Approximate Bayesian Computation (ABC) in population genetics

Matthieu Foll

04.04.2011

Reviews in Computational Biology

Outline

- ABC: what for?
- The link with population genetics: history of a controversy
- A “user-centered” review of ABC criticisms (including an introduction to ABC)
- Conclusion

ABC: what for?

- Approximate Bayesian Computation
- Monte Carlo method to approximate posterior distributions or likelihood surfaces from a model
- Numerical tool for solving problems within a statistical framework

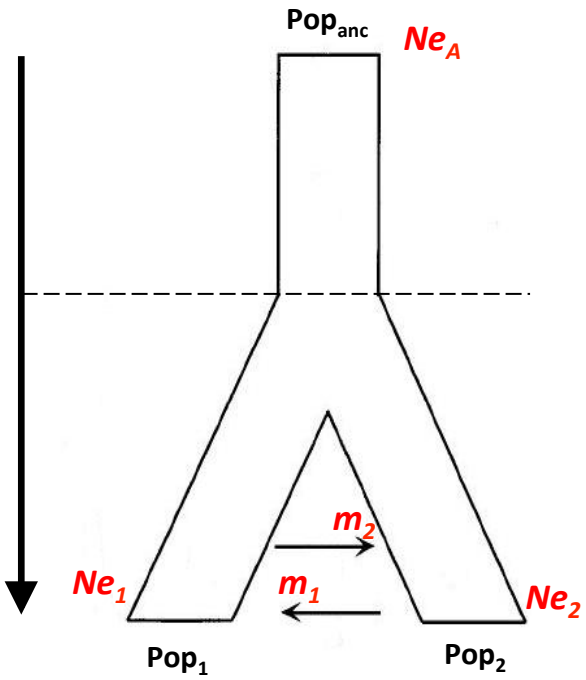
ABC: what for?

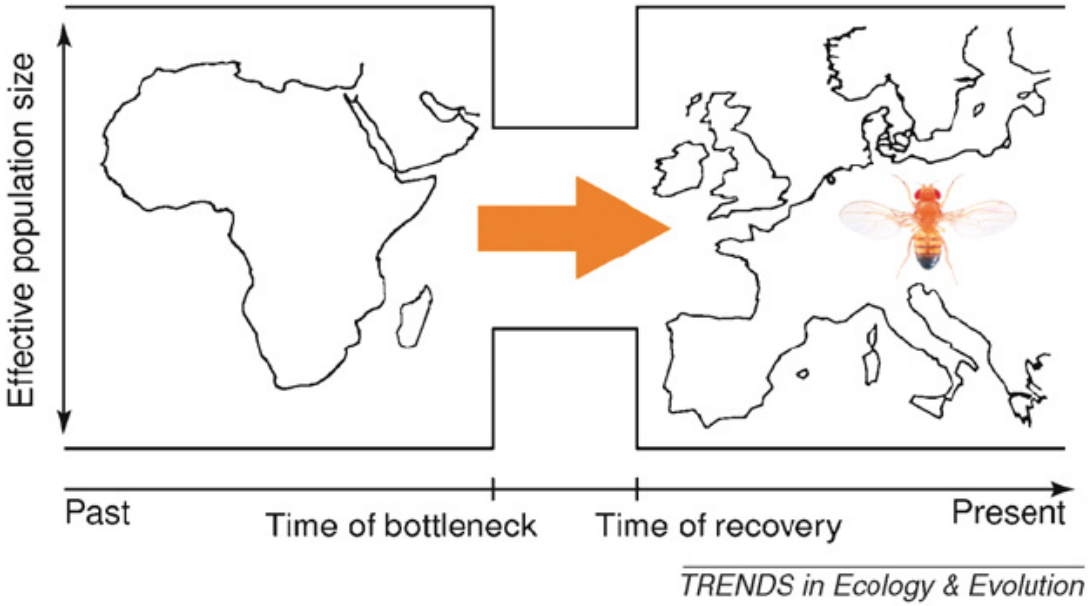
- We have a data set D, we suppose that we can represent the problem by a model M, determined by some parameters Φ
- $P(\Phi|D) \propto P(D|\Phi) * P(\Phi)$
  Posterior ∝ Likelihood * Prior
- Posterior distribution?
- Monte Carlo Markov Chain is the most widely used method for this purpose

In population genetics?

- Infer the demographic history of populations from genetic data

$\Phi = \{Ne_1, Ne_2, Ne_A, m_1, m_2, t\}$





History of a controversy

- How it started?
  - 2002: Knowles criticizes Nested Clade Phylogeographical Analysis (NCPA)
    Phylogeography: “studying patterns of genetic variation in a geographical context via gene trees”
  - 2004: answer, and improvements from Templeton
  - 2007: Strong charge against NCPA

Statistical phylogeography

L. LACEY KNOWLES and WAYNE P. MADDISON

Statistical phylogeography: methods of evaluating and minimizing inference errors

ALAN R. TEMPLETON

THE AUTOMATION AND EVALUATION OF NESTED CLADE PHYLOGEOGRAPHIC ANALYSIS

Mahesh Panchal[1] and Mark A. Beaumont[2]

The coup de grâce for the nested clade phylogeographic analysis?

RÉMY J. PETIT

When ABC is invited to the party

- 2008: Templeton starts to attack model based approaches as a reply

REPLY

Nested clade analysis: an extensively validated method for strong phylogeographic inference

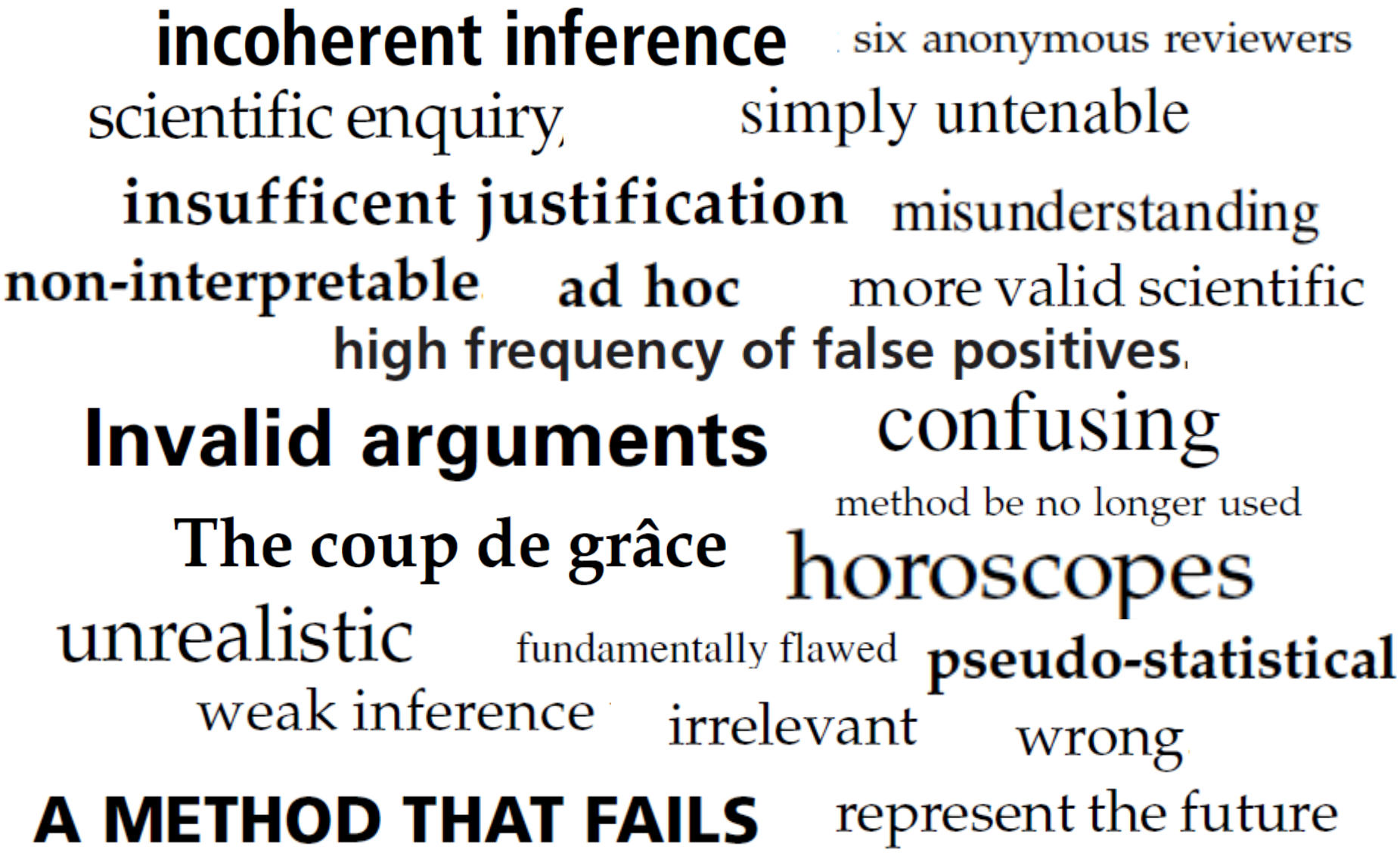
ALAN R. TEMPLETON

- 2009: Strong charge against ABC

Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation

ALAN R. TEMPLETON

2008: the war is open



Endless discussions follow…

COMMENTARY

COMMENT

REPLY

LETTER

Reply to

Letters

Letters Response

General comment

- Answering criticisms about one method by pointing out weaknesses of another is not very efficient
  - Sounds childish and personal
  - Create suspicions on your defense
- “My method is working” should be enough
- At the end, it sounds like “maybe my method is not working, but it’s the only one available”

ABC vs. NCPA: Comparing apples and oranges

- Statistics
  - Model based statistics
    - Bayesian statistics
      - Computational methods
        - » EM
        - » IS
        - » MCMC
        - » RJ-MCMC
        - » ABC
        - …
- Biology
  - Evolutionary biology
    - Molecular phylogenetics
      - Phylogeography
        - » NCPA

***James O. Berger*[a], *Stephen E. Fienberg*[b], *Adrian E. Raftery*[c], *and Christian P. Robert*[d,1]**

[a]*Department of Statistical Sciences, Duke University, Durham, NC 27708-0251;* [b]*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3891;* [c]*Department of Statistics, University of Washington, Seattle, WA 98195-4320; and* [d]*Centre De Recherche en Mathématiques de la Décision, Université Paris-Dauphine, 75775 Paris Cedex 16, France*

Molecular Ecology (2009) **18**, 319–331 doi: 10.1111/j.1365-294X.2008.04026.x

Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation

ALAN R. TEMPLETON

Department of Biology, Washington University, St. Louis, MO 63130-4899, USA

Disentangling ABC criticisms

- ABC is a Monte Carlo method to approximate posterior distributions or likelihood surfaces from a model
- Majority of criticisms are also aimed more generally against
  - Model-based statistics
  - Bayesian statistics

- Some are really ABC specific
  - Invalid
    - Why?
  - Valid
    - Solutions exist
    - Unsolved: perspective?

- Statistics
  - Model based statistics
    - Bayesian statistics
      - Computational methods
        - » EM
        - » IS
        - » MCMC
        - » RJ-MCMC
        - » ABC
        - …

Model-based/Bayesian criticisms

- Model-based methods do not cover the entire "hypothesis space"
  - compare only a small number of potentially mis-specified and subjectively chosen models
  - Back to the debate in the 1930s between Fisher and Neyman-Pearson hypothesis testing
- In ABC "parameter ranges and distributions are only guessed based upon the subjective opinion of the investigators"
  - Classical criticisms of prior in Bayesian statistics

When (real) mathematicians are invited

PNAS

Coherent and incoherent inference in phylogeography and human evolution

Alan R. Templeton[1]

Department of Biology, Washington University, St. Louis, MO 63130

Edited* by Eviatar Nevo, Institute of Evolution, Haifa, Israel, and approved March 3, 2010 (received for review September 16, 2009)

A hypothesis is nested within a more general hypothesis when it is a special case of the more general hypothesis. Composite hypoth-

Incoherent inference is formally illogical inference and represents a mathematical error.

PNAS

LETTER

Incoherent phylogeographic inference

***James O. Berger*[a], *Stephen E. Fienberg*[b], *Adrian E. Raftery*[c], *and Christian P. Robert*[d,1]**

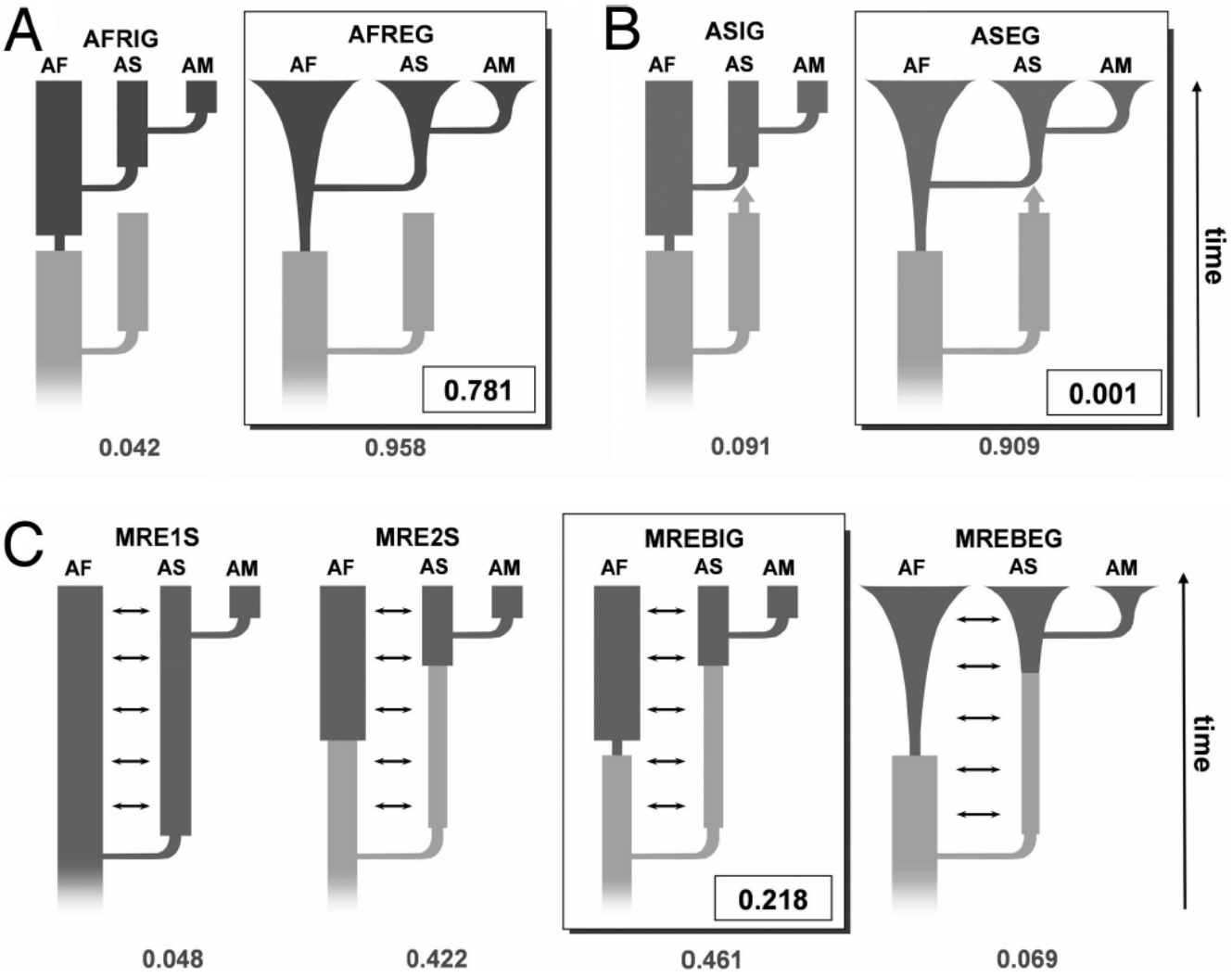
[a]*Department of Statistical Sciences, Duke University, Durham, NC 27708-0251;* [b]*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3891;* [c]*Department of Statistics, University of Washington, Seattle, WA 98195-4320; and* [d]*Centre De Recherche en Mathématiques de la Décision, Université Paris-Dauphine, 75775 Paris Cedex 16, France*

"ABC is simply a numerical computational technique; attacking it as incoherent is similar to calling calculus incoherent if it is used to compute the wrong thing"

Invalid Bayesian criticisms

- In “ABC there is no null hypothesis, which complicates the computation of sampling error”
- “The posterior probabilities that emerge from ABC [are] mathematically impossible ... to be probabilities”
- “The probability of the nested special case must be less than or equal to the probability of the general model within which the special case is nested”

Typical « nested hypotheses » example



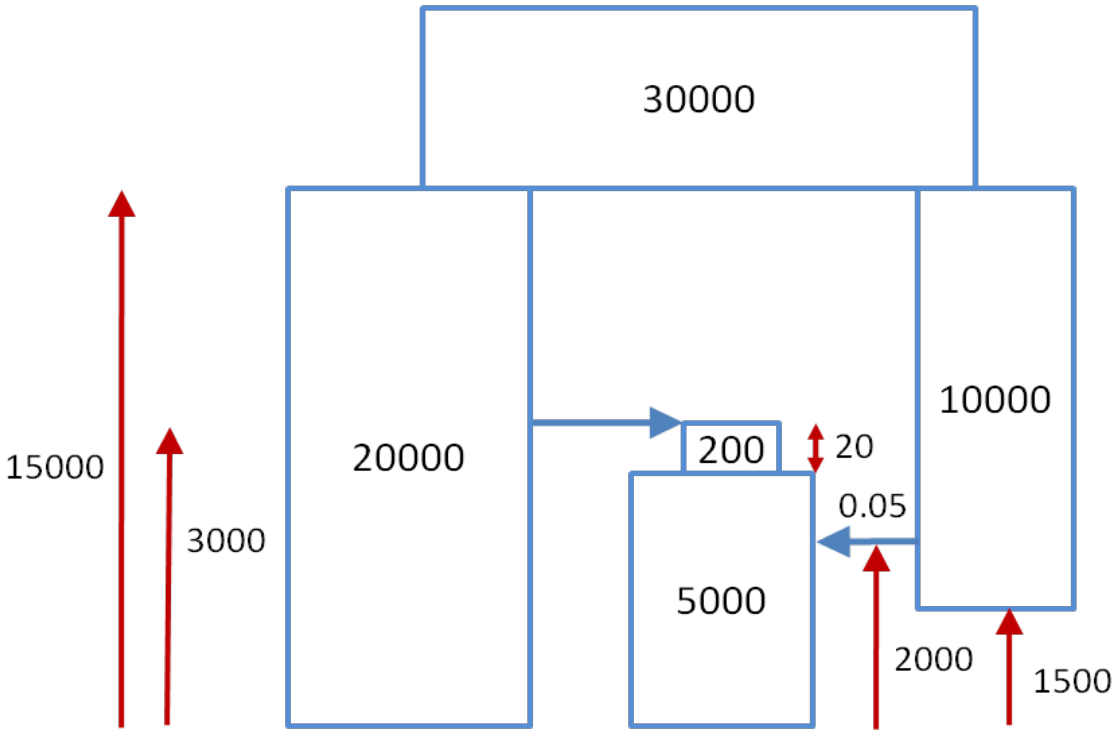
Statistical evaluation of alternative models of human evolution

Nelson J. R. Fagundes[†‡§], Nicolas Ray[§], Mark Beaumont[¶], Samuel Neuenschwander[§‖], Francisco M. Salzano[‡††], Sandro L. Bonatto[†,††], and Laurent Excoffier[§††]

OK, but how does ABC work?

- We have a data set D
- We suppose that we can represent the problem by a model M
- M is determined by some parameters Φ
- $P(\Phi|D) \propto P(D|\Phi) * P(\Phi)$
  Posterior distribution?
- We can not calculate the likelihood but…
- …we are able to simulate M

Easy to simulate: coalescent theory



Applications note / Genetic and population analysis

fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios

Excoffier L[1,2,*], Foll, M[1,2]

Rejection algorithm

1. Generate Φ at random (=prior)
2. Simulate D’ from M with parameters Φ
3. Accept Φ if D=D’ and return to 1
4. Stop when sufficient data sets have been accepted

The ABC algorithm

1. Generate Φ at random (=prior)
2. Simulate D′ from M with parameters Φ
3. Accept Φ if $d(D,D') < \varepsilon$ and return to 1
4. Stop when sufficient data sets have been accepted

The ABC algorithm with summaries

- Choose summary statistics S to represent D and calculate s for D.
- Generate Φ at random (=prior)
- Simulate D′ from M with parameters Φ
- Calculate s′ for D′
- Accept Φ if $d(s,s') < \varepsilon$ and return to 2
- Stop when sufficient data sets have been accepted.

The Annals of Statistics
1984, Vol. 12, No. 4, 1151–1172

SPECIAL INVITED PAPER

BAYESIANLY JUSTIFIABLE AND RELEVANT FREQUENCY CALCULATIONS FOR THE APPLIED STATISTICIAN[1]

BY DONALD B. RUBIN

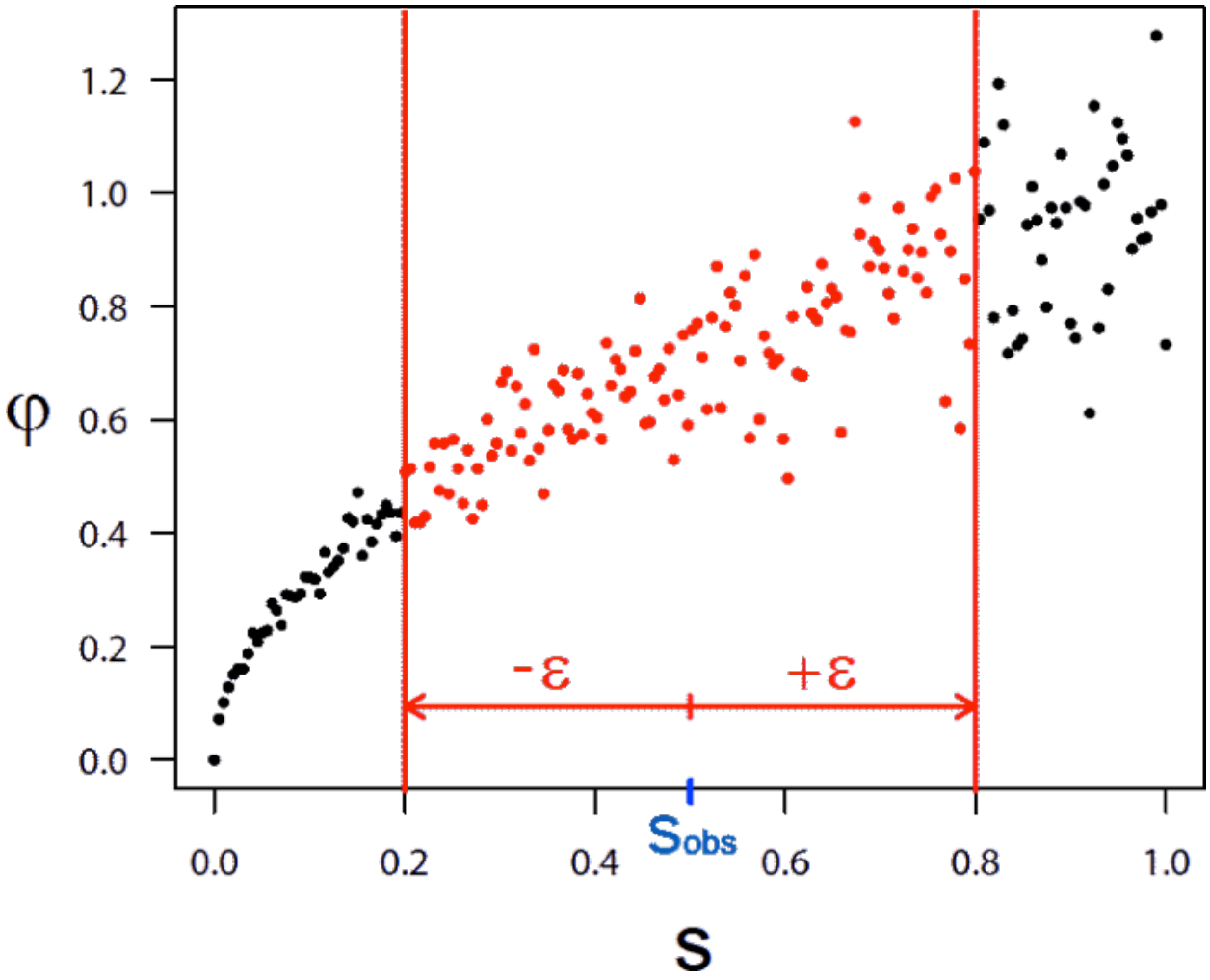
Suppose we first draw equally likely values of θ from $p(\theta)$, and label these $\theta_1, \cdots, \theta_s$.

For each θ_j, we now draw an X from $f(X \mid \theta = \theta_j)$; label these $X_1, \cdots, X_s$. The X_j represent possible values

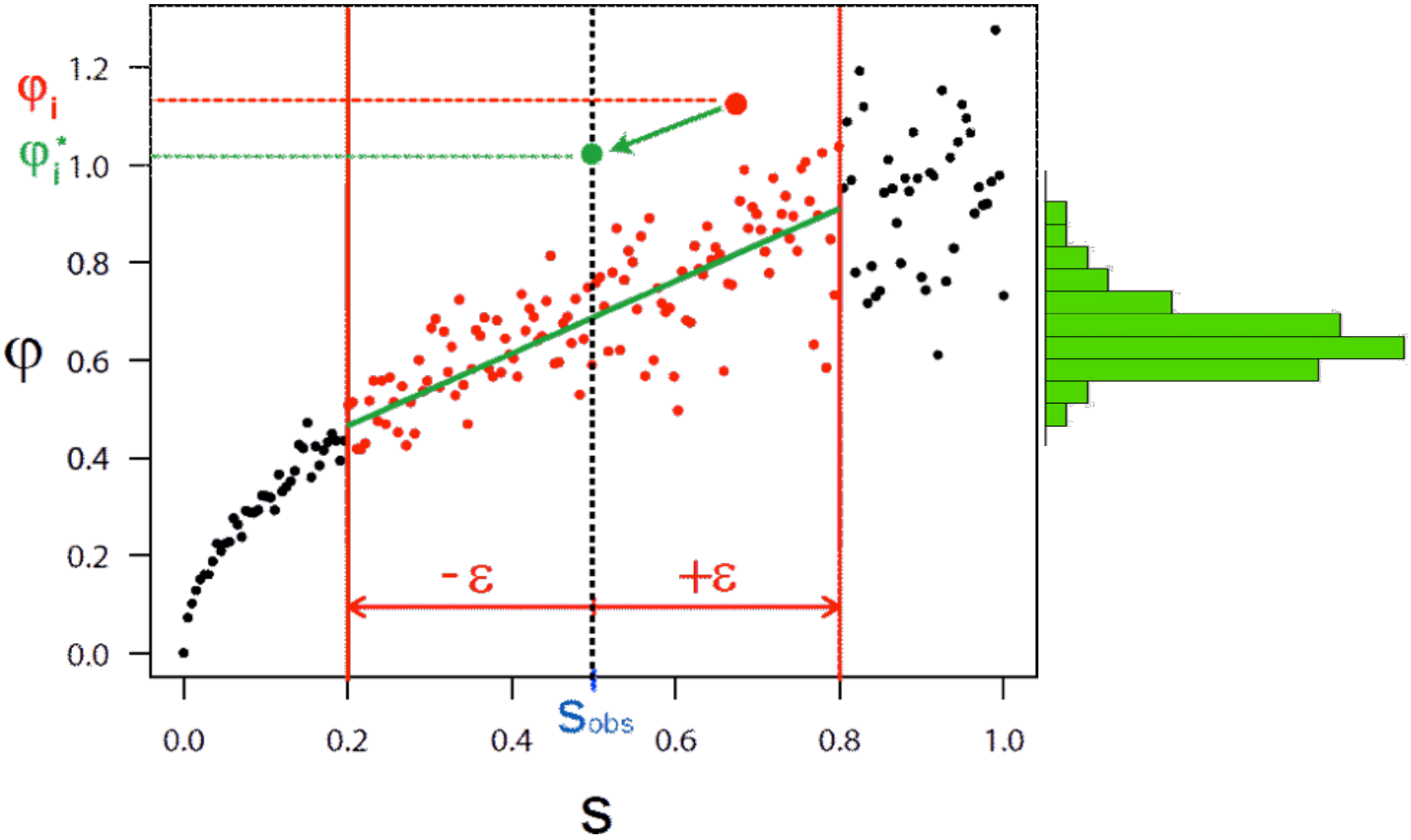
Suppose we collect together all X_j that match the observed X, and then all θ_j that correspond to these X_j.

formally, this collection of θ values represents the posterior distribution of θ.

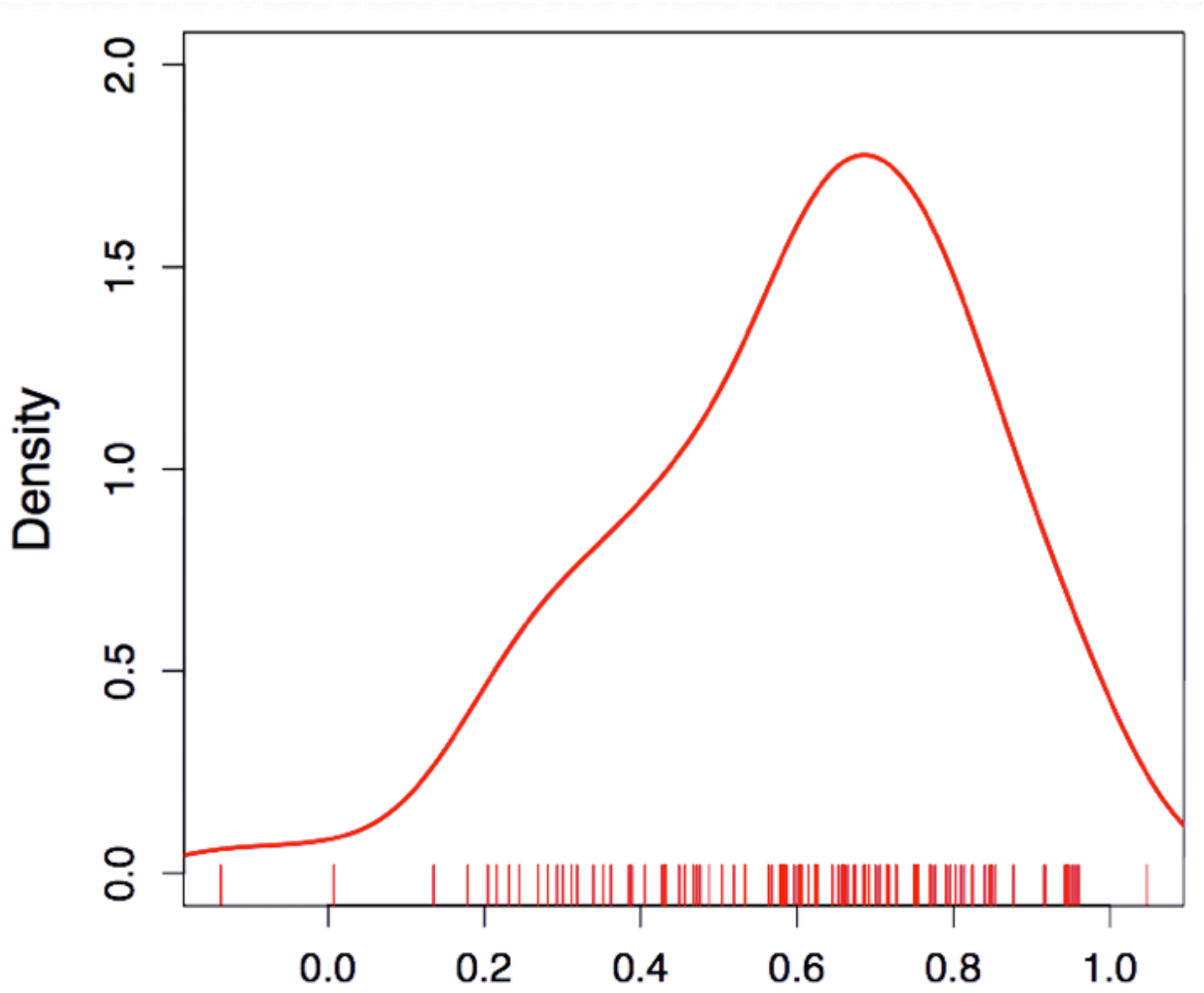
Example



Refinement (regression)



Posterior distribution



Model choice

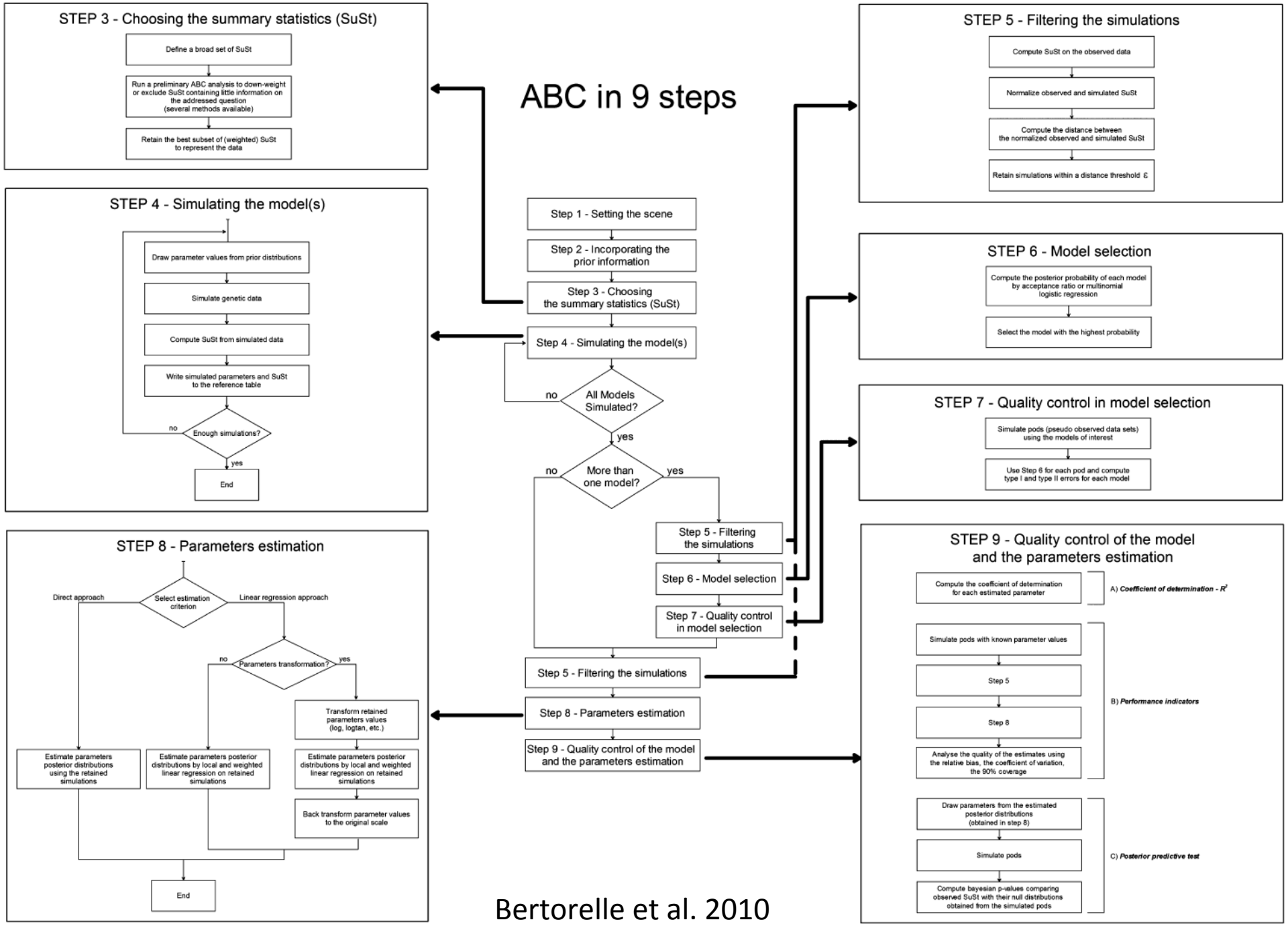
- Generate M and Φ_M at random (=prior)
- Simulate D' from M with parameters Φ_M

...

- The posterior probabilities of each model is approximated by the fraction of simulations produced by each of them

Improvements

- Once you have found a good region, explore it a bit more !
  - ABC-MCMC
- First do a quick tour of the world, and then go back to your favorite regions to explore a bit more !
  - SMC, ABC-PMC
- Automatic choice of independent summaries that best explain your data
  - PLS
- Non-linear regression



Bertorelle et al. 2010

Already reviewed extensively

Annual Review of Ecology, Evolution, and Systematics

Volume 41, 2010

Approximate Bayesian Computation in Evolution and Ecology

Mark A. Beaumont

Review

Cell PRESS

Approximate Bayesian Computation (ABC) in practice

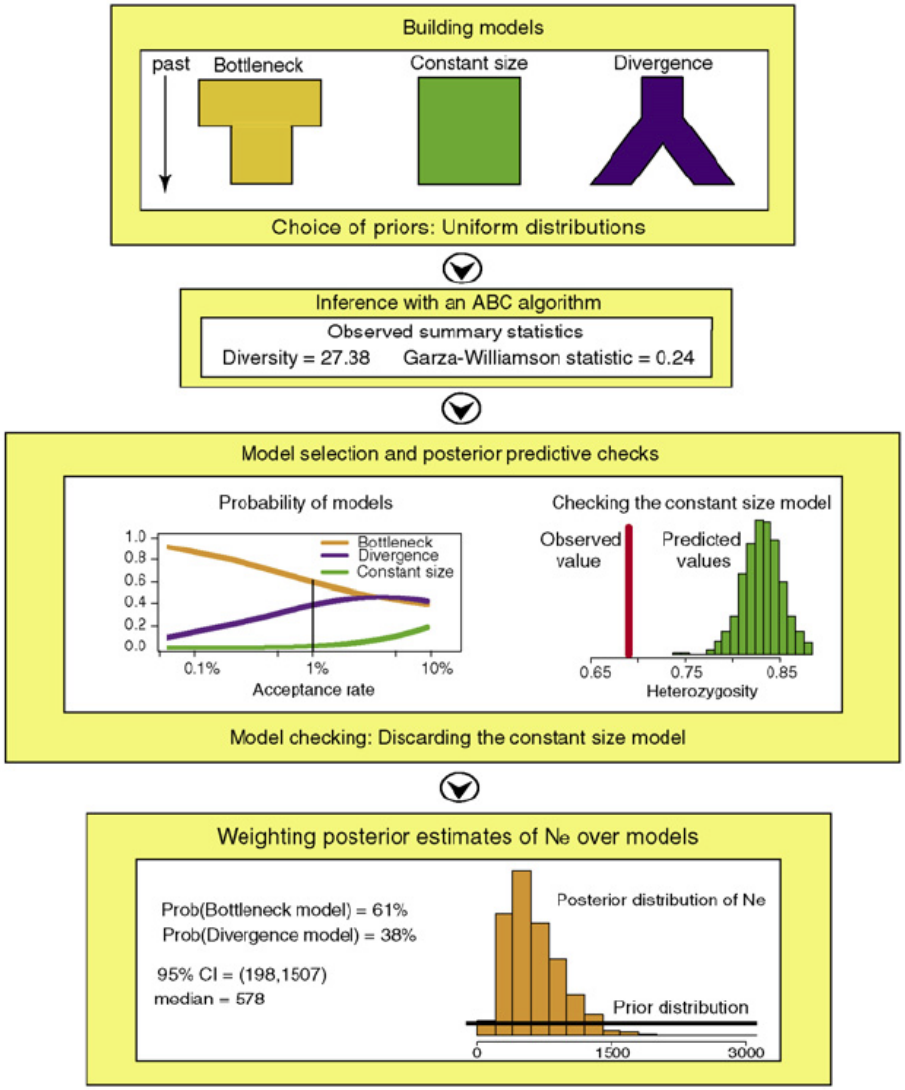
Katalin Csilléry[1], Michael G.B. Blum[1], Oscar E. Gaggiotti[2] and Olivier François[1]

INVITED REVIEW

ABC as a flexible framework to estimate demography over space and time: some cons, many pros

G. BERTORELLE,* A. BENAZZO* and S. MONA*†‡

Typical population genetics application



Approximate Bayesian Computation (ABC) in practice

Katalin Csilléry[1], Michael G.B. Blum[1], Oscar E. Gaggiotti[2] and Olivier François[1]

TRENDS in Ecology & Evolution

Invalid ABC-specific criticisms

- Simulation weighting
- Posterior densities and Bayesian model choice
- Sample size

- Already reviewed:

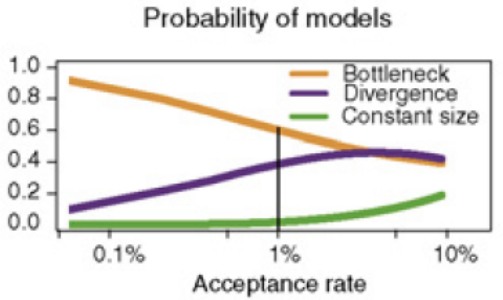
REPLY

In defence of model-based inference in phylogeography

MARK A. BEAUMONT,[1] RASMUS NIELSEN,[2] CHRISTIAN ROBERT,[3] JODY HEY,[4] OSCAR GAGGIOTTI,[5] LACEY KNOWLES,[6] ARNAUD ESTOUP,[7] MAHESH PANCHAL,[8] JUKKA CORANDER,[9] MIKE HICKERSON,[10] SCOTT A. SISSON,[11] NELSON FAGUNDES,[12] LOUNÈS CHIKHI,[13] PETER BEERLI,[14] RENAUD VITALIS,[15] JEAN-MARIE CORNUET,[7] JOHN HUELSENBECK,[2] MATTHIEU FOLL,[16,17] ZIHENG YANG,[18] FRANCOIS ROUSSET,[19] DAVID BALDING[20] and LAURENT EXCOFFIER[16,17]

Valid ABC criticisms

- Summary statistics
  - Choice
  - Sufficiency
- Approximation due to d(D, D′)<ε: choice of acceptance rate
- (linear) regression
- Model choice: wrong estimate of Bayes Factors (Didelot *et al.* 2011)



- Need to do it properly, in particular quality control step, which is *one* solution to most of these problems

Conclusion

- Important issue in population genetics and phylogeography (2008: >1700 NCPA citations)
- Interesting controversy
- For users, need to separate criticisms
  - ABC specific (valid/unvalid)
  - Model-based/Bayesian statistics (valid/unvalid)