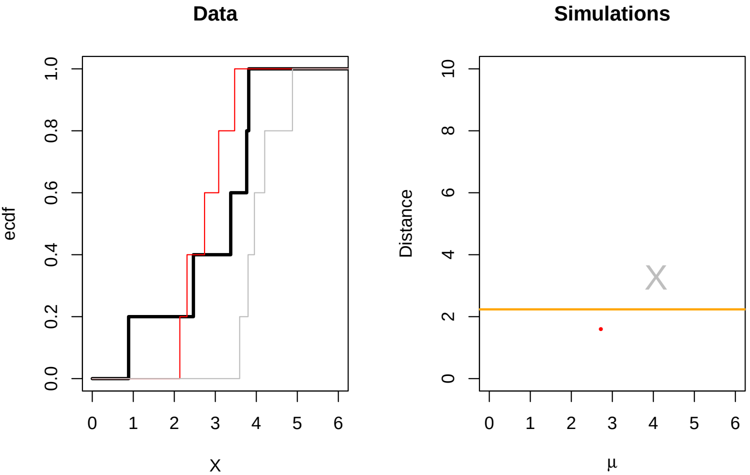
ABC algorithm example

- Model: 5 draws from $N(\mu, 1)$
- Data is ordered draws: $y_1 \leq y_2 \leq \ldots \leq y_5$
- Prior: $\text{Uniform}(0, 6)$

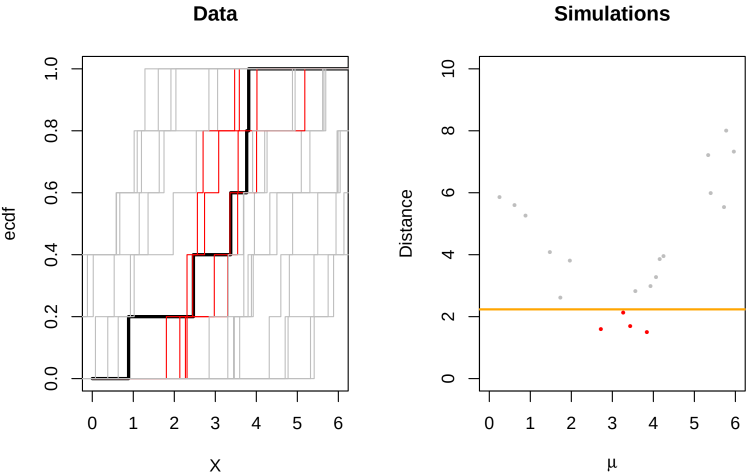
ABC algorithm example



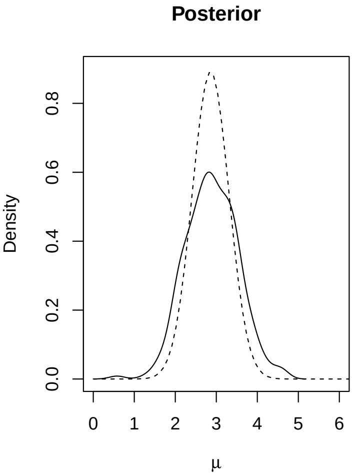
ABC algorithm example

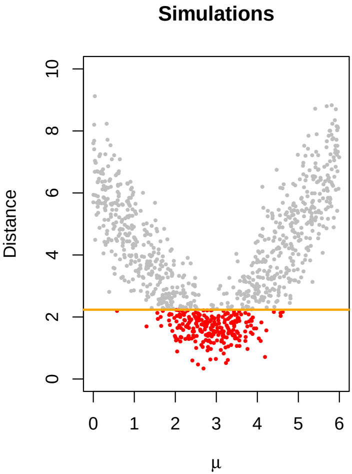


ABC algorithm example

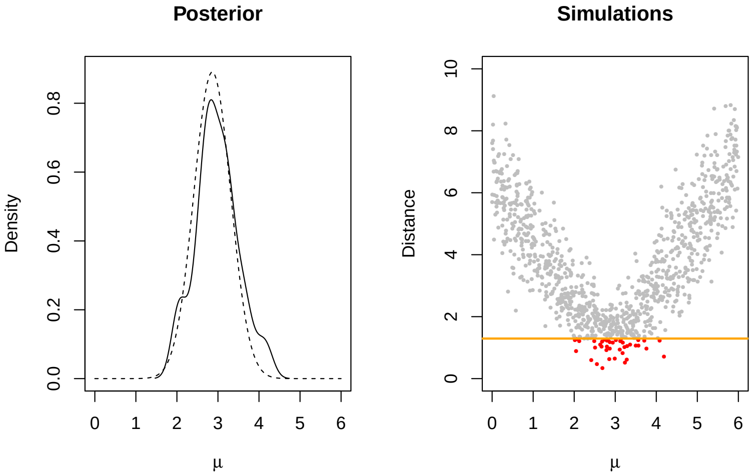


ABC algorithm example

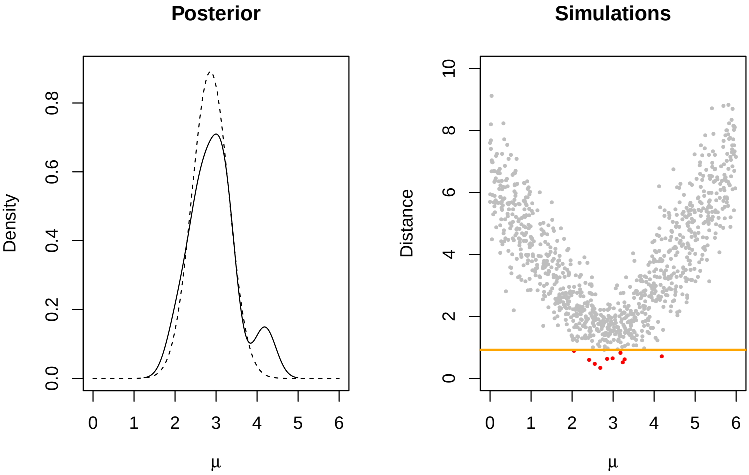




ABC algorithm example



ABC algorithm example



Tuning ABC

- ABC has several tuning choices
- e.g. threshold h and distance function d
- Affect quality of approximate results
- Some key choices discussed next

ABC algorithm - effect of h

- h too large
  - Some accepted simulations are far from observations
  - Distribution of accepted θs poor approx of posterior
- h too small
  - Distribution of accepted θs better approx
  - But too few acceptances to learn distribution well!
- So choosing h is a trade-off between two source of error

Choice of h in practice

- Often rather than choose h in advance, the number k of desired acceptances is specified (e.g. 200)
- Then h is chosen to achieve k acceptances
- So chosen **after** distances computed
- This seems a good pragmatic approach

- Also some asymptotic theory available
- Blum (2010) “Approximate Bayesian Computation: A Nonparametric Perspective” (on h)
- Biau et al (2014) “New insights into Approximate Bayesian Computation” (on k)

Choice of h in practice

- Often rather than choose h in advance, the number k of desired acceptances is specified (e.g. 200)
- Then h is chosen to achieve k acceptances
- So chosen **after** distances computed
- This seems a good pragmatic approach

- Also some asymptotic theory available
- Blum (2010) "Approximate Bayesian Computation: A Nonparametric Perspective" (on h)
- Biau et al (2014) "New insights into Approximate Bayesian Computation" (on k)

Choice of d in practice

- Euclidean distance often used i.e.

$$d(a,b) = \left[\sum_i (a_i - b_i)^2\right]^{1/2}$$

  (where i indexes data components)

- Not sensible if data on widely different scales
- A popular alternative is weighted Euclidean distance,

$$d(a,b) = \left[\sum_i \left(\tfrac{a_i - b_i}{\sigma_i}\right)^2\right]^{1/2}$$

- Here σ_i could be standard deviation of ith data component samples
- Many other distances possible - impact on results modest in general

Choice of d for repeated observations

- Special case: data is IID (e.g. repeated time series)
- Recent work has found good ABC distance measures here:
  - Kernel MMD (Park et al 2016)
  - Wasserstein distance (Bernton et al 2017)
  - Kullback-Leibler divergence (Jiang et al 2018)

Summary statistics

Summary statistics in ABC

- Earlier algorithm accepts when $d(y, y_{\text{obs}}) \leq h$
- In practice the data is usually reduced to a vector of **summary statistics** $s = S(y)$
- Acceptance occurs when $d(s, s_{\text{obs}}) \leq h$ (where $s_{\text{obs}} = S(y_{\text{obs}})$)
- Clearly necessary for non-numeric data such as genetic sequences
- Also turns out to be necessary more generally

Need for summary statistics

- Beaumont et al (2002), reviewing early work on ABC:
  "A crucial limitation of the. . . method is that only a small number of summary statistics can usually be handled. Otherwise, either acceptance rates become prohibitively low or the tolerance. . . must be increased, which can distort the approximation."
- Crucial that only a small number of summaries used
- Recognised from earliest work on ABC

Curse of dimensionality

- Quote is about a **curse of dimensionality** problem in ABC
- Informal statement:
  - More summary statistics means more opportunities for mismatches between $S(y)$ and $S(y_{\text{obs}})$
  - So distances $d(S(y), S(y_{\text{obs}}))$ typically larger
  - Need large h, which causes approximation error
- Formal statement:
  - ABC converges to correct posterior as $h \to 0$, $N \to \infty$
  - Asymptotic rate of convergence worsens with $\dim y$
  - Proved (at least partially) for most varieties of ABC

Curse of dimensionality

- Quote is about a **curse of dimensionality** problem in ABC
- Informal statement:
  - More summary statistics means more opportunities for mismatches between $S(y)$ and $S(y_{\text{obs}})$
  - So distances $d(S(y), S(y_{\text{obs}}))$ typically larger
  - Need large h, which causes approximation error
- Formal statement:
  - ABC converges to correct posterior as $h \to 0$, $N \to \infty$
  - Asymptotic rate of convergence worsens with $\dim y$
  - Proved (at least partially) for most varieties of ABC

Curse of dimensionality example

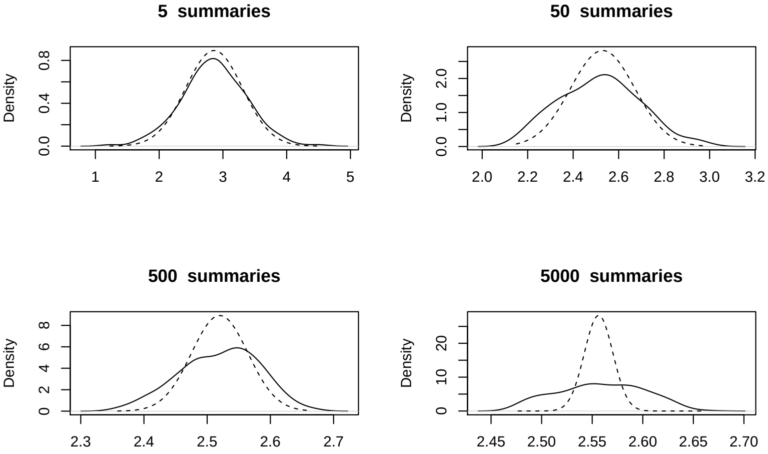
- Same example as before but with higher dimensional data

- Model: d draws from $N(\mu, 1)$
- Data is ordered draws: $y_1 \leq y_2 \leq \ldots \leq y_d$
- Uniform prior on $[0, 6]$

- $N = 10^4$ ABC iterations
- $k = 200$ acceptances

- Density estimate of ABC output compared to true posterior

Curse of dimensionality example



Need for summary statistics

- Dimension of data usually high so must be replaced with **low-dimensional summaries**
- Now ABC approximates $\pi(\theta|s_{\text{obs}})$
- i.e. posterior conditional on observed summaries
- We want this to be similar to $\pi(\theta|y_{\text{obs}})$
- So we want **informative summaries** about θ
- How to meet both requirements?

Sufficient statistics

- **Sufficient statistics** satisfy $\pi(\theta|s_{\text{obs}}) = \pi(\theta|y_{\text{obs}})$
- Low dimensional sufficient statistics would be ideal for ABC
- However they essentially only exist for exponential family models
- Very few intractable likelihood models are in this class
- So generally we must use **insufficient statistics** and accept some loss of information

Example analysis

Application

- We analyse the “**human**” dataset from the `abc` R package
- Population genetic data from Italy population (16 individuals)
- A coalescent model investigated with a population bottleneck. Parameters are:
  - Ne: Effective population size
  - a: Intensity of bottleneck (ratio of pop size before and during)
  - duration: Bottleneck duration
  - start: Start of bottleneck

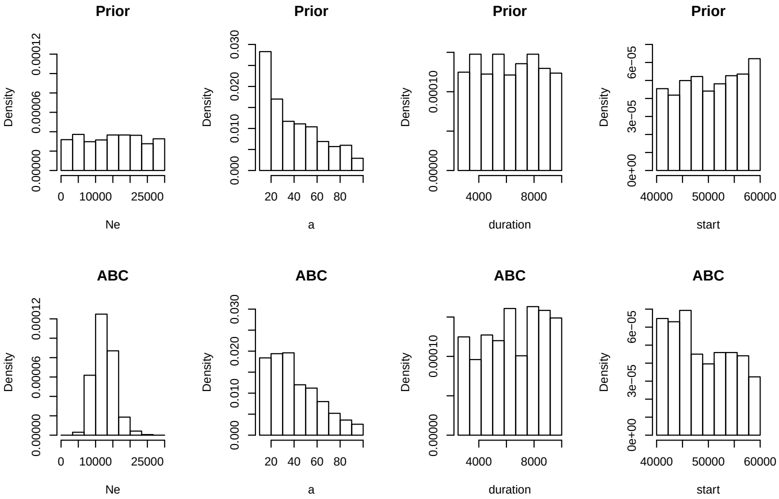
- Data is genetic sequences from various regions of genome
- 3 summary statistics used, believed to be informative about demographic history
  - Average nucleotide diversity, $\bar{\pi}$
  - Mean of Tajima's D
  - Variance of Tajima's D

ABC setup

- $N = 50,000$ ABC iterations
- $k = 500$ acceptances (1%)
- Weighted Euclidean distance used

```
library(abc)
data(human)
## Initialise data
sumstats = subset(stat.3pops.sim, models=="bott")
params = par.italy.sim
sumstats.obs = stat.voight[2,]
## Tuning choice
mytol = 500/nrow(sumstats) # i.e. 500 acceptances
## Do ABC
abc.out = abc(target=sumstats.obs, param=params,
              sumstat=sumstats, tol=mytol,
              method="rejection")
```

ABC output histograms



Choosing summary statistics

Recap

- We typically have high dimensional data y
- Want summary statistics $S(y)$ which are:

  (1) **Low dimensional**

  (2) **Informative about** θ
- How to choose these?

Choosing summary statistics

- One option is to make a subjective choice
- More automatic methods have been proposed:
  - **Subset selection**
    Find best subset of many *candidate summaries*. e.g. run ABC for each subset on test datasets and minimise error.
  - **Projection**
    Find projections of many *data features* $z(y)$ which are informative about θ. e.g. fit $\theta \sim N(Az(y), \Sigma)$ or use machine learning
  - **Auxiliary model**
    Use a tractable *auxiliary model*. e.g. let $S(y)$ be its MLEs.
- These methods typically do better than subjective choice
- But no obvious best method
- Lots of user input still required in choosing/tuning/testing method

Choosing summary statistics

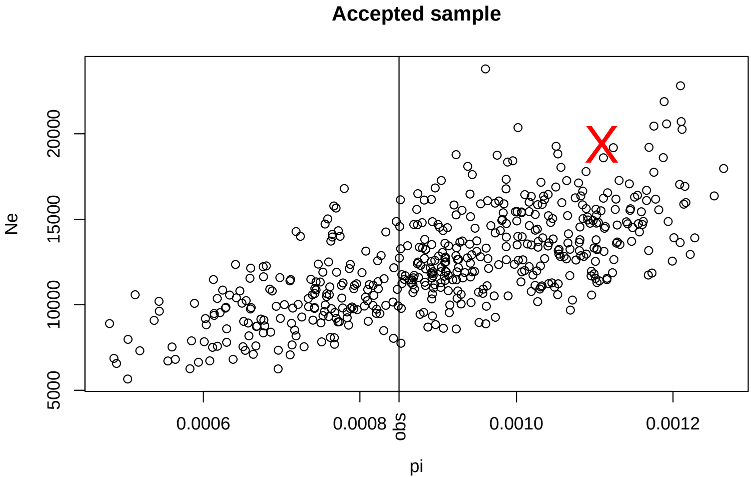
- One option is to make a subjective choice
- More automatic methods have been proposed:
  - **Subset selection**
    Find best subset of many *candidate summaries*. e.g. run ABC for each subset on test datasets and minimise error.
  - **Projection**
    Find projections of many *data features* $z(y)$ which are informative about θ. e.g. fit $\theta \sim N(Az(y), \Sigma)$ or use machine learning
  - **Auxiliary model**
    Use a tractable *auxiliary model*. e.g. let $S(y)$ be its MLEs.
- These methods typically do better than subjective choice
- But no obvious best method
- Lots of user input still required in choosing/tuning/testing method

Post-processing

Idea

- ABC accepts a sample $\theta_1, \theta_2, \ldots, \theta_k$
- The associated summary statistics are $s_1, s_2, \ldots, s_k$
- Can we **correct** θ_i to take account of the difference between s_i and s_{obs}?

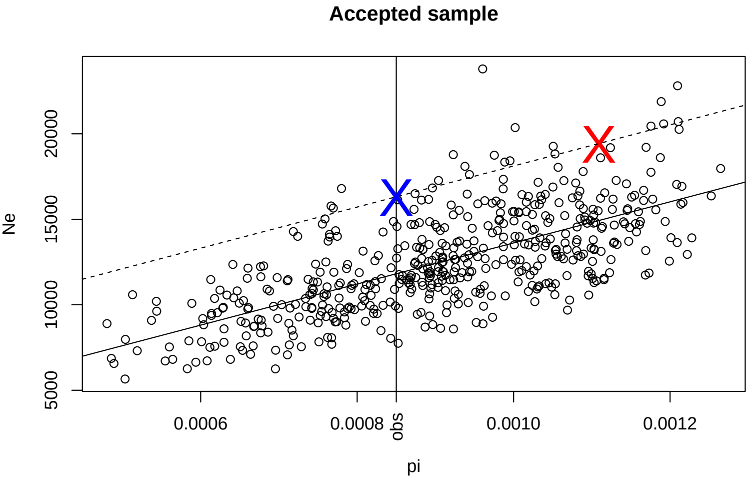
Illustration: human dataset



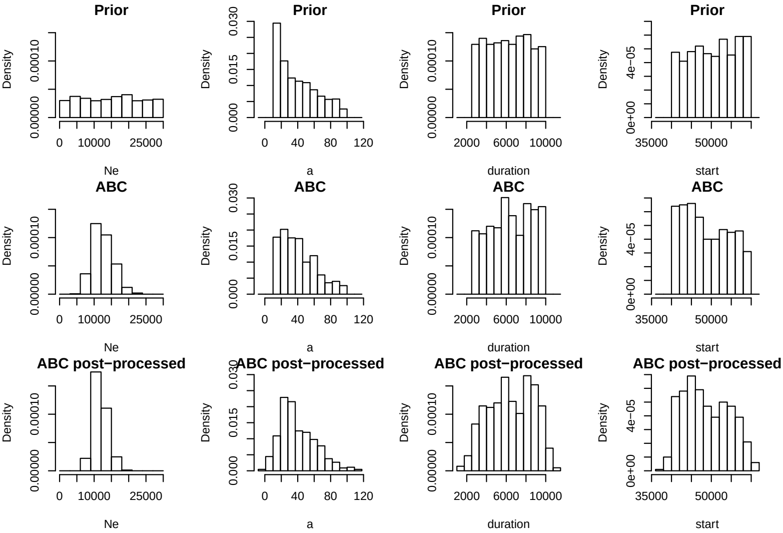
Approach

- Fit a model to the accepted (θ_i, s_i) pairs
- e.g. regression $\theta \sim N(As + b, \Sigma)$ (Beaumont et al 2002)
- So $E(\theta|s) = As + b$
- Now correct θ_i to $\theta_i - E(\theta|s_i) + E(\theta|s_{\text{obs}})$
- i.e. to $\theta_i + A(s_{\text{obs}} - s_i)$

Illustration: human dataset



Example: human dataset



Variations

- Regression correction can be applied to output of most ABC algorithms
- Methods using more flexible regression models have been proposed
- e.g. heteroskedastic regression, neural networks (Blum and François 2010)
- Similar ideas for ABC model choice (Beaumont et al 2008)

Usefulness

- Originally hoped post-processing would reduce curse of dimensionality
- However it's been proved that the rate of convergence issue is essentially unchanged (Blum 2010)
- Nonetheless post-processing sometimes improves results greatly in practice (Blum et al 2013)
- Li and Fearnhead (2018) give some asymptotic support for use in case of large data
- But Frazier et al (2018) argue it's poor under misspecified models

More efficient ABC algorithms

Inefficiency of rejection ABC

- Always samples θ from prior distribution
- Typically posterior is much more concentrated than prior
- Therefore most simulations will be very poor
- We'd like to propose better θ values
- Ideally learn good proposal distribution during algorithm

ABC importance sampling

Input: y_{obs}, $h \geq 0$, $d(\cdot, \cdot)$, **importance density** $g(\theta)$
For $i = 1, 2, \ldots, N$:

1. Sample parameter vector θ_i from $g(\theta)$
2. Simulate data from $\pi(y|\theta_i)$
3. If $d(y, y_{\text{obs}}) \leq h$ accept θ_i with weight $w_i = \pi(\theta)/g(\theta)$

Output: accepted (θ_i, w_i) pairs

Monte Carlo inference now possible as for standard importance sampling
Issue: need a sensible choice of g

ABC-SMC (very rough overview)

- Run standard ABC with threshold h_1
- Use output to choose importance density $g(\theta)$
- Run ABC importance sampling with threshold $h_2 < h_1$
- And so on

- Adaptively learns $g(\theta)$
- Various schemes along these lines
- (Technical point: some are population Monte Carlo methods and others are true SMC algorithms)

ABC-MCMC

0. Initialise some θ_0 and simulate corresponding y. Let $t = 0$.
1. Propose θ' from $q(\cdot|\theta_t)$ e.g. $\theta' \sim N(\theta_t, \Sigma)$
2. Simulate y' from model conditional on θ'
3. Calculate $\alpha = \min\left[1, \frac{\pi(\theta')q(\theta'|\theta_t)\mathbb{1}[d(s(y'),s_{\text{obs}})\leq h]}{\pi(\theta_t)q(\theta_t|\theta')\mathbb{1}[d(s(y),s_{\text{obs}})\leq h]}\right]$
4. With probability α accept

   Acceptance: Let $\theta_{t+1} = \theta'$ and $y = y'$
   Rejection: Let $\theta_{t+1} = \theta_t$ and leave y unchanged
5. Increment t and return to 1

- Discard initial results as burn-in
- Remainder can be used as Monte Carlo sample

Comparison of ABC algorithms

- All sample from **same** approximate distribution given h etc.
- So can choose based on efficiency/convenience
- ABC rejection
  - Least efficient
  - But simple to implement, esp parallelisation
  - Simulations can be reused for other analyses e.g. checking performance on simulated datasets
  - Model choice version easy
- ABC-SMC
  - Allows adaptation of h
  - Complicated to code
  - Some parallelisation possible
  - Lots of tuning choices
  - Model choice version easy

Comparison of ABC algorithms, continued

- ABC-MCMC
  - Must fix h in advance (some research on varying it)
  - Parallelisation not possible
  - Several tuning choices
  - Model choice version tricky
  - Some convergence theory exists
    (which shows alternative "one-hit" MCMC kernel is more efficient)

Software for ABC

Software options

- python: “ABCPy” and “PyABC” (many algorithms, very up to date)
- python: “ABCSysBio” (sequential ABC algorithm, systems biology)
- R: “abc” package (rejection sampling)
- R: “easyABC” package (multiple algorithms)
- standalone: “DIY-ABC” (rejection sampling, population genetics)
- standalone: “pop-ABC” (rejection sampling, population genetics)
- . . .
- Or code your own!
    - Especially feasible for rejection ABC as algorithm v simple

Alternatives to ABC

Other likelihood-free methods

- Indirect inference
  - Tries to find θ minimising average distance $d(s, s_{\text{obs}})$
  - Can be viewed as maximum likelihood analogue of ABC
- Synthetic likelihood
- Likelihood-free expectation propagation
- History matching
- Conditional density estimation (e.g. via random forests or deep learning)
- Bayesian optimisation of estimated likelihoods
- Likelihood ratio estimation

Conclusion

Summary

- ABC is a likelihood-free method for inference
- Useful for generative models with intractable likelihood
- Idea is to find θ values that produce $S(y) \approx S(y_{\text{obs}})$
- These approximate posterior

- Informative low-dimensional summaries crucial for method to work well
- Lots of algorithms/methodology exists to improve the method
- Research in more general likelihood-free methods very active!

Strengths

- The **only** way to do inference in some situations!
- Lots of freedom in what model can be used
    - Just need to be able to simulate data in reasonable time
- Simplicity
    - Simplest ABC algorithm very easy to understand/implement

Weaknesses

- Tuning requirements
  - Acceptance threshold, summary statistics, algorithm specific choices. . .
- Results are approximate
  - And it's hard to quantify how approximate
- Computationally expensive
  - Since very large number of simulations often required
- Only possible for a small number of parameters (up to around 10)
  - ABC curse of dimensionality limits number of summary statistics
  - Identifiability generally requires at least one summary statistic for each parameter

References

Review papers

- Beaumont et al “Approximate Bayesian computation in evolution and ecology” (2010)
- Bertorelle et al “ABC as a flexible framework to estimate demography over space and time: some cons, many pros” (2010)
- Csillery et al “Approximate Bayesian computation (ABC) in practice” (2010)
- Marin et al “Approximate Bayesian computational methods” (2011)
- Sunnaker et al “Approximate Bayesian computation” (2013) (basis of the ABC wikipedia page!)
- Baragatti and Pudlo “An overview on approximate Bayesian computation” (2014)
- Lintusaari et al “Fundamentals and recent developments in approximate Bayesian computation” (2017)
- Handbook of ABC (2018)

Bonus material!

ABC for model choice

Bayesian model choice

- Suppose there are several proposed models for the data
- i.e. $M_1, M_2, \dots, M_k$
- We'll usually consider $k = 2$ or 3
- Each has a pdf $\pi(y|\theta, M_i)$ and a prior $\pi(\theta|M_i)$
- n.b. θ may represent **different** set of parameters for each model
- We also have prior model weights $\pi(M_1), \pi(M_2), \dots$

- We want posterior models weights $\pi(M_1|y_{\text{obs}}), \dots$
- Maybe also parameter estimates $\pi(\theta|y_{\text{obs}}, M_1), \dots$
- All based on Bayes theorem

Difficulties with Bayesian model choice

- **Computational**

  Likelihood-based calculation of posterior model weights notoriously hard in sufficiently complicated problems

  Motivates methods like reversible-jump MCMC etc

- **Robustness**

  Results can be very sensitive to details of prior distributions

  Sensitivity analysis required

- **Interpretation**

  Depends on whether we assume one model really is correct

  Or that we search for the best approximation

- However still useful in practice!

ABC model choice

- Some or all models may have intractable likelihoods
- Human dataset example:
  - 3 coalescent models compared representing different demographic histories

  1. Model 1: bottleneck
  2. Model 2: constant population
  3. Model 3: exponential population growth

  - Each model has different number of parameters
- ABC algorithms can be adapted to this problem

ABC model choice: rejection sampling

```
Input: $y_{\text{obs}}, h, d(\cdot, \cdot), S(\cdot)$

For $i = 1, 2, \ldots, N$:
    1 Sample model $m_i$ from model prior
    2 Sample parameter vector $\theta_i$ from prior $\pi(\theta | m_i)$
    3 Simulate data from $\pi(y | \theta_i, m_i)$
    4 If $d(S(y), s_{\text{obs}}) \leq h$ accept $(m_i, \theta_i)$
Output: accepted $(m_i, \theta_i)$ values
```

Estimate posterior weight of model M_i by its frequency in output

Example: human dataset

- Equal prior model weights
- $N = 150,000$ ABC iterations
- $k = 500$ acceptances (0.3%)
- Results for 3 sets of observed data: Hausa (Cameroon), Chinese and Italian (each 16 individuals)
- Same ABC simulations used for each analysis
- Output proportions:

Population	Bottleneck	Constant	Exp growth
Hausa	0.012	0.288	0.702
Chinese	0.776	0.226	0.000
Italian	0.966	0.036	0.000

- Especially important to do sensitivity analyses etc. See abc package vignette for worked details.

Example: human dataset

- Equal prior model weights
- $N = 150,000$ ABC iterations
- $k = 500$ acceptances (0.3%)
- Results for 3 sets of observed data: Hausa (Cameroon), Chinese and Italian (each 16 individuals)
- Same ABC simulations used for each analysis
- Output proportions:

Population	Bottleneck	Constant	Exp growth
Hausa	0.012	0.288	0.702
Chinese	0.776	0.226	0.000
Italian	0.966	0.036	0.000

- Especially important to do sensitivity analyses etc. See `abc` package vignette for worked details.

Summary statistics for ABC model choice

- Can't simply use good parameter inference summaries
- Example: suppose $x_1, x_2, \ldots, x_n$ iid $N(\mu, 1)$

  Then sample mean good to infer μ

  But sample variance needed for model comparison
- Various summary statistic selection methods can be generalised to model choice

ABC model choice and classification

- ABC model choice similar to classification
- We simulate data-label pairs (y_i, m_i)
- Aim is to infer label for a further point y_{obs}
- Many statistics/machine learning methods for this
- Could be used to choose ABC summary statistics (Prangle et al 2013)
- Or to replace ABC entirely (Pudlo et al 2014 advocate random forest classifiers instead)

Weighted distances

Why weight distances

- So far ABC simulations are accepted or rejected
- Discards some information about distance of accepted simulations
- Instead we can **weight** close matches higher
- We introduce a **ABC kernel** $K(x)$
- Then $K([s - s_{\text{obs}}]/h)$ maps summaries s to a weight
- h affects the scale, acting as a “bandwidth”
- Examples:
  - Uniform kernel: $K(x) = \begin{cases} 1 & \text{for } x^T x \leq 1 \\ 0 & \text{otherwise} \end{cases}$
  - Gaussian kernel: $K(x) = \exp(-x^T x)$
- So uniform kernel gives an accept/reject algorithm
- (n.b. can easily include weight terms in kernels)

ABC algorithm with kernel

Input: y_{obs}, $h \geq 0$, $S(\cdot)$, **ABC kernel** $K(x)$.

For $i = 1, 2, \ldots, N$:

1. Sample parameter vector θ_i from prior $\pi(\theta)$
2. Simulate data from $\pi(y|\theta_i)$
3. Let $w_i = K([S(y) - s_{\text{obs}}]/h)$

Output: accepted (θ_i, w_i) pairs

Use for Monte Carlo as in importance sampling

n.b. distance $d(\cdot, \cdot)$ no longer used – $K(\cdot)$ performs similar role.

Kernels in other ABC algorithms

- Same idea can be used in all the algorithms described earlier
- Can help ABC-MCMC/ABC-SMC work well

- Alternatively can modify algorithms to **accept** with probability w_i

- Little theory on best choice of kernel
- Practice suggests it's not as important to results as summary statistics

Effect of kernel

- Assume K is a pdf
- Then ABC samples from posterior for a **misspecified model**

$$\text{summary statistics} \sim \text{model of interest} + hz$$

- where z is an independent draw from $K(\cdot)$
- See “Approximate Bayesian computation (ABC) gives exact results under the assumption of model error” - Wilkinson 2013
- Can occasionally allow exact inference if model can be expressed in this form
- Or can be used to capture effect of model misspecification

Sequential ABC analysis

Motivation

- Time series setting
- We have data at every time point
- If model has a helpful structure we can do inference **sequentially**
- i.e. analyse data at $t = 1$ by ABC, then data at $t = 2$, ...
- Each step has low dimensional data! So low approximation error

- Can be done by an ABC version of **particle filtering**
- For a review see Jasra (2014) "Approximate Bayesian computation for a class of time series models"

Challenges

- Sequential ABC is promising
- Algorithms have been developed with good theoretical properties
- But still challenging in practice
- Can be very computationally demanding
- Prone to getting “stuck” at outlying observations
- Tuning well is difficult
- Data at each time point could be complicated, requiring introduction of summaries