Tutorial on Approximate Bayesian Computation

Michael Gutmann

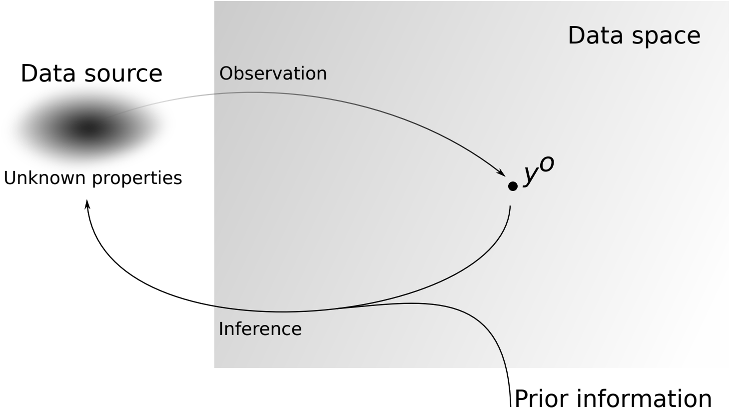
http://homepages.inf.ed.ac.uk/mgutmann

School of Informatics, University of Edinburgh

24th June 2018

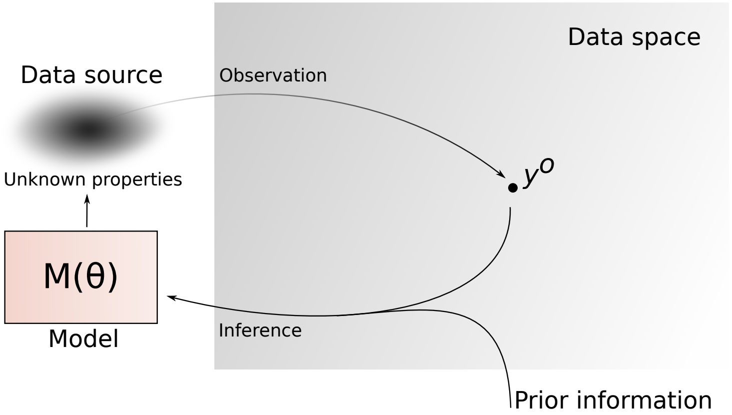
General problem considered

- Given data $\mathbf{y}^o$, draw conclusions about properties of its source
- If available, possibly take prior information into account



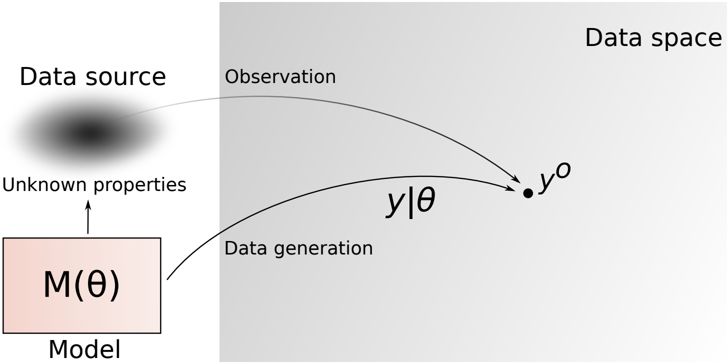
Model-based approach

- Set up a model with potential properties $\boldsymbol{\theta}$ (parameters)
- See which $\boldsymbol{\theta}$ are reasonable given the observed data



Likelihood function

- Measures agreement between $\boldsymbol{\theta}$ and the observed data $\mathbf{y}^o$
- Probability to generate data $\mathbf{y}$ like $\mathbf{y}^o$ if property $\boldsymbol{\theta}$ holds



Likelihood function

- For discrete random variables:

$$L(\boldsymbol{\theta}) = \Pr(\mathbf{y} = \mathbf{y}^o | \boldsymbol{\theta}) \tag{1}$$

- For continuous random variables:

$$L(\boldsymbol{\theta}) = \lim_{\epsilon \to 0} \frac{\Pr(\mathbf{y} \in B_\epsilon(\mathbf{y}^o) | \boldsymbol{\theta})}{\mathrm{Vol}(B_\epsilon(\mathbf{y}^o))} \tag{2}$$

Performing statistical inference

- If $L(\boldsymbol{\theta})$ is known, the inference problem becomes an optimisation or sampling problem
- Maximum likelihood estimation

$$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

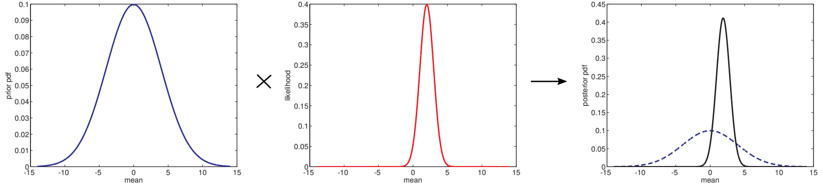
- Bayesian inference

$$p(\boldsymbol{\theta}|\mathbf{y}^o) \propto p(\boldsymbol{\theta}) \times L(\boldsymbol{\theta})$$
$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

possibly follwed by sampling or optimisation

Textbook case

- model $\equiv$ family of probability density/mass functions $p(\mathbf{y}|\boldsymbol{\theta})$
- Likelihood function $L(\boldsymbol{\theta}) = p(\mathbf{y}^o|\boldsymbol{\theta})$
- In simple cases, closed form expressions for the posterior possible



prior likelihood posterior

Simulator-based models

- Not all models are specified as family of pdfs $p(\mathbf{y}|\boldsymbol{\theta})$.
- Here: simulator-based models:

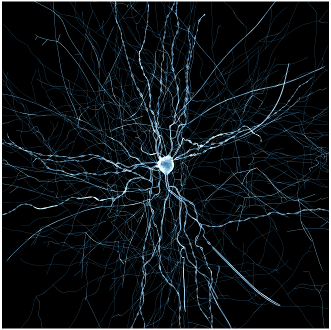
  models that are specified by a (stochastic) mechanism for generating data

Other names for simulator-based models

- Models specified via a data generating mechanism occur in multiple and diverse scientific fields.
- Different communities use different names for simulator-based models:
  - Generative models
  - Implicit models
  - Stochastic simulation models
  - Generative (latent-variable) models
  - Probabilistic programs

Examples

- Evolutionary biology:
  Simulating evolution
- Neuroscience:
  Simulating neural circuits
- Astrophysics:
  Simulating the formation of galaxies, stars, or planets
- Health science:
  Simulating the spread of an infectious disease
- Computer vision:
  Simulating facial expressions
- ...



Simulated neural activity in rat somatosensory cortex
(Figure from `https://bbp.epfl.ch/nmc-portal`)

Strengths of simulator-based models

- Direct implementation of hypotheses of how the observed data were generated.
- Modelling by replicating the mechanisms of nature that produced the observed/measured data. ("analysis by synthesis")
- Neat interface with scientific models (e.g. from physics or biology).
- Possibility to emulate real-world experiments on the computer

Weaknesses of simulator-based models

- Generally elude analytical treatment.
- May easily be made more complicated than necessary.

Weaknesses of simulator-based models

- Generally elude analytical treatment.
- May easily be made more complicated than necessary.
- Statistical inference (parameter learning) is difficult

Weaknesses of simulator-based models

- Generally elude analytical treatment.
- May easily be made more complicated than necessary.
- Statistical inference (parameter learning) is difficult

  Main reason: *Likelihood function is too expensive to evaluate*

Implicit definition of the likelihood function

- To compute the likelihood function, we needed to compute the probability that the simulator generates data close to $\mathbf{y}^o$,

$$\Pr(\mathbf{y} = \mathbf{y}^o | \boldsymbol{\theta}) \quad \text{or} \quad \Pr(\mathbf{y} \in B_\epsilon(\mathbf{y}^o) | \boldsymbol{\theta})$$

- Typically no analytical expression available.
- But we can empirically test whether simulated data equals $\mathbf{y}^o$ or is in $B_\epsilon(\mathbf{y}^o)$.
- This property will be exploited to perform inference for simulator-based models.

Exact inference for discrete random variables

- For discrete random variables, sampling from the exact posterior is possible without having to evaluate the likelihood function.
- Two equivalent perspectives:

  (1) via conditioning

  (2) via rejection sampling

Exact inference for discrete random variables

Conditioning perspective:

- By definition, the posterior is obtained by conditioning $p(\boldsymbol{\theta}, \mathbf{y})$ on the event $\mathbf{y} = \mathbf{y}^o$:

$$p(\boldsymbol{\theta}|\mathbf{y}^o) = \frac{p(\boldsymbol{\theta}, \mathbf{y}^o)}{p(\mathbf{y}^o)} = \frac{p(\boldsymbol{\theta}, \mathbf{y} = \mathbf{y}^o)}{p(\mathbf{y} = \mathbf{y}^o)} \tag{3}$$

- Can be used for sampling from the posterior without evaluating the likelihood function.

Exact inference for discrete random variables

Conditioning perspective:

- By definition, the posterior is obtained by conditioning $p(\boldsymbol{\theta}, \mathbf{y})$ on the event $\mathbf{y} = \mathbf{y}^o$:

$$p(\boldsymbol{\theta}|\mathbf{y}^o) = \frac{p(\boldsymbol{\theta}, \mathbf{y}^o)}{p(\mathbf{y}^o)} = \frac{p(\boldsymbol{\theta}, \mathbf{y} = \mathbf{y}^o)}{p(\mathbf{y} = \mathbf{y}^o)} \tag{3}$$

- Can be used for sampling from the posterior without evaluating the likelihood function.
- Generate tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i) \sim p(\boldsymbol{\theta}, \mathbf{y})$:
  - $\boldsymbol{\theta}_i \sim p_{\boldsymbol{\theta}}$ (iid from the prior)
  - $\mathbf{y}_i \sim p(\mathbf{y}|\boldsymbol{\theta}_i)$ (run the simulator with param $\boldsymbol{\theta}_i$)

Exact inference for discrete random variables

Conditioning perspective:

- By definition, the posterior is obtained by conditioning $p(\boldsymbol{\theta}, \mathbf{y})$ on the event $\mathbf{y} = \mathbf{y}^o$:

$$p(\boldsymbol{\theta}|\mathbf{y}^o) = \frac{p(\boldsymbol{\theta}, \mathbf{y}^o)}{p(\mathbf{y}^o)} = \frac{p(\boldsymbol{\theta}, \mathbf{y} = \mathbf{y}^o)}{p(\mathbf{y} = \mathbf{y}^o)} \tag{3}$$

- Can be used for sampling from the posterior without evaluating the likelihood function.
- Generate tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i) \sim p(\boldsymbol{\theta}, \mathbf{y})$:
  - $\boldsymbol{\theta}_i \sim p_{\boldsymbol{\theta}}$ (iid from the prior)
  - $\mathbf{y}_i \sim p(\mathbf{y}|\boldsymbol{\theta}_i)$ (run the simulator with param $\boldsymbol{\theta}_i$)
- Condition on $\mathbf{y} = \mathbf{y}^o \Leftrightarrow$ retain the tuples where $\mathbf{y}_i = \mathbf{y}^o$

Exact inference for discrete random variables

Conditioning perspective:

- By definition, the posterior is obtained by conditioning $p(\boldsymbol{\theta}, \mathbf{y})$ on the event $\mathbf{y} = \mathbf{y}^o$:

$$p(\boldsymbol{\theta}|\mathbf{y}^o) = \frac{p(\boldsymbol{\theta}, \mathbf{y}^o)}{p(\mathbf{y}^o)} = \frac{p(\boldsymbol{\theta}, \mathbf{y} = \mathbf{y}^o)}{p(\mathbf{y} = \mathbf{y}^o)} \tag{3}$$

- Can be used for sampling from the posterior without evaluating the likelihood function.
- Generate tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i) \sim p(\boldsymbol{\theta}, \mathbf{y})$:
  - $\boldsymbol{\theta}_i \sim p_{\boldsymbol{\theta}}$ (iid from the prior)
  - $\mathbf{y}_i \sim p(\mathbf{y}|\boldsymbol{\theta}_i)$ (run the simulator with param $\boldsymbol{\theta}_i$)
- Condition on $\mathbf{y} = \mathbf{y}^o \Leftrightarrow$ retain the tuples where $\mathbf{y}_i = \mathbf{y}^o$
- The $\boldsymbol{\theta}_i$ of the retained tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i)$ are samples from the posterior $p(\boldsymbol{\theta}|\mathbf{y}^o)$.

Exact inference for discrete random variables

Rejection sampling perspective:

- If you retain (accept) the samples $\boldsymbol{\theta}_i \sim p_{\boldsymbol{\theta}}$ with probability

$$L(\boldsymbol{\theta}_i)/\max L(\boldsymbol{\theta}),$$

  the retained samples follow a distribution proportional to $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})L(\boldsymbol{\theta})$.

Exact inference for discrete random variables

Rejection sampling perspective:

- If you retain (accept) the samples $\boldsymbol{\theta}_i \sim p_{\boldsymbol{\theta}}$ with probability

$$L(\boldsymbol{\theta}_i)/\max L(\boldsymbol{\theta}),$$

  the retained samples follow a distribution proportional to $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})L(\boldsymbol{\theta})$.

Exact inference for discrete random variables

Rejection sampling perspective:

- If you retain (accept) the samples $\boldsymbol{\theta}_i \sim p_{\boldsymbol{\theta}}$ with probability

$$L(\boldsymbol{\theta}_i)/\max L(\boldsymbol{\theta}),$$

  the retained samples follow a distribution proportional to $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})L(\boldsymbol{\theta})$.
- Key point: since $L(\boldsymbol{\theta}_i) = \Pr(\mathbf{y} = \mathbf{y}^o|\boldsymbol{\theta}_i)$ we can implement the accept/reject step by
  - drawing $\mathbf{y}_i \sim p(\mathbf{y}|\boldsymbol{\theta}_i)$
  - checking whether $\mathbf{y}_i = \mathbf{y}^o$.

Exact inference for discrete random variables

Rejection sampling perspective:

- If you retain (accept) the samples $\boldsymbol{\theta}_i \sim p_{\boldsymbol{\theta}}$ with probability

$$L(\boldsymbol{\theta}_i) / \max L(\boldsymbol{\theta}),$$

  the retained samples follow a distribution proportional to $p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) L(\boldsymbol{\theta})$.
- Key point: since $L(\boldsymbol{\theta}_i) = \Pr(\mathbf{y} = \mathbf{y}^o | \boldsymbol{\theta}_i)$ we can implement the accept/reject step by
  - drawing $\mathbf{y}_i \sim p(\mathbf{y}|\boldsymbol{\theta}_i)$
  - checking whether $\mathbf{y}_i = \mathbf{y}^o$.
- Allows us to sample from the posterior without evaluating the likelihood function.

Limitations

- Only applicable to discrete random variables.
- And even for discrete random variables:
  Computationally not feasible in higher dimensions
- Reason: *The probability of the event $\mathbf{y}_{\boldsymbol{\theta}} = \mathbf{y}^o$ becomes smaller and smaller as the dimension of the data increases.*
- Only a small fraction of the simulated tuples will be accepted.
  - The small number of accepted samples do not represent the posterior well.
  - Large Monte Carlo errors

Approximations to make inference feasible

- Settle for approximate yet computationally feasible inference.
- Introduce two types of approximations:

1. Instead of working with the whole data, work with lower dimensional summary statistics $\mathbf{t}_{\boldsymbol{\theta}}$ and $\mathbf{t}^o$,

$$\mathbf{t}_{\boldsymbol{\theta}} = T(\mathbf{y}_{\boldsymbol{\theta}}) \qquad \mathbf{t}^o = T(\mathbf{y}^o). \tag{4}$$

2. Instead of requiring $\mathbf{t}_{\boldsymbol{\theta}} = \mathbf{t}^o$, require that $\Delta_{\boldsymbol{\theta}} = d(\mathbf{t}^o, \mathbf{t}_{\boldsymbol{\theta}})$ is less than ϵ. (d may or may not be a metric)

Approximation of the likelihood function

Likelihood function:

$$L(\boldsymbol{\theta}) = \lim_{\epsilon \to 0} L_\epsilon(\boldsymbol{\theta}) \qquad\qquad L_\epsilon(\boldsymbol{\theta}) = \frac{\Pr(\mathbf{y} \in B_\epsilon(\mathbf{y}^o)|\boldsymbol{\theta})}{\mathrm{Vol}(B_\epsilon(\mathbf{y}^o))}$$

The two approximations are equivalent to:

1. Replacing $\Pr\left(\mathbf{y} \in B_\epsilon(\mathbf{y}^o) \mid \boldsymbol{\theta}\right)$ with $\Pr\left(\Delta_{\boldsymbol{\theta}} \leq \epsilon \mid \boldsymbol{\theta}\right)$
2. Not taking the limit $\epsilon \to 0$

Approximation of the likelihood function

Likelihood function:

$$L(\boldsymbol{\theta}) = \lim_{\epsilon \to 0} L_\epsilon(\boldsymbol{\theta}) \qquad L_\epsilon(\boldsymbol{\theta}) = \frac{\Pr(\mathbf{y} \in B_\epsilon(\mathbf{y}^o)|\boldsymbol{\theta})}{\mathrm{Vol}(B_\epsilon(\mathbf{y}^o))}$$

The two approximations are equivalent to:

1. Replacing $\Pr\left(\mathbf{y} \in B_\epsilon(\mathbf{y}^o) \mid \boldsymbol{\theta}\right)$ with $\Pr\left(\Delta_{\boldsymbol{\theta}} \leq \epsilon | \, \boldsymbol{\theta}\right)$
2. Not taking the limit $\epsilon \to 0$

They define an approximate/surrogate likelihood function $\tilde{L}_\epsilon(\boldsymbol{\theta})$

$$\tilde{L}_\epsilon(\boldsymbol{\theta}) \propto \Pr\left(\Delta_{\boldsymbol{\theta}} \leq \epsilon | \, \boldsymbol{\theta}\right)$$

Rejection ABC algorithm

- The two approximations yield the rejection algorithm for approximate Bayesian computation (ABC).
- Do N times:
  1. $\boldsymbol{\theta}_i \sim p_{\boldsymbol{\theta}}$ (iid from the prior)
  2. $\mathbf{y}_i \sim p(\mathbf{y}|\boldsymbol{\theta}_i)$ (run the simulator with param $\boldsymbol{\theta}_i$)
  3. Compute the discrepancy $\Delta_i = d(T(\mathbf{y}^o), T(\mathbf{y}_i))$

  Retain the $\boldsymbol{\theta}_i$ with $\Delta_i \leq \epsilon$
- This is *the* basic ABC algorithm.

Properties

- Rejection ABC algorithm produces samples $\boldsymbol{\theta} \sim \tilde{p}_\epsilon(\boldsymbol{\theta}|\mathbf{y}^o)$,

$$\tilde{p}_\epsilon(\boldsymbol{\theta}|\mathbf{y}^o) \propto p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\tilde{L}_\epsilon(\boldsymbol{\theta})$$

- Inference is approximate due to
  - the summary statistics T and distance d
  - $\epsilon > 0$
  - the finite number of samples (Monte Carlo error)

Some current research themes in ABC

- Broad classification into research on
  (1) statistical efficiency, (2) computational efficiency, (3) theoretical analysis of the algorithms

Some current research themes in ABC

- Broad classification into research on
  (1) statistical efficiency, (2) computational efficiency, (3) theoretical analysis of the algorithms
- Choice of summary statistics, distance and threshold

Some current research themes in ABC

- Broad classification into research on (1) statistical efficiency, (2) computational efficiency, (3) theoretical analysis of the algorithms
- Choice of summary statistics, distance and threshold
- Using the tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i)$ to

Some current research themes in ABC

- Broad classification into research on (1) statistical efficiency, (2) computational efficiency, (3) theoretical analysis of the algorithms
- Choice of summary statistics, distance and threshold
- Using the tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i)$ to
  - model the summary statistics, discrepancy, or data generative process conditional on $\boldsymbol{\theta}$ (model for the likelihood function)

Some current research themes in ABC

- Broad classification into research on (1) statistical efficiency, (2) computational efficiency, (3) theoretical analysis of the algorithms
- Choice of summary statistics, distance and threshold
- Using the tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i)$ to
  - model the summary statistics, discrepancy, or data generative process conditional on $\boldsymbol{\theta}$ (model for the likelihood function)
  - model $\boldsymbol{\theta}|\mathbf{y}$ (model for the posterior)

Some current research themes in ABC

- Broad classification into research on (1) statistical efficiency, (2) computational efficiency, (3) theoretical analysis of the algorithms
- Choice of summary statistics, distance and threshold
- Using the tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i)$ to
  - model the summary statistics, discrepancy, or data generative process conditional on $\boldsymbol{\theta}$ (model for the likelihood function)
  - model $\boldsymbol{\theta}|\mathbf{y}$ (model for the posterior)
- Generation of tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i)$ that are suitable for the above (e.g. not sampling $\boldsymbol{\theta}_i$ from the prior but a proposal distribution)

Some current research themes in ABC

- Broad classification into research on
  (1) statistical efficiency, (2) computational efficiency, (3) theoretical analysis of the algorithms
- Choice of summary statistics, distance and threshold
- Using the tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i)$ to
  - model the summary statistics, discrepancy, or data generative process conditional on $\boldsymbol{\theta}$ (model for the likelihood function)
  - model $\boldsymbol{\theta}|\mathbf{y}$ (model for the posterior)
- Generation of tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i)$ that are suitable for the above (e.g. not sampling $\boldsymbol{\theta}_i$ from the prior but a proposal distribution)
- Theoretical analysis of the nature of the approximations

Some current research themes in ABC

- Broad classification into research on (1) statistical efficiency, (2) computational efficiency, (3) theoretical analysis of the algorithms
- Choice of summary statistics, distance and threshold
- Using the tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i)$ to
  - model the summary statistics, discrepancy, or data generative process conditional on $\boldsymbol{\theta}$ (model for the likelihood function)
  - model $\boldsymbol{\theta}|\mathbf{y}$ (model for the posterior)
- Generation of tuples $(\boldsymbol{\theta}_i, \mathbf{y}_i)$ that are suitable for the above (e.g. not sampling $\boldsymbol{\theta}_i$ from the prior but a proposal distribution)
- Theoretical analysis of the nature of the approximations
- Applications to solve inference problems!

References

Longer tutorial:

https://michaelgutmann.github.io/assets/slides/Gutmann-2016-05-16.pdf

Review papers:

- Lintusaari et al. Fundamentals and Recent Developments in Approximate Bayesian Computation. *Systematic Biology* 2017.
- Sunnaker et al. Approximate Bayesian Computation. *PLoS Computational Biology*, 2013.
- Marin et al. Approximate Bayesian computational methods. *Statistics and Computing*, 2012.
- Hartig et al. Statistical inference for stochastic simulation models – theory and application.*Ecology Letters*, 2011.
- Beaumont. Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, 2010.