Approximate Bayesian Computation (Sunnåker et al., 2013)

CSC 696H1 Fall 2022 - Paper Presentation
(Moyeen Uddin)

Outline

1. Background
2. Approximate Bayesian Computation (ABC) Definition
3. Criticism against ABC
4. Some Remedies and counter arguments
    a. Basically, this paper refers to other papers that experiments with these remedies.
5. General guidelines as to how to use ABC
    a. Eg: what to keep in mind
6. Conclusion

Bayes Formula

$$p(\theta \mid D) = \frac{p(\theta)\; p(D \mid \theta)}{p(D)}$$

Approximation of an Integration by Sampling

1. Plain Monte-Carlo
2. Metropolis-Hasting MC
3. Etc.

- ABC also employs a similar kind of sampling technique.
- It works with the likelihood function (ie: approximates it)

Motivation

1. Sometimes, computing the likelihood function is expensive (even infeasible in some cases)
2. Sometimes, the analytical formula for the Likelihood function is elusive (eg: no closed form)

Benefit:

1. Bypass the computation of Likelihood Function. (eg: approximates the Likelihood function by holding some assumptions)
2. Widens the realms of models for statistical inference

Analysis of complex problems in:

Biological sciences (e.g., in population genetics, ecology, epidemiology, and systems biology).

However, worsens the challenges of parameter estimation and model selection.

Implicit vs. Explicit Models

Typically we know, both, the **prior** and **likelihood** of the joint,

$$p(\theta, \mathcal{D}) = p(\theta)p(\mathcal{D} \mid \theta)$$

- We call this an **explicit model**
- An **implicit model** lacks a closed-form joint
- Models are usually implicit because we don't know the likelihood

Two common reasons for implicit likelihood:

1) Need to integrate nuisance variables,

Can address this with standard inference

$$p(\mathcal{D} \mid \theta) = \int p(\theta, \eta)p(\mathcal{D} \mid \eta, \theta)\, d\eta$$

2) Likelihood is based on simulation **Topic of this paper**

A Basic Monte Carlo Rejection Sampler

A1: Generate $\theta \sim p(\theta)$ from prior

A2: Accept θ with probability $h = p(\mathcal{D} \mid \theta)$

A3: Return to A1

- It's trivial to show that this has the correct *target distribution,*

$$\theta \sim p(\theta \mid \mathcal{D})$$

- Special case of a Rejection Sampler with proposal $\theta \sim p(\theta)$

- In general, find an upper bound $c \geq p(\mathcal{D} \mid \theta)$ and accept with prob. h/c

What are some issues with this sampler?

Likelihood-Free Monte Carlo

B1: Generate $\theta \sim p(\theta)$ from prior

B2: Simulate $\mathcal{D}'$ from model with input θ

B3: Accept θ if $\mathcal{D}' = \mathcal{D}$; Return to B1

- Unlike rejection sampler, never need to evaluate likelihood
- Probability of acceptance is proportional to $p(\mathcal{D})$
- Prohibitively low acceptance for high-dimensional data
- **Idea** Make acceptance criteria weaker… accept within some distance:

$$\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon$$

Likelihood-Free Monte Carlo

B1: Generate $\theta \sim p(\theta)$ from prior

B2: Simulate $\mathcal{D}'$ from model with input θ

B3: Accept θ if $\mathcal{D}' = \mathcal{D}$; Return to B1

- Unlike rejection sampler, never need to evaluate likelihood
- Probability of acceptance is proportional to $p(\mathcal{D})$
- Prohibitively low acceptance for high-dimensional data
- **Idea** Make acceptance criteria weaker… accept within some distance:

$$\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon$$

Likelihood-Free Monte Carlo

C1: Generate $\theta \sim p(\theta)$ from prior

C2: Simulate $\mathcal{D}'$ from model with input θ

C3: Calculate distance $\rho(\mathcal{D}', \mathcal{D})$

C4: Accept θ if $\rho(\mathcal{D}', \mathcal{D}) \leq \epsilon$; Return to C1

- Will have higher acceptance than Algorithm B
- Target distribution is approximation of true posterior,

$$p(\theta \mid \rho(\mathcal{D}, \mathcal{D}') \leq \epsilon) \approx p(\theta \mid \mathcal{D})$$

- This still won't work in high-dimensional data…too many rejections
- **Idea** Test a *statistic S* instead…

Likelihood-Free Monte Carlo

D1: Generate $\theta \sim p(\theta)$ from prior

D2: Simulate $\mathcal{D}'$ from model with input θ

D3: Compute statistic S' of $\mathcal{D}'$

D4: Calculate distance $\rho(\mathcal{S}', \mathcal{S})$

D5: Accept θ if $\rho(\mathcal{S}', \mathcal{S}) \leq \epsilon$; Return to D1

- Typically higher acceptance rate than Algorithm C
- Target distribution is an even rougher approximation of true posterior,

$$p(\theta \mid \rho(S, S') \leq \epsilon) \approx p(\theta \mid \mathcal{D})$$

- Finding statistics that make this a good approximation is hard
- Standard statistics: mean, median, min, max, etc.

Previous Works

1. Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applies statistician. The Annals of Statistics 12: 1151–1172.
    a. A hypothetical sampling mechanism that yields a sample from the posterior distribution
    b. Coincides exactly with that of the ABC-rejection scheme
2. Diggle PJ, Gratton J (1984) Monte Carlo methods of inference for implicit statistical models. Journal of the Royal Statistical Society, Series B 46: 193–227.
    a. Defining a grid in the parameter space and using it to approximate the likelihood by running several simulations for each grid point
    b. The approximation was then improved by applying smoothing techniques to the outcomes of the simulations.

ABC - Rejection Algorithm

1. A set of parameter points is first sampled from the prior distribution.
   a. $\theta \sim p(\theta)$
2. Given a sampled parameter point θ, a dataset *D'* is then simulated under the statistical model M specified by θ.
   a. Generate *D'*
3. If the generated *D'* is too different from the observed data *D*, the sampled parameter value is discarded.
4. Outcome: A set of parameter points (ie: some θ values)

$$\rho(D', D) \leq \varepsilon$$

Modification

A common approach to lessen this problem is to replace D with a set of lower dimensional summary statistics S(D).

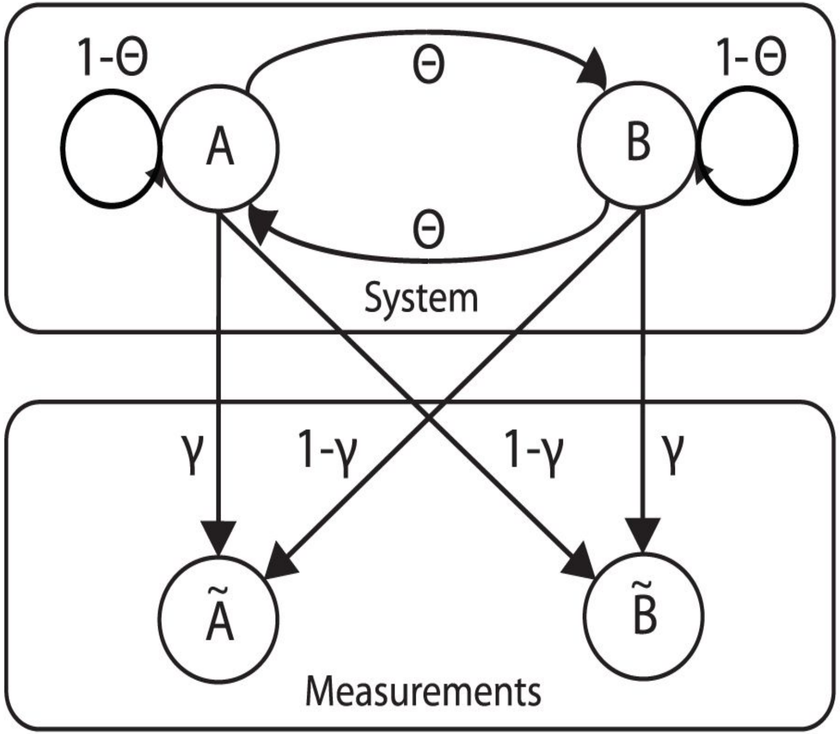
$$\rho(\, S(\, D' \,)\, ,\, S(\, D \,)\,) \leq \varepsilon$$

If the summary statistics are sufficient with respect to the model parameters θ, the efficiency increase obtained in this way does not introduce any error

Reality

1. It is typically impossible, outside the exponential family of distributions, to identify a finite-dimensional set of sufficient statistics.
2. Nevertheless, informative, but possibly non-sufficient, summary statistics are often used in applications where inference is performed with ABC methods.

Example: A dynamic bistable hidden Markov Model



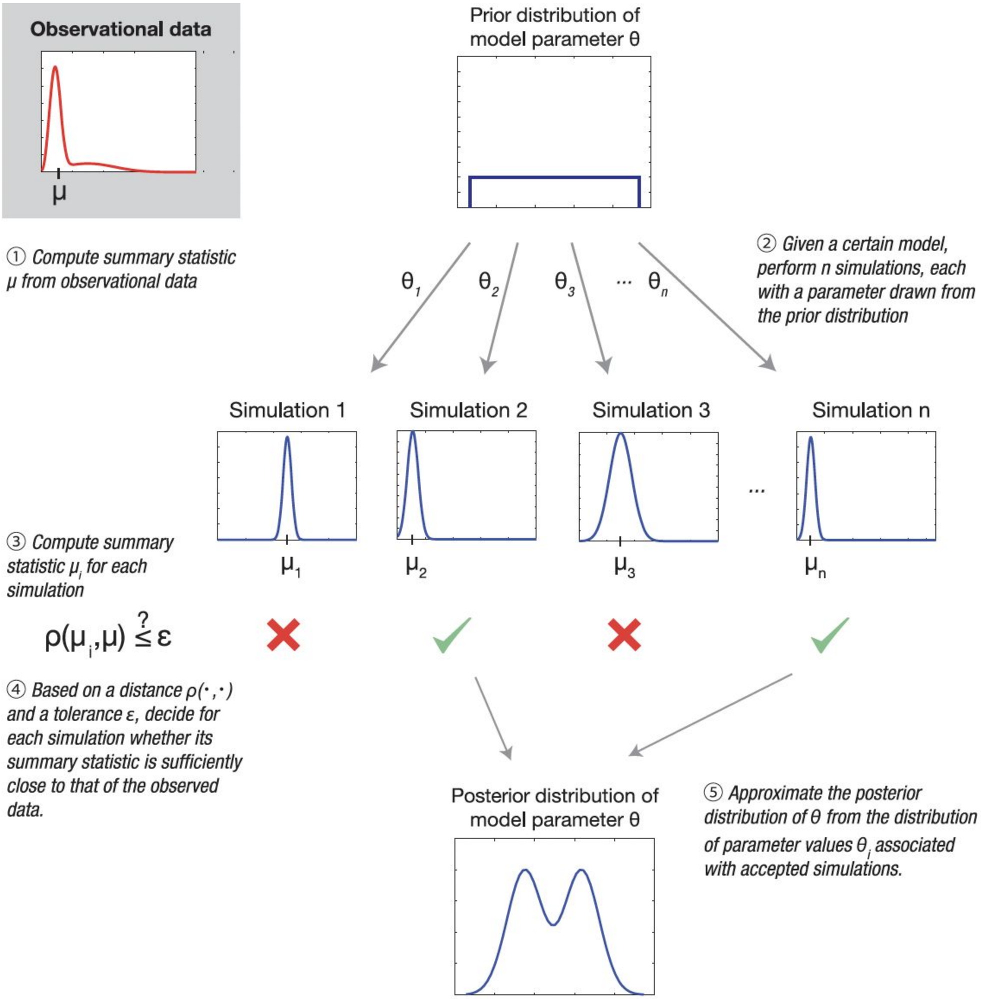


Figure 1. Parameter estimation by Approximate Bayesian Computation: a conceptual overview.
doi:10.1371/journal.pcbi.1002803.g001

Table 1. Example of ABC rejection algorithm.

i	θ_i	Simulated Datasets (Step 2)	Summary Statistic $\omega_{S,i}$ (Step 3)	Distance ρ ($\omega_{S,i}$,ω_E) (Step 4)	Outcome (Step 4)
1	0.08	AABAAAABAABAAABAAAAA	8	2	accepted
2	0.68	AABBABABAAABBABABBAB	13	7	rejected
3	0.87	BBBABBABBBBABABBBBBA	9	3	rejected
4	0.43	AABAAAAABBABBBBBBBBA	6	0	accepted
5	0.53	ABBBBBAABBABBABAABBB	9	3	rejected

doi:10.1371/journal.pcbi.1002803.t001

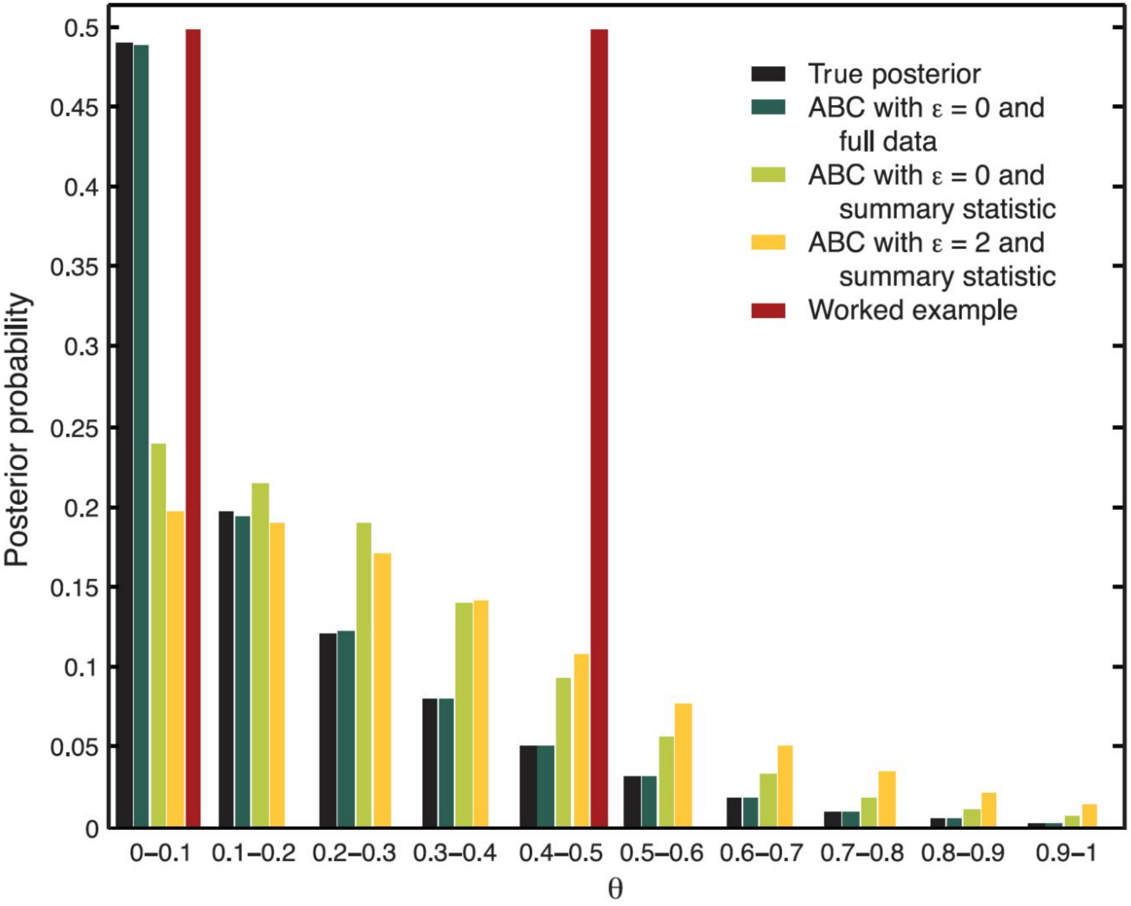


Figure 3. Posterior of θ obtained in the example (red), compared with the true posterior distribution (black), and ABC simulations with large *n*. The use of the insufficient summary statistic ω introduces a bias, even when requiring $\varepsilon = 0$ (light green).
doi:10.1371/journal.pcbi.1002803.g003

Techniques of Bayesian Model Comparison

Bayes Factor: $B_{1,2}$

$$\frac{p(M_1 | D)}{p(M_2 | D)} = \frac{p(D|M_1)\, p(M_1)}{p(D|M_2)\, p(M_2)} = B_{1,2} \frac{p(M_1)}{p(M_2)}$$

1. Computational improvements for ABC in the space of models have been proposed, such as constructing a particle filter in the joint space of models and parameters
2. In practice, these measures can be highly sensitive to the choice of parameter prior distributions and summary statistics, and thus conclusions of model comparison should be drawn with caution.

Pitfalls and Remedies

1. Like many other inference tactics, ABC resorts to some approximations and assumptions in order to make the inference computationally tractable
   a. Eg: $\varepsilon > 0$ is set
2. Sufficient statistics are typically not available, and instead, other summary statistics are used, which introduces an additional bias due to loss of information
3. Some more:
   a. Didelot X, Everitt RG, Johansen AM, Lawson DJ (2011) Likelihood-free estimation of model evidence
   b. Robert CP, Cornuet J-M, Marin J-M, Pillai NS (2011) Lack of confidence in approximate Bayesian computation model choice

Criticism of ABC

On top of usual criticism against Bayesian framework, some of the general pitfalls are particularly relevant for ABC methods; since they can handle more complex model.

Approximation of the Posterior

$$p(\theta \mid D) \qquad\qquad p(\theta \mid \rho(D', D) \leq \varepsilon)$$

1. Sufficiently small tolerance + sensible distance measure:
   a. Reasonably well approximation of the target posterior
2. Large tolerance:
   a. We essentially get back the prior as posterior approximation
3. Empirical studies: Difference between the two as a function of ε.
4. Theoretical results: for an ε-dependent upper bound for the error in parameter estimates.
5. Accuracy of the posterior (defined as the expected quadratic loss) as a function of ε has been also investigated

Gaps…

1. However, the convergence of the distributions when ε approaches zero, and how it depends on the distance measure used, is an important topic that has yet to be investigated in greater detail.
2. It remains difficult to disentangle errors introduced by this approximation from errors due to model mis-specification.
3. more…

Remedies

1. As an attempt to correct some of the error due to a non-zero ε, the usage of local linear weighted regression with ABC to reduce the variance of the posterior estimates has been suggested
2. Obtained regression coefficients are used to correct sampled parameters in the direction of observed summaries.
3. An improvement was suggested in the form of nonlinear regression using a feed-forward neural network model. However, this turned out to be not always consistent with the prior.

Comments

1. The idea of using a non-zero tolerance ε is not inherently flawed: under the assumption of measurement errors, the optimal ε can in fact be shown to be not zero
2. Bias caused by a non-zero tolerance can be characterized and compensated by introducing a specific form of noise to the summary statistics
3. Asymptotic consistency for such “noisy ABC” has been established, together with formulas for the asymptotic variance of the parameter estimates for a fixed tolerance

Choice and Sufficiency of Summary Statistics

1. Low-dimensional sufficient statistics are optimal
2. Heuristic to obtain useful low-dimensional summary statistics
3. Use of a set of poorly chosen summary statistics will often lead to inflated credible intervals due to the implied loss of information
4. A review of methods for choosing summary statistics
   a. Blum MGB, Nunes MA, Prangle D, Sisson SA (2012) A comparative review of dimension reduction methods in approximate Bayesian computation. [29]

Suggestions

1. One Approach: Using many different statistics?
   a. Turned out to be not a great idea. Accuracy and stability of ABC appears to decrease rapidly with an increasing numbers of summary statistics
2. A better strategy: focus on the relevant statistics only
   a. Relevancy depending on the whole inference problem, on the model used, and on the data at hand

Two Approaches

1. An algorithm has been proposed for identifying a representative subset of summary statistics
    a. Iteratively assessing whether an additional statistic introduces a meaningful modification of the posterior
    b. One of the challenges here is that a large ABC approximation error may heavily influence the conclusions about the usefulness of a statistic at any stage of the procedure.
2. Another method decomposes into two main steps
    a. First, a reference approximation of the posterior is constructed by minimizing the entropy.
    b. Sets of candidate summaries are then evaluated by comparing the ABC-approximated posteriors with the reference posterior.

Slightly different approaches:

1. "Partial Least Squares Regression" based approach
   a. Uses information from all the candidate statistics, each being weighted appropriately.
2. Constructing summaries in a semi-automatic manner
   a. Based on the observation that the optimal choice of summary statistics, when minimizing the quadratic loss of the parameter point estimates, can be obtained through the posterior mean of the parameters, which is approximated by performing a linear regression based on the simulated data.

Comments:

Methods for the identification of summary statistics that could also simultaneously assess the influence on the approximation of the posterior would be of substantial value.

However, none of the methods - discussed so far - assess the choice of summaries for the purpose of model selection

Bayes Factor with ABC and Summary Statistics

Combination of insufficient summary statistics and ABC for model selection can be problematic

$$B_{1,2} = \frac{p(D|M_1)}{p(D|M_2)} = \frac{p(D|S(D),M_1)\,p(S(D)|M_1)}{p(D|S(D),M_2)\,p(S(D)|M_2)} = \frac{p(D|S(D),M_1)}{p(D|S(D),M_2)} B^s_{1,2}.$$

Thus, a summary statistic $S(D)$ is sufficient for comparing two models M_1 and M_2 if and only if:

$$p(D|S(D),M_1) = p(D|S(D),M_2),$$

Findings

1. Sufficiency for M1 or M2 alone, or for both models, does not guarantee sufficiency for ranking the models
2. However, it was also shown that any sufficient summary statistic for a model M in which both M1 and M2 are nested is valid for ranking the nested models
3. Alternatively, necessary and sufficient conditions on summary statistics for a consistent Bayesian model choice have recently been derived [34], which can provide useful guidance.
4. If there is no dimensionality reduction going on, then ABC does not suffer these drawbacks.

Indispensable Quality Controls

Specifically,

1. Choice of competing models/hypotheses,
2. Number of simulations,
3. Choice of summary statistics,
4. Acceptance threshold

All these, cannot currently be based on general rules, but the effect of these choices should be evaluated and tested in each study [10].

Quality Controls (Continued…)

1. Quantification of the fraction of parameter variance explained by the summary statistics
2. A common class of methods aims at assessing whether or not the inference yields valid results, regardless of the actually observed data.
3. One can generate a large number of artificial datasets. Quality and robustness of ABC inference can be assessed in a controlled setting, by gauging how well the chosen ABC inference method recovers the true parameter values, and also models if multiple structurally different models are considered simultaneously.

1. Another class of methods assesses whether the inference was successful in light of the given observed data, for example by comparing the posterior predictive distribution of summary statistics to the summary statistics observed
2. cross-validation techniques [36] and predictive checks [37,38] represent promising future strategies to evaluate the stability and out-of-sample predictive validity of ABC inferences.
    a. particularly important when modeling large datasets, because then the posterior support of a particular model can appear overwhelmingly conclusive, even if all proposed models in fact are poor representations of the stochastic system underlying the observation data.
    b. Out-of-sample predictive checks can reveal potential systematic biases within a model and provide clues on to how to improve its structure or parametrization.

1. ABC allows, by construction, estimation of the discrepancies between the observed data and the model predictions, with respect to a comprehensive set of statistics.
2. These statistics are not necessarily the same as those used in the acceptance criterion.
3. Resulting discrepancy distributions have been used for selecting models that are in agreement with many aspects of the data simultaneously [39], and model inconsistency is detected from conflicting and codependent summaries.
4. Another quality-control-based method for model selection employs ABC to approximate the effective number of model parameters and the deviance of the posterior predictive distributions of summaries and parameters [

General Risks in Statistical Inference Exacerbated in ABC

1. Not specific to ABC, but also relevant for other statistical methods as well.
2. However, the flexibility offered by ABC to analyze very complex models makes them highly relevant to discuss here.

Prior Distribution and Parameter Ranges

1. Theoretical results regarding objective priors are available, which may for example be based on the principle of indifference or the principle of maximum entropy
2. In principle, uninformative and flat priors that exaggerate our subjective ignorance about the parameters may still yield reasonable parameter estimates.
3. However, Bayes factors are highly sensitive to the prior distribution of parameters
4. Conclusions on model choice based on Bayes factor can be misleading unless the sensitivity of conclusions to the choice of priors is carefully considered.

Small number of models

1. Model-based methods have been criticized for not exhaustively covering the hypothesis space
2. An upper limit to the number of considered candidate models is typically set by the substantial effort required to define the models and to choose between many alternative options
3. No commonly accepted ABC-specific procedure for model construction, so experience and prior knowledge are used instead
4. Sensible characterization of complex systems will always necessitate a great deal of detective work and use of expert knowledge from the problem domain.

Another Criticism and Counter-Argument

1. Some opponents of ABC contend that since only few models—subjectively chosen and probably all wrong—can be realistically considered, ABC analyses provide only limited insight
2. However, there is an important distinction between identifying a plausible null hypothesis and assessing the relative fit of alternative
3. Since useful null hypotheses, that potentially hold true, can extremely seldom be put forward in the context of complex models, predictive ability of statistical models as explanations of complex phenomena is far more important than the test of a statistical null hypothesis in this context.

Table 2. Potential risks and remedies in ABC-based statistical inference

Error Source	Potential Issue	Solution	Subsection
Nonzero tolerance ε	The inexactness introduces a bias in the computed posterior distribution.	Theoretical/practical studies of the sensitivity of the posterior distribution to the tolerance. Noisy ABC.	Approximation of the posterior
Nonsufficient summary statistics	The information loss causes inflated credible intervals.	Automatic selection/semi-automatic identification of sufficient statistics. Model validation checks (e.g., Templeton 2009 [19]).	Choice and sufficiency of summary statistics
Small number of models/mis-specified models	The investigated models are not representative/lack predictive power.	Careful selection of models. Evaluation of the predictive power.	Small number of models
Priors and parameter ranges	Conclusions may be sensitive to the choice of priors. Model choice may be meaningless.	Check sensitivity of Bayes factors to the choice of priors. Some theoretical results regarding choice of priors are available. Use alternative methods for model validation.	Prior distribution and parameter ranges
Curse-of-dimensionality	Low parameter acceptance rates. Model errors cannot be distinguished from an insufficient exploration of the parameter space. Risk of overfitting.	Methods for model reduction if applicable. Methods to speed up the parameter exploration. Quality controls to detect overfitting.	Curse-of-dimensionality
Model ranking with summary statistics	The computation of Bayes factors on summary statistics may not be related to the Bayes factors on the original data, which may therefore render the results meaningless.	Only use summary statistics that fulfill the necessary and sufficient conditions to produce a consistent Bayesian model choice. Use alternative methods for model validation.	Bayes factor with ABC and summary statistics
Implementation	Low protection to common assumptions in the simulation and the inference process.	Sanity checks of results. Standardization of software.	Indispensable quality controls

doi:10.1371/journal.pcbi.1002803.t002

Large Datasets

1. In some ABC-based analyses, part of the data have to be omitted
2. It has been proposed alternatively to combine the Metropolis-Hastings algorithm with ABC, which was reported to result in a higher acceptance rate than for plain ABC
3. Naturally, such an approach inherits the general burdens of MCMC methods, such as
   a. Difficulty to assess convergence,
   b. Correlation among the samples from the posterior [23],
   c. Relatively poor parallelizability

Large datasets (Continued…)

1. The ideas of Sequential Monte Carlo (SMC) and Population Monte Carlo (PMC) methods have been adapted to the ABC setting [23,45].
   a. General idea is to iteratively approach the posterior from the prior through a sequence of target distributions.
2. Benefit:
   a. Compared to ABC-MCMC, is that the samples from the resulting posterior are independent
   b. Tolerance levels must not be specified prior to the analysis, but are adjusted adaptively
3. It is relatively straightforward to parallelize a number of steps in ABC algorithms based on rejection sampling and sequential Monte Carlo methods.

Curse-of-Dimensionality

1. Can require an extremely large number of parameter points
2. Error of the ABC estimators as functions of the dimension of the summary statistics
3. Investigation done on how the dimension of the summary statistics is related to the mean squared error for different correction adjustments to the error of ABC estimators.
4. Another scheme is to project (possibly high-dimensional) data into estimates of the parameter posterior means; now having the same dimension as the parameters, are then used as summary statistics for ABC
5. One should account for the possibility of overfitting
6. Although no computational method (based on ABC or not) seems to be able to break the curse-of-dimensionality, methods have recently been developed to handle high-dimensional parameter spaces under certain assumptions (e.g., based on polynomial approximation on sparse grids)

Curse-of-Dimensionality (Continued…)

1. For certain problems, it might therefore be difficult to know whether the model is incorrect or, whether the explored region of the parameter space is inappropriate.
2. A more pragmatic approach is to cut the scope of the problem through model reduction

Table 3. Software incorporating ABC.

Software	Keywords and Features	Reference
DIY-ABC	Software for fit of genetic data to complex situations. Comparison of competing models. Parameter estimation. Computation of bias and precision measures for a given model and known parameters values.	[53]
ABC R package	Several ABC algorithms for performing parameter estimation and model selection. Nonlinear heteroscedastic regression methods for ABC. Cross-validation tool.	[54]
ABC-SysBio	Python package. Parameter inference and model selection for dynamical systems. Combines ABC rejection sampler, ABC SMC for parameter inference, and ABC SMC for model selection. Compatible with models written in Systems Biology Markup Language (SBML). Deterministic and stochastic models.	[55]
ABCtoolbox	Open source programs for various ABC algorithms including rejection sampling, MCMC without likelihood, a particle-based sampler, and ABC-GLM. Compatibility with most simulation and summary statistics computation programs.	[56]
msBayes	Open source software package consisting of several C and R programs that are run with a Perl "front-end." Hierarchical coalescent models. Population genetic data from multiple co-distributed species.	[57]
PopABC	Software package for inference of the pattern of demographic divergence. Coalescent simulation. Bayesian model choice.	[58]
ONeSAMP	Web-based program to estimate the effective population size from a sample of microsatellite genotypes. Estimates of effective population size, together with 95% credible limits.	[59]
ABC4F	Software for estimation of F-statistics for dominant data.	[60]
2BAD	Two-event Bayesian ADmixture. Software allowing up to two independent admixture events with up to three parental populations. Estimation of several parameters (admixture, effective sizes, etc.). Comparison of pairs of admixture models.	[61]

doi:10.1371/journal.pcbi.1002803.t003

Conclusion

1. ABC: Well-Founded. However, reliable application of ABC requires additional caution to be considered, due to the approximations and biases introduced at the different stages of the approach.
2. ABC: Best suited for inference about parameters or predictive inferences about observables in the presence of a single or few candidate model(s).
3. Since the computation of the likelihood function is bypassed, it can be tempting to attack high-dimensional problems using ABC, but inevitably this comes bundled with new challenges that investigators need to be aware of at each step of their analyses.

Acknowledgement

1. Professor Jason Pacheco’s slides on “Implicit Models”
2. Caleb Dahlke: For providing tremendous feedback and clarifications

Reference:

1. Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). [12] [csc 696h1] [f22] [paper] [reading] [Approximate Bayesian Computation]. PLOS Computational Biology, 9(1), e1002803. https://doi.org/10.1371/journal.pcbi.1002803
2. Professor Jason Pacheco’s slides on Implicit Models