Approximate Bayesian Computation (ABC) in practice

olivier.francois@imag.fr

Journées MAS 2010 – Bordeaux

Outline

- Short introduction to likelihood-free inference and Approximate Bayesian Computation for complex models
- The likelihood is not available analytically: Inference is based on Monte-Carlo simulations and summary statistics instead of the full data
- Rejection algorithm
- Part 1: Conditional density estimation algorithm
- Part 2: An exact hierarchical Bayes model
- Application: Demographic inference in coalescent models

Example of use of computer simulations in population genetics

- The use of simulations and summary statistics has a long tradition in population genetics

CAVALLI-SFORZA, L. L., and ZEI, G. 1967. Experiments with an artificial population. Pp. 473–478 in J. F. CROW and J. V. NEEL (eds.), *Third world congress of human genetics: proceedings*. Johns Hopkins, Baltimore.

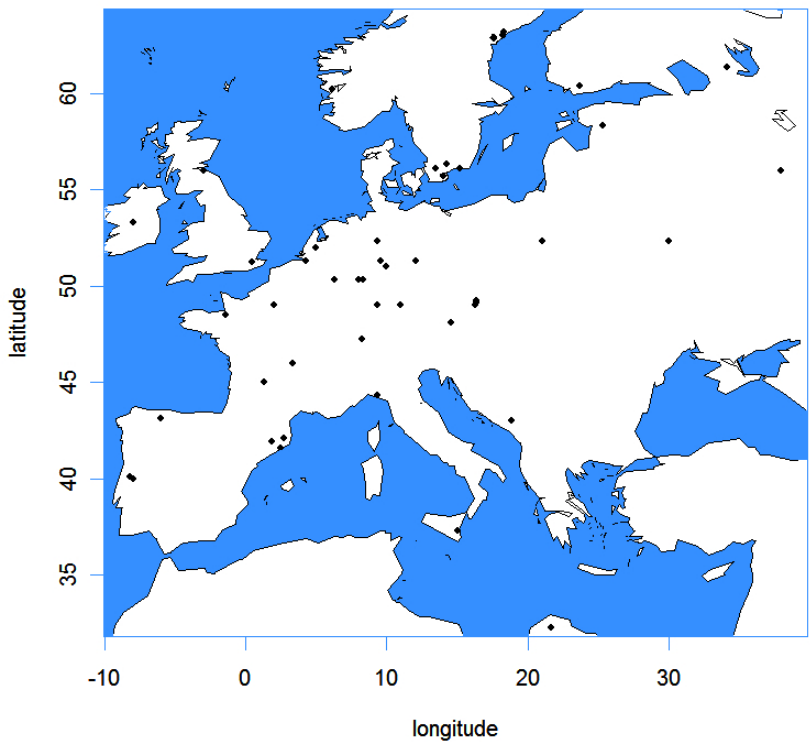
- Example: Dating the (latest) expansion of *Arabidopsis thaliana* in Europe

Arabidopsis thaliana

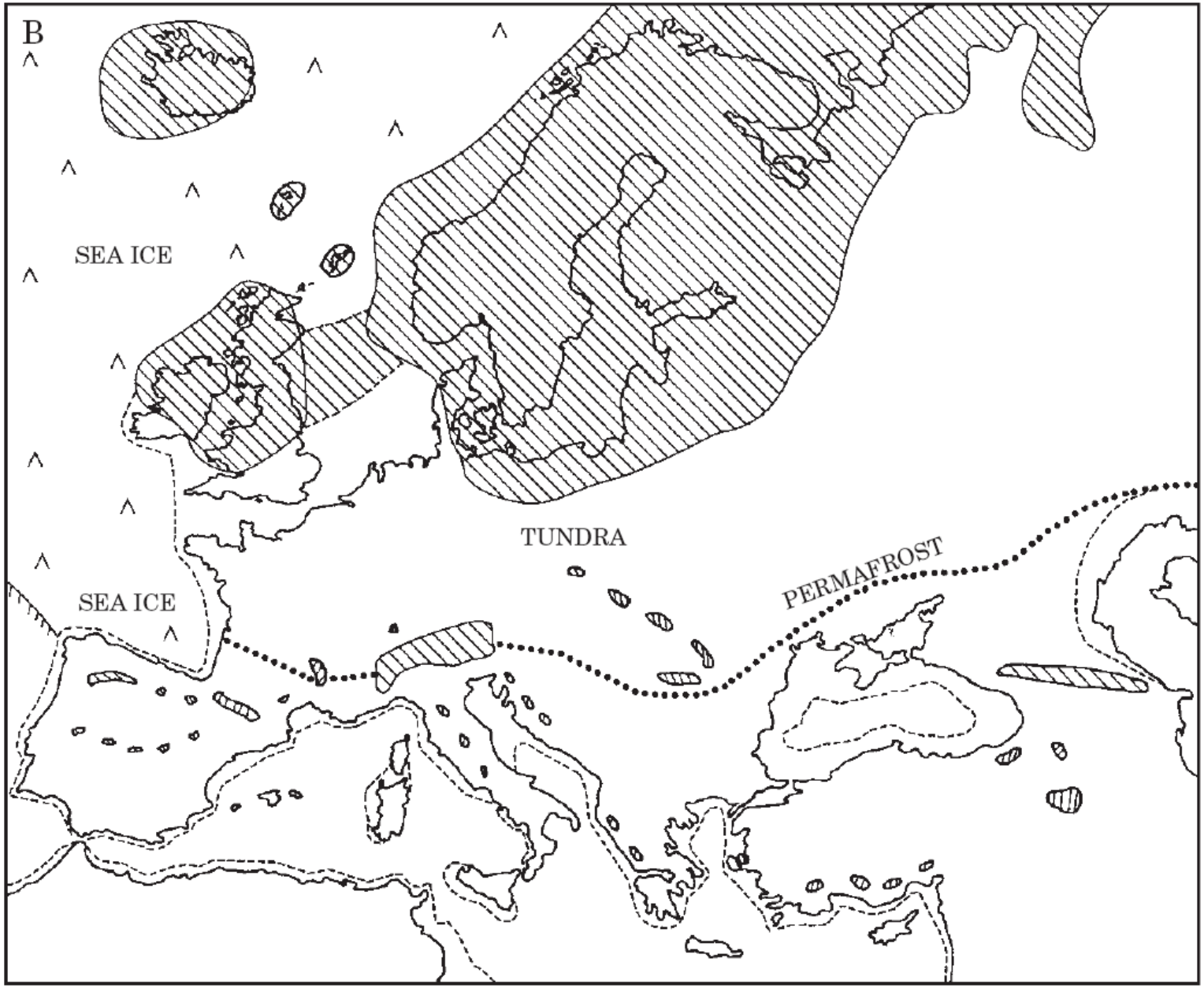
- Model organism for plant genetic research (flowering and adaptation).
- Small genome: 5 chromosomes, 150 Mbp.
- Lifestyle: weedy species.

Genomic data: Nordborg et al 2005

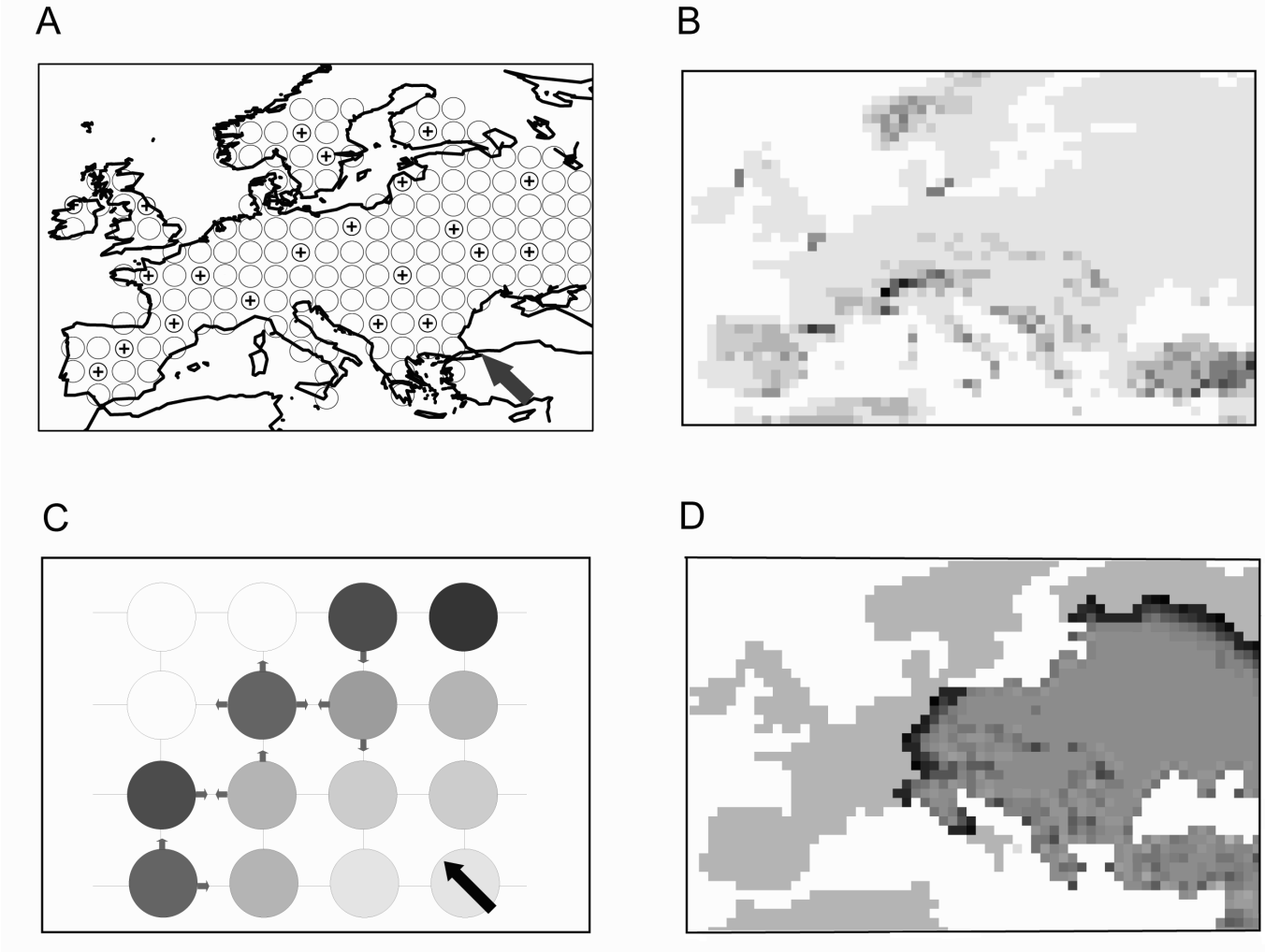
- 95 plants sampled world-wide (76 European individuals)
- 876 alignments of intra and inter-genic sequences (480,000 bp for each individual)



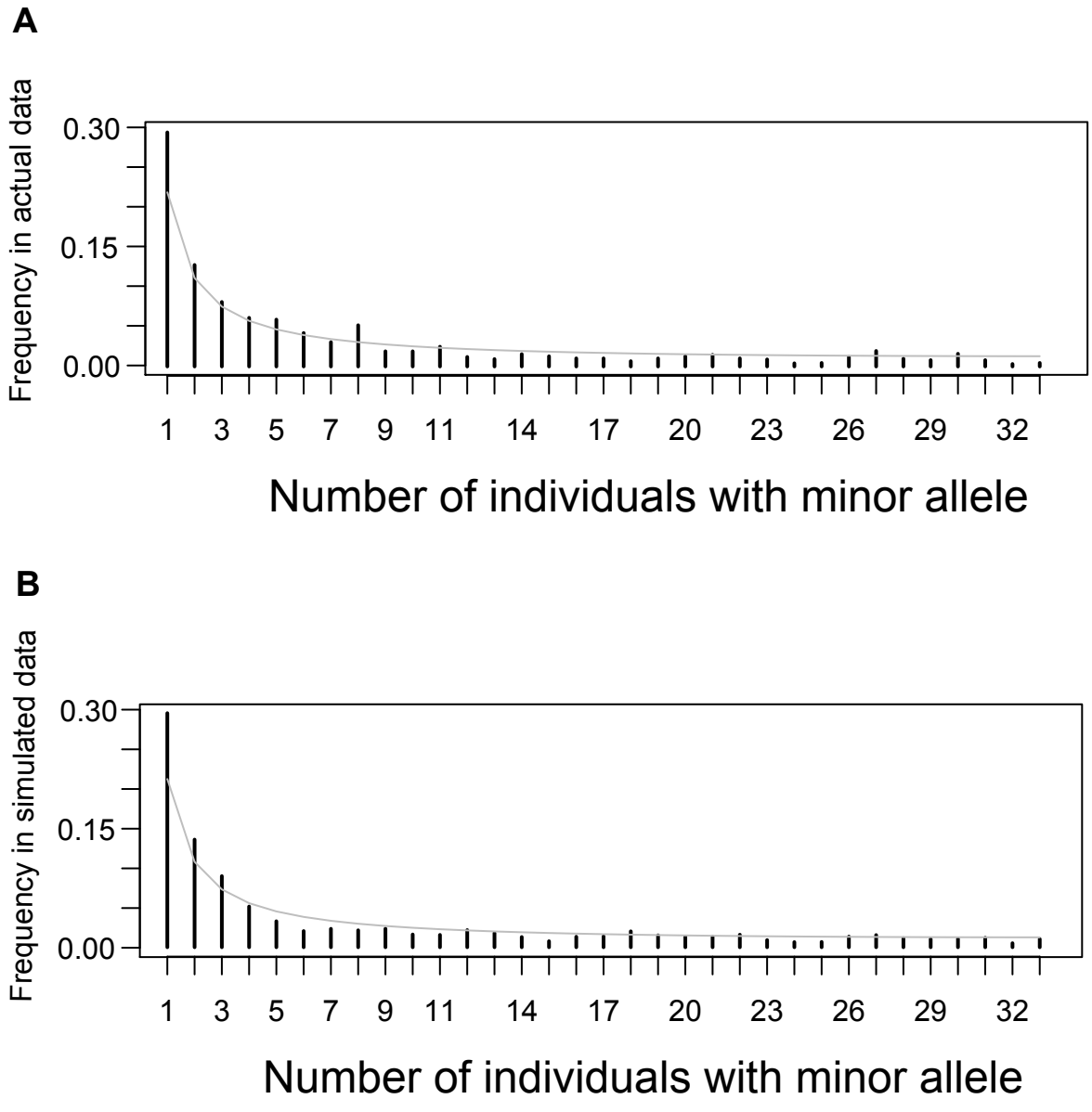
Europe, 18Ky BP



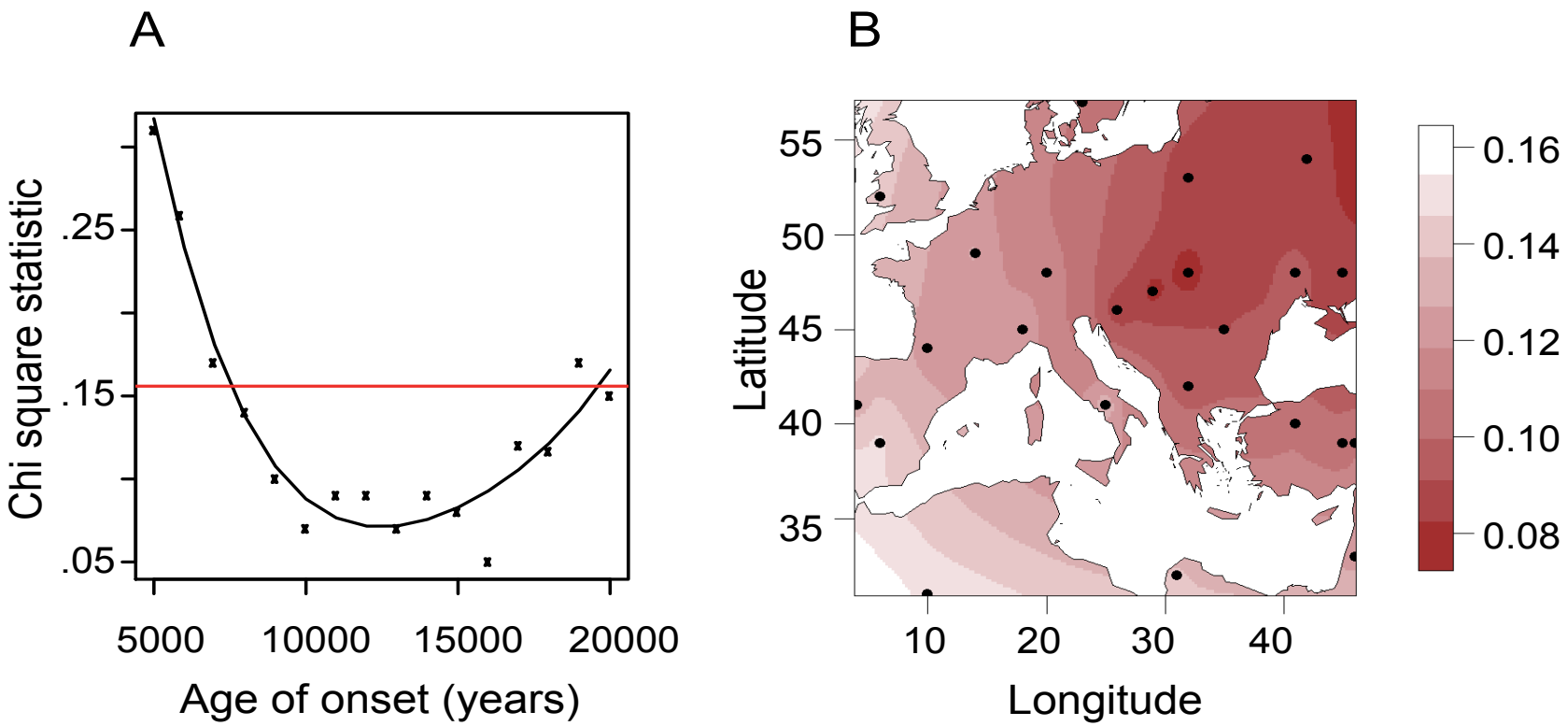
Past demography: Wave-of-advance models



Best fitting scenario: southeastern origin, $\sim$ 10-12 Ky ago



χ^2 distance between the actual and simulated frequency spectra



Summary

- Origin to the South-East of Europe (Black Sea)
- Speed $\sim$ 0.9km/year
- Onset of expansion $\sim$ 10,000 y ago
- Suggest a correlation with the expansion of Agriculture

OPEN ACCESS Freely available online PLoS GENETICS

Demographic History of European Populations of *Arabidopsis thaliana*

Olivier François[1*], Michael G. B. Blum[2], Mattias Jakobsson[3,4], Noah A. Rosenberg[3,4]

Bayesian inference

- Parameter $\theta = (\theta_1, \ldots, \theta_J)$, $J \geq 1$.
- Data $y = (y_1, \ldots, y_n)$, $n \geq 1$.
- $p(\theta)$ is the prior distribution.
- $p(y|\theta)$ is the likelihood or sampling distribution.
- We use the Bayes formula to compute the posterior distribution

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Inferring allele frequencies

- Prior distribution on the allele frequency: $\theta \sim$ beta(1,1) (uniform)
- Data: We observe the derived allele $y = 9$ times in a sample of size $n = 20$ (frequency = .45)
- Sampling distribution

$$p(y|\theta) = \text{binom}(n, \theta)(y) \propto \theta^y (1-\theta)^{n-y}$$

- The posterior distribution is

$$p(\theta|y) = \text{beta}(y+1, n+1-y)(\theta).$$

ABC of Approximated Bayesian Computation

- Suppose that y is discrete. The rejection algorithm

  Repeat
  1. sample θ from the prior distribution $p(\theta)$;
  2. sample y_s from the sampling distribution $p(y|\theta)$;

  Until $(y_s = y)$
  return(θ)

  generates samples from the posterior distribution.

$$p_y(\theta) = \sum_{s=1}^{\infty} (1 - p(y))^{s-1} p(y, \theta) = p(\theta|y).$$

Approximate Bayesian Computation tolerates an imperfect match

- The algorithm

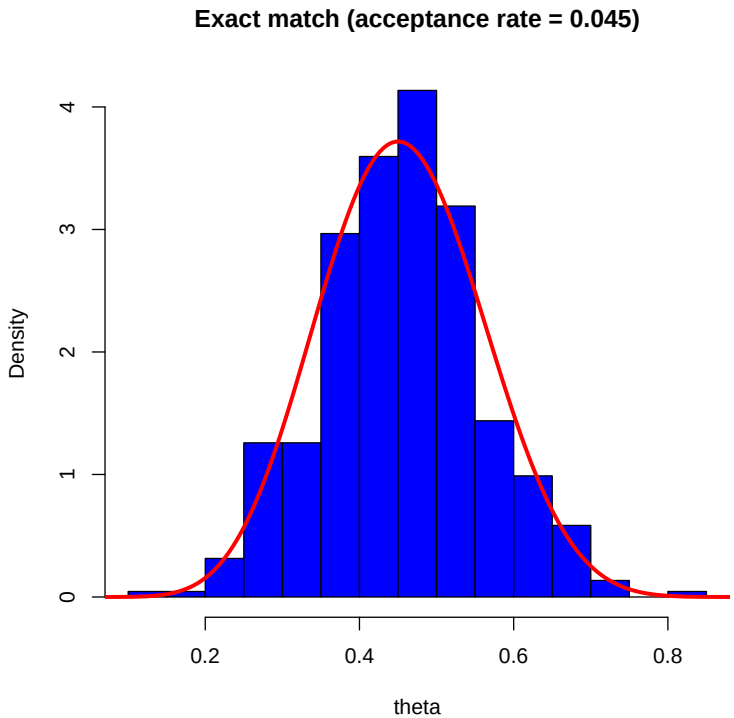
  Repeat
  1. sample θ from the prior distribution $p(\theta)$;
  2. sample y_s from the sampling distribution $p(y|\theta)$;

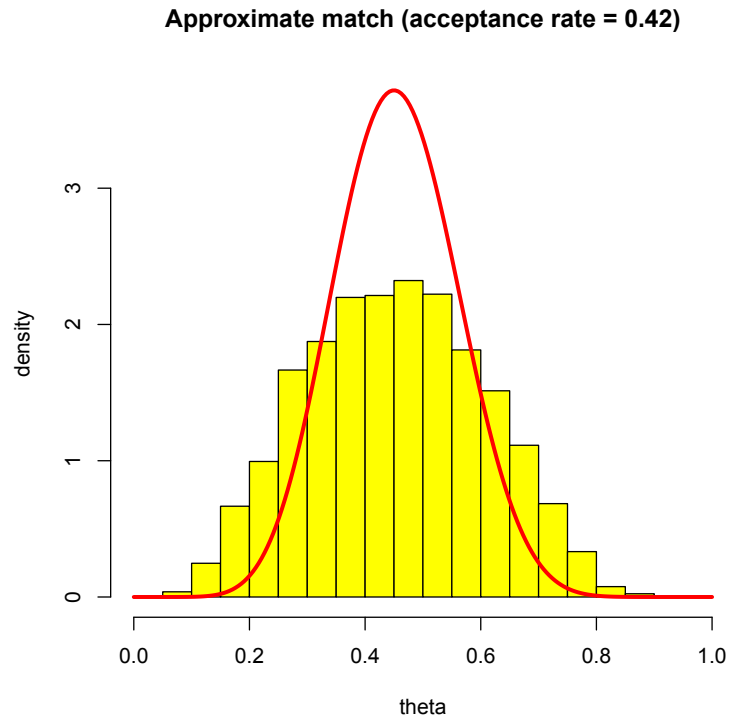
  Until $(|y_s - y| < \epsilon)$
  return(θ)

  generates samples from an approximation of the posterior distribution

$$p_\epsilon(\theta \mid y) \propto \Pr(|y_s - y| < \epsilon \mid \theta) p(\theta).$$

Exact vs approximate matching ($\epsilon = 5$)





Brief history of ABC

- Frequentist statistics: *Implicit models* (Diggle and Gratton 1984), *indirect inference* (Gourieroux et al 1993)
- Bayesian statistics: Tavaré et al (1997); Pritchard et al (1999)
- Termed ABC after (Beaumont et al 2002)
- *ABC easy as 123?*

Part 1. ABC viewed as a conditional density estimation algorithm

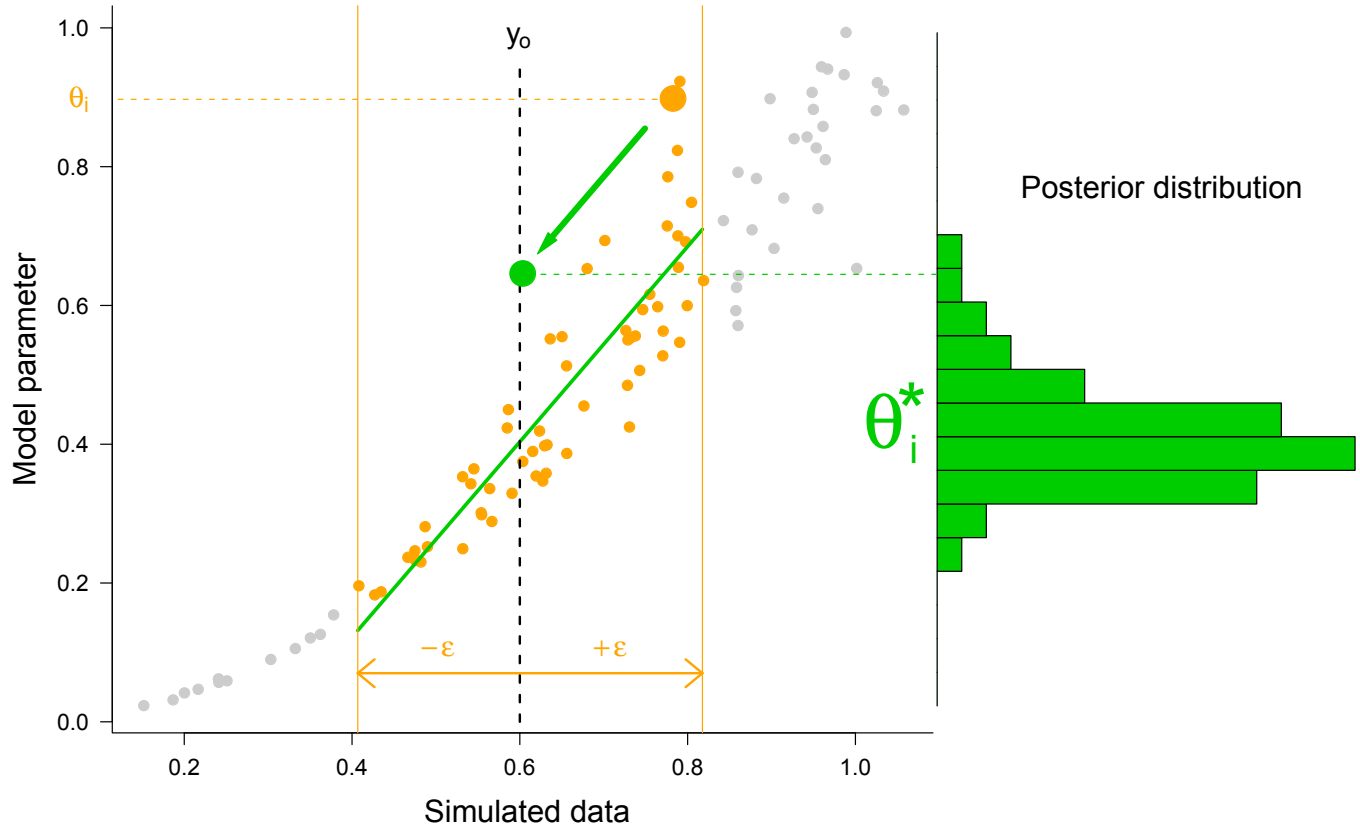
ABC and conditional density estimation (Beaumont et al 2002)

- Let K be the uniform distribution on $(0, 1)$. The empirical distribution is

$$\hat{p}_\epsilon(\theta \mid y) \propto \frac{1}{n\epsilon} \sum_{s=1}^{n} K\left(\frac{|y_s - y|}{\epsilon}\right) \delta_{\theta_s}.$$

- Other kernels can be considered: Epanechnikov, Gaussian, etc
- A second kernel can be used to produced an estimate of the conditional density

Regression adjustment on posterior estimates



Statistical models of the relationship between y_s and θ can be used to correct the discrepancy between the simulated and observed data.

Regression adjustment

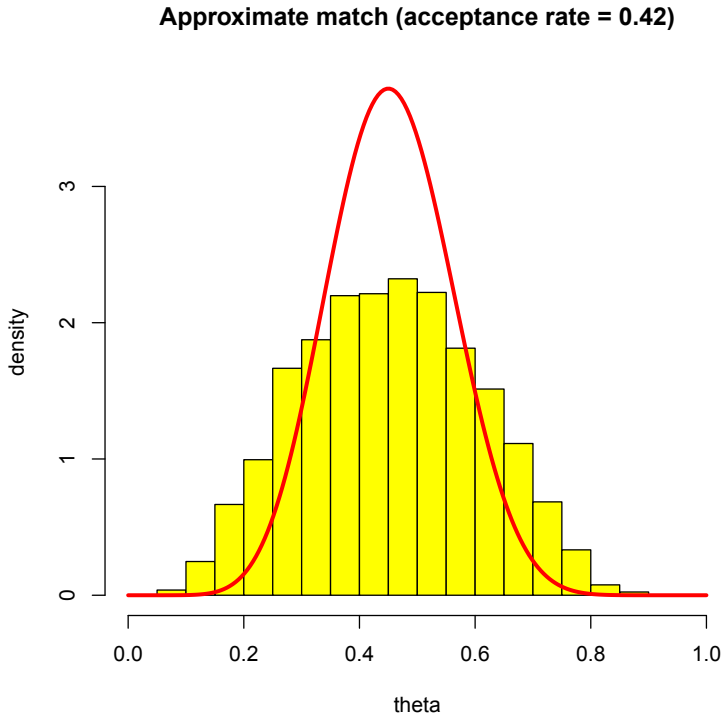
- Linear model

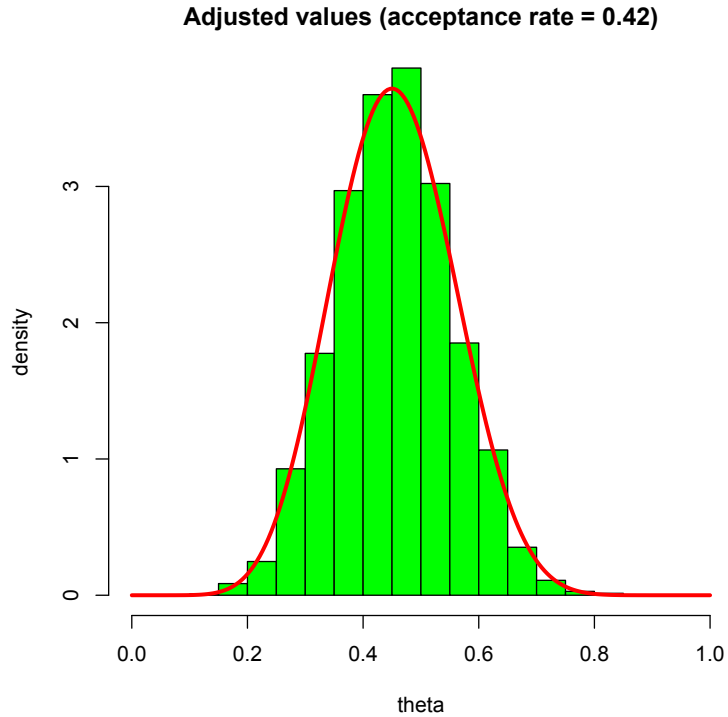
$$\theta_s = \alpha + \beta(y - y_s) + \xi_s\,, \quad s = 1, \ldots, N_\epsilon$$

  where the (y_s) check the condition $(|y - y_s| < \epsilon)$.
- Fit the linear model and correct θ as follows

$$\theta_s^* = \hat{\alpha} + \xi_s = \theta_s - \hat{\beta}(y - y_s)$$

Adjusted values



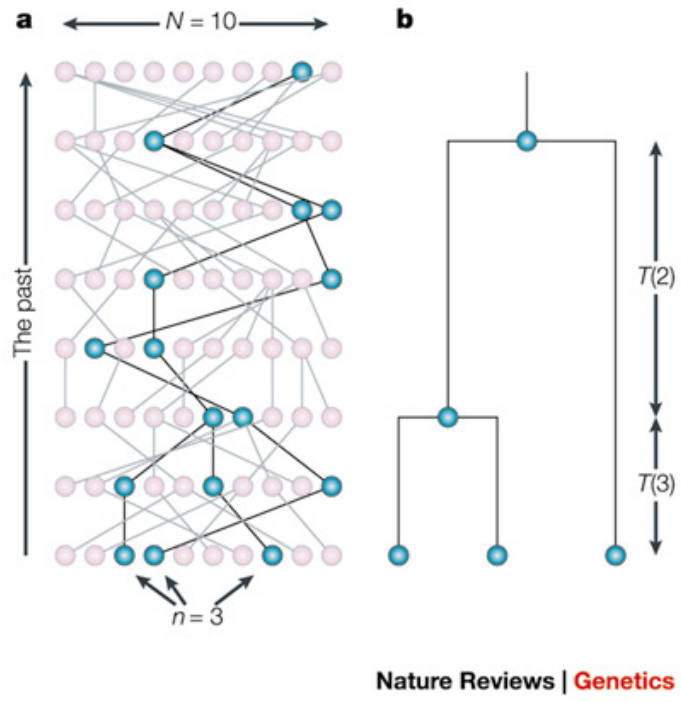


- Regression adjustments can significantly improve the accuracy of the approximation

Theoretical results

- Blum (2010) obtained the asymptotic bias and variance of the estimators of the posterior distribution based on rejection sampling and linear adjustment as $n_{\text{simu}} \to \infty$ (and $\epsilon \to 0$).
- Much remains to be done in order to propose a method (of practical value) for calibrating the tolerance or error rate (open problem!).

Example: Demographic inference in population genetics: Estimating the effective population size



- Coalescence time τ for 2 genes (unit = generation)

$$\Pr(\tau > k) = (1 - 1/N)^k$$

Coalescence time of 2 genes

- Diffusion approximation: If $k = \lfloor tN \rfloor$ and $\tau = \lfloor TN \rfloor$

$$\Pr(\tau > k) = \Pr(T > t) \to \exp(-t)$$

- The coalescent: The number of ancestors in a sample of n genes is a continuous time Markov chain on $\{n, \ldots, 1\}$

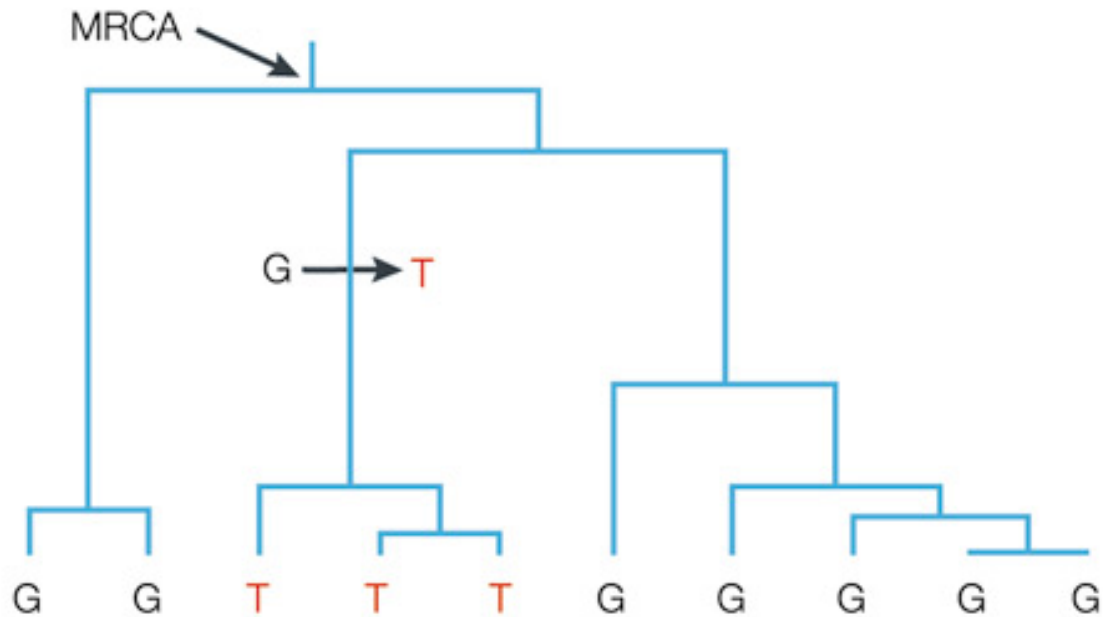
$$\lambda_{i,i-1} = \frac{i(i-1)}{2} \quad i = n, \ldots, 2$$

Mutations

- Mutation rate $\theta = 2\mu N$ where μ is the mutation probability per generation
- The number of mutations in the tree is equal to the number of polymorphic sites in the sample

$$S_n \sim \text{Poisson}(\theta L_n/2)$$

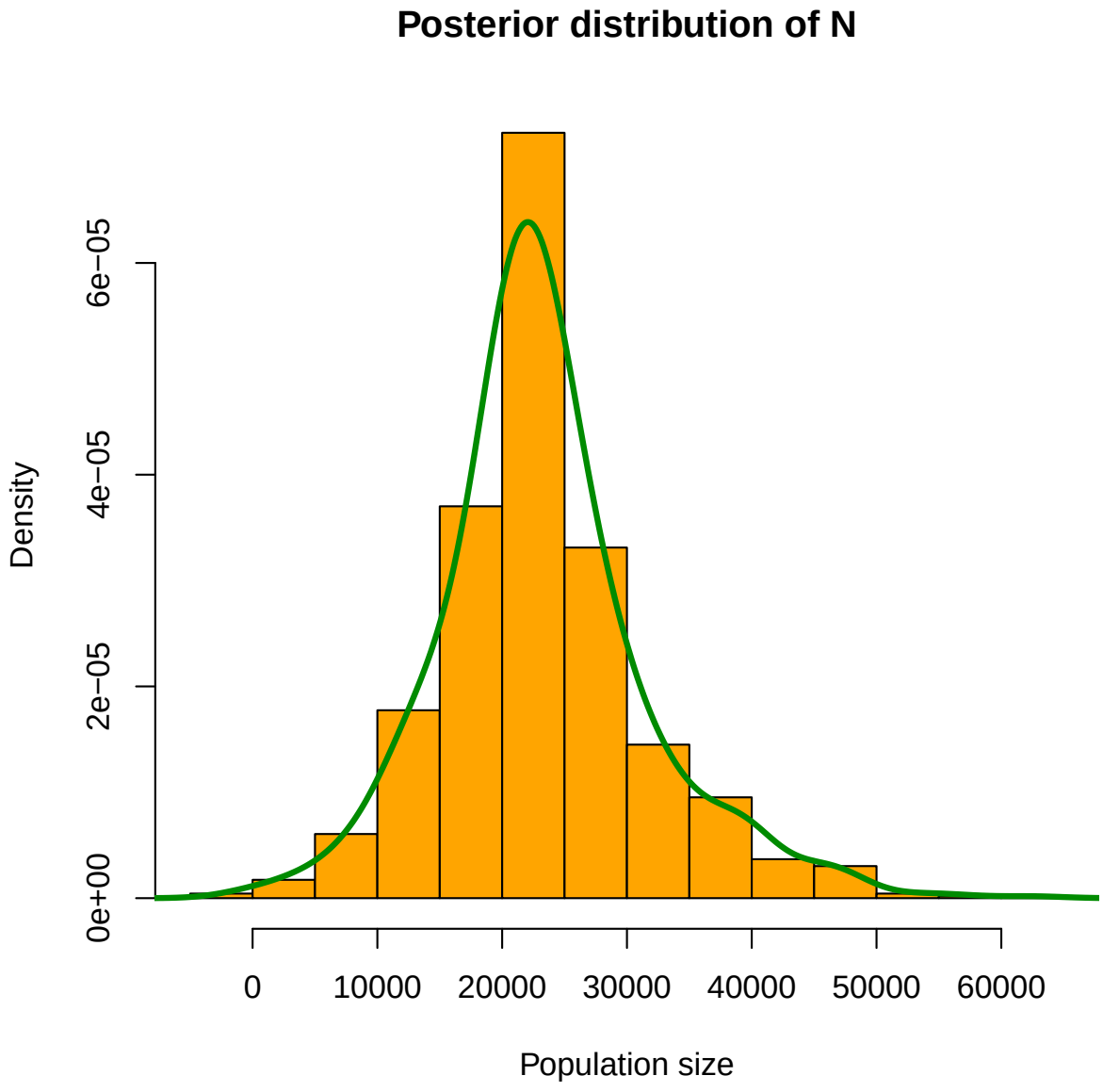
where L_n is the length of the tree for n genes.



Nature Reviews | Genetics

Inference of the effective population size, N

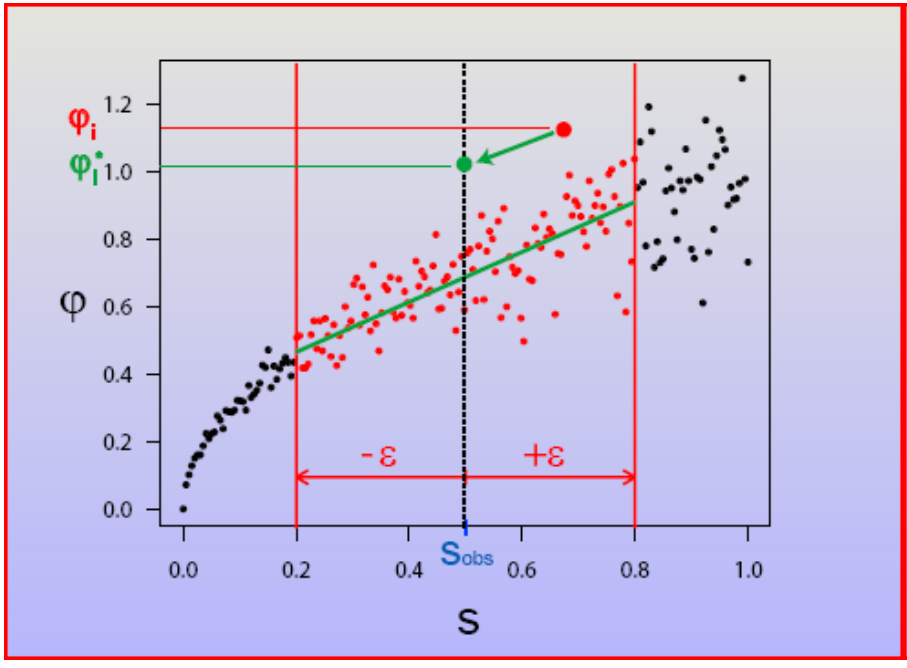
- Data: $n = 100$ non-recombining DNA sequences of length $L = 1000$ bp ($\mu = 10^{-3}$)
- Observation: $S = 21$ polymorphic sites
- Prior distribution on $\theta = 2N\mu$: uniform over the interval (0,100)
- `ms` command line
  ms 100 10000 -t tbs $<$ theta | samplestats $>$ results.txt

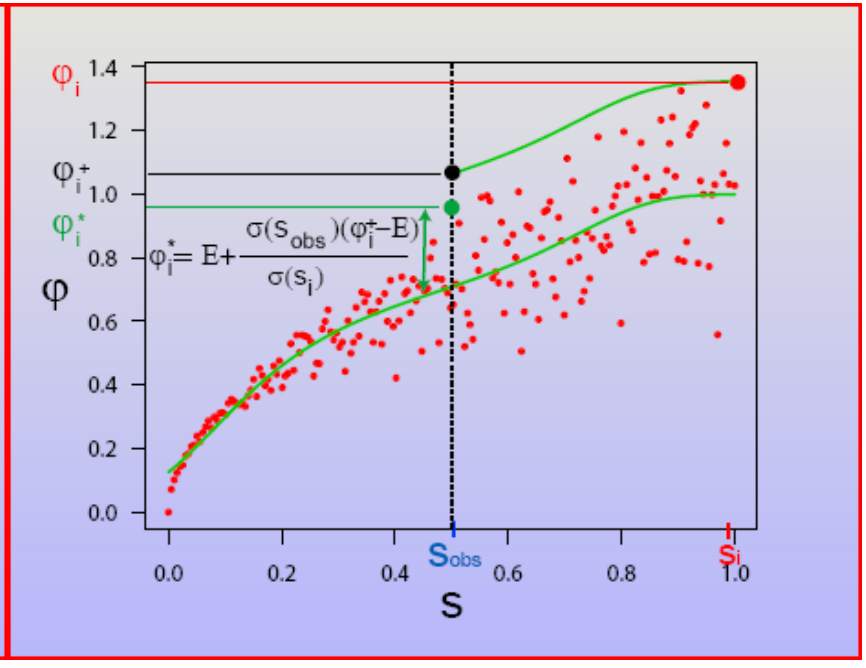

Posterior distribution of N
Density
0e+00
2e−05
4e−05
6e−05
0
10000
20000
30000
40000
50000
60000
Population size

Non-linear regression and dimension reduction

- y can be of large dimension and rejection ABC methods suffer from the curse of dimensionality
- ABC is wasteful. A typical usage may involve millions of simulations, and it retains only a few thousands.
- Recycle the wasted data by using non-linear adjustment (Blum and F 2010)

Non-linear regression and dimension reduction





Reducing the dimensionality of the data

- This can be achieved by many techniques
- GAM, ridge regression, lasso, projection pursuit
- Regularized feedforward neural networks (Ripley 1996)

Regression model

- Feedforward neural nets are 2-layer models
- The first layer performs a non-linear projection on a subspace of lower dimensionality. Build H linear combinations from the sample ($H < d$, d the dimension of the data)

$$s_h = w_{0h} + \sum_{k=1}^{d} w_{kh} y_k, \quad h = 1, \ldots, H$$

- Perform non-linear regression on the H projections

$$\theta = w_{00} + \sum_{h=1}^{H} w_{h0} \phi(s_h) + \text{error} = g_w(s) + \text{error}\,.$$

Regularization and correction

- Weighted (local) regularized least squares criterion

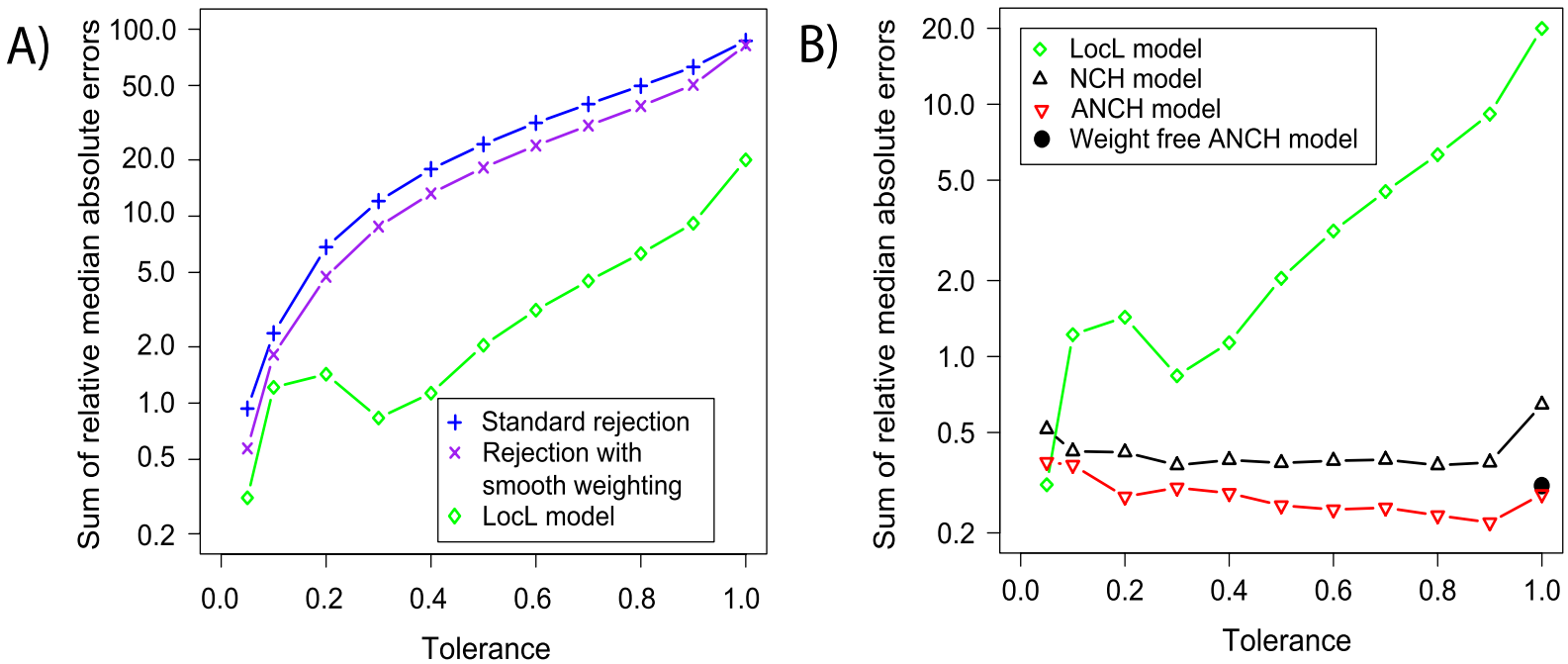
$$\sum_{s=1}^{n}(\theta_s - g_w(\mathbf{s}_s))^2 K_\epsilon(\|\mathbf{s}_s - \mathbf{s}\|) + \lambda\|w\|^2$$

- There is an automatic choice of summary statistics
- Non-linear adjustment

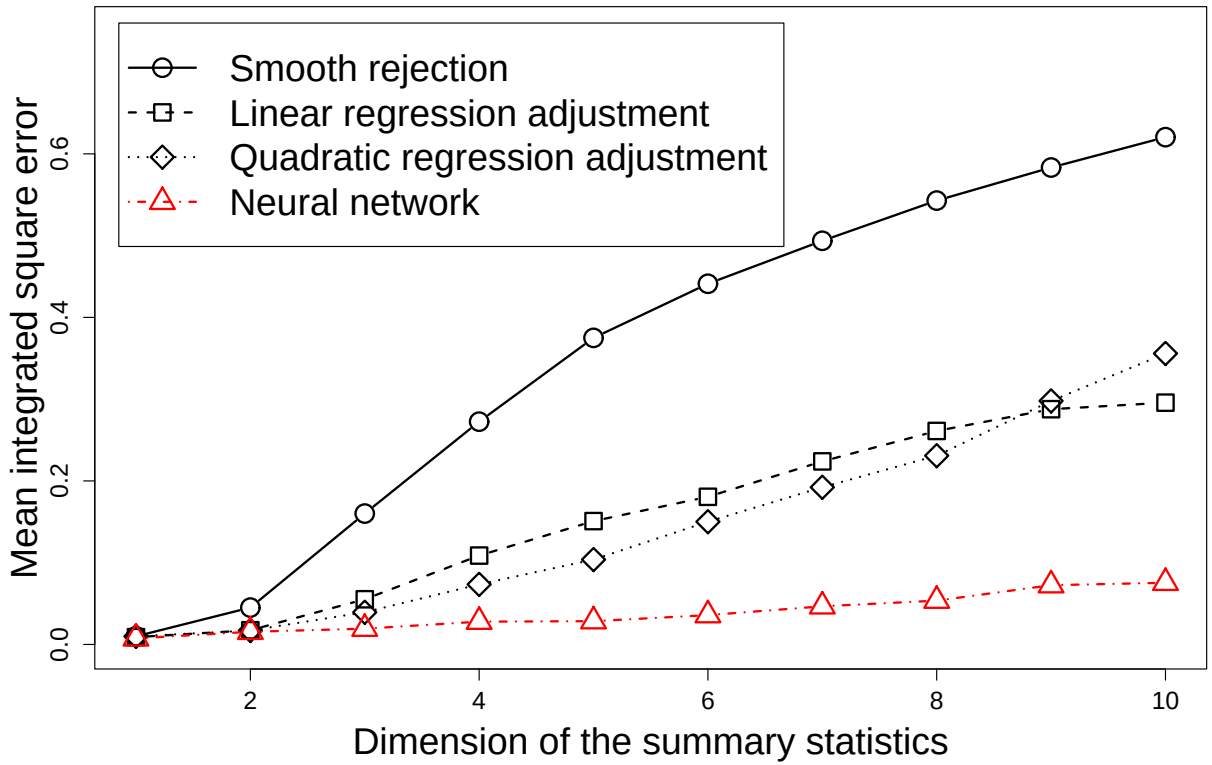
$$\theta_s^* = \theta_s + g_w(\mathbf{s}) - g_w(\mathbf{s}_s)$$

The benefit of using non-linear models

- Back to the population genetic example



Dimension reduction (artificial example)



Part 2. Markov Chains and Sequential Monte Carlo samplers

ABC as a hierarchical model (Marjoram and Tavaré 2003)

- ABC gives "exact" inference under a different model (Wilkinson 2009)

$$\theta \sim p(\theta) \qquad \text{prior distribution}$$

$$y_s|\theta \sim p(.|\theta) \qquad \text{sampling distribution}$$

$$y|y_s \sim K_\epsilon(|y_s - y|) \equiv p(.|y_s) \qquad \text{error distribution}$$

- The parameter ϵ may also be drawn from a prior distribution (Bortot and Sisson 2007)

ABC with MCMC (Metropolis-Hastings)

- An iterative algorithm starting from an initial value θ_1, at time $t = 1$.
- At time t, the algorithm generates a candidate value θ according to $q(\theta|\theta_t)$, where q is some proposal distribution.
- Then it simulates a data set y_s from the sampling distribution with parameter θ, and sets $\theta_{t+1} = \theta$ with probability

$$\alpha = \min\left\{1, K_\epsilon(|y_s - y|)\frac{p(\theta)q(\theta_t|\theta)}{p(\theta_t)q(\theta|\theta_t)}\right\}$$

otherwise $\theta_{t+1} = \theta$.

ABC with MCMC

- ABC-MCMC faces convergence issues (not many applications so far)
- Sequential Monte Carlo samplers, adaptive techniques, iterative importance sampling, offer new horizons
- Sisson (2007) and correctors (Beaumont et al 2010, Toni et al 2009, etc...)

ABC-Population Monte Carlo (Beaumont et al 2010, Toni et al 2009)

Given a decreasing sequence of errors $\epsilon_1 \geq \cdots \geq \epsilon_T = \epsilon$, and samples of parameters $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_N^{(t)})$

- At time $t = 1$, for $i = 1, \ldots, N$, simulate $\theta_i^{(1)} \sim p(\theta)$ and $y_s \sim p(y|\theta_i^{(1)})$ until $\|y_s - y\| < \epsilon_1$.
- Set $\omega_i^{(1)} = 1/N$ and $\sigma_2^2 = 2\mathrm{var}(\theta_i^{(1)})$.

ABC-Population Monte Carlo cont.

- At time $2 \leq t \leq T$, for $i = 1, \ldots, N$. **Repeat**: select $\theta_i^\star$ from the $\theta_j^{(t-1)}$'s with probabilities $\omega_j^{(t-1)}$.
- Generate $\theta_i^{(t)}$ according to $N(\theta_i^\star, \sigma_t^2)$, $y_s \sim p(y|\theta_i^{(t)})$ **until** $\|y_s - y\| < \epsilon_t$.
- Set

$$\omega_i^{(t)} \propto \frac{p(\theta_i^{(t)})}{\sum_{j=1}^N \omega_j^{(t-1)} N(\theta_j^{(t)}; \theta_j^{(t-1)}, \sigma_t^2)}$$

and $\sigma_{t+1}^2 = 2\mathrm{var}(\theta^{(t)})$ (minimize the KL divergence between the target and the proposal).

Remarks

Importance sampling (cf Beaumont et al 2010)

- Let $\hat{p}_t(\theta^{(t)}) = \sum_{j=1}^{N} \omega_j^{(t-1)} N(\theta_j^{(t)}; \theta_j^{(t-1)}, \sigma_t^2)$.
- Let $I = E[\omega^{(t)} h(\theta^{(t)})]$, we have

$$\begin{aligned} I &\propto \int\int h(\theta^{(t)}) \frac{p(\theta^{(t)})}{\hat{p}(\theta^{(t)})} \hat{p}(\theta^{(t)}) p(\theta^{(t)}|y) p(\theta^{(t-1)}) d\theta^{(t)} d\theta^{(t-1)} \\ &= \text{independent on } \theta^{(t-1)} \end{aligned}$$

Algorithm complexity: $O(N^2 \times T)$

Example

- Population divergence model (IM model) with 4 scaled demographic parameters N_a, N_1, N_2 and τ_{div}
- 12 summary statistics

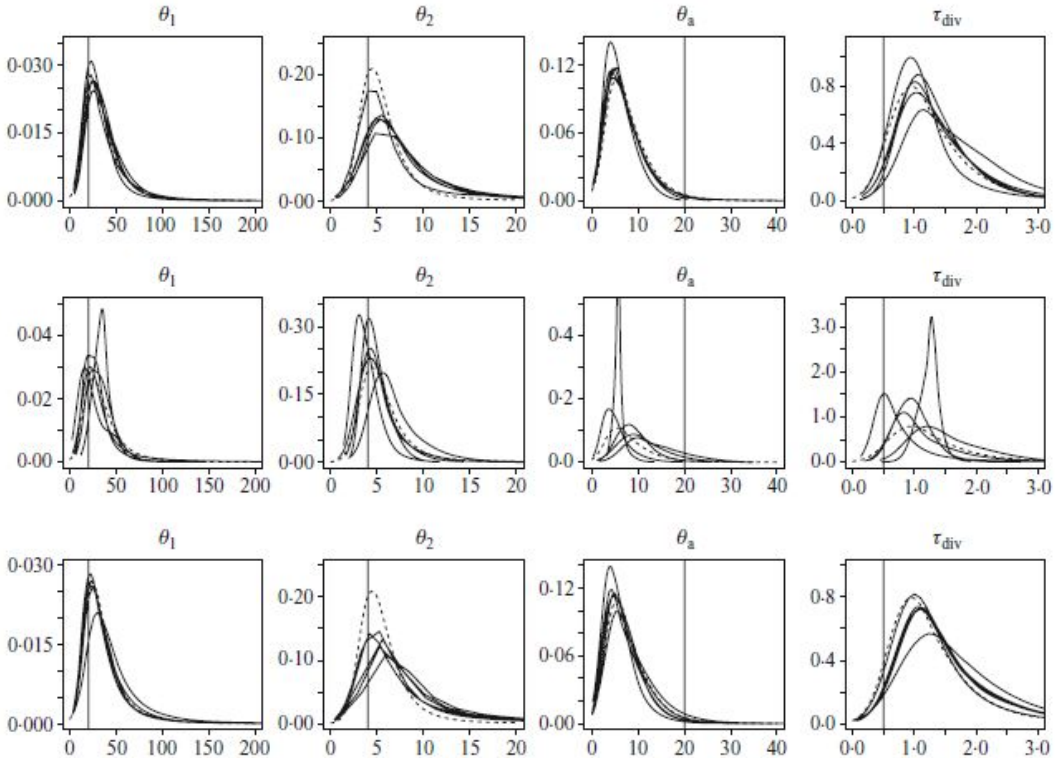


Fig. 2. Variability of the different alternative schemes evaluated through five replicates of the density estimates of the posterior distributions of four identifiable parameters of the population genetics model. First line: population Monte Carlo version; second line: tempered Markov chain Monte Carlo version; third line: Beaumont et al.'s (2002) version of the approximate Bayesian computation algorithm, against the reference posterior, obtained by 5×10^5 simulations in Beaumont et al.'s (2002) version (dotted line). The vertical lines identify the true values of the parameters.

Adaptive ABC-Sequential Monte Carlo (Del Moral et al 2008)

- Step 0. Set $t = 0$. For $i = 1, \ldots, N$, sample $\theta_i^{(0)} \sim p(\theta)$ and $y_{s,i}^{k,0} \sim p(y|\theta_i^{(0)})$.
- Step 1. Set $t = t + 1$. If $\epsilon_{t-1} = 0$ stop, otherwise choose ϵ_t by solving the implicit equation

$$\#\{\text{alive particles at time } t\} = \alpha\#\{\text{alive particles at time } t-1\}$$

where the weights depend on ϵ_t through the update rule

$$\omega_i^{(t)} \propto \omega_i^{(t-1)} \frac{\#\{\text{accepted samples at time } t-1\}}{\#\{\text{accepted samples at time } t\}}$$

If $\epsilon_t < \epsilon$, then set $\epsilon_t = \epsilon$.

Adaptive ABC-Sequential Monte Carlo cont.

- Step 2. If $ESS(\omega^{(t)}) < N_T$, then resample N particles from the vector $(\theta^{(t-1)}, y_s^{(t-1)})$ with weights $\omega^{(t)}$, and reset $\omega_i^{(t)} = 1/N$.
- Step 3. For $i = 1, \ldots, N$, sample

$$(\theta^{(t)}, y_s^{(t)}) \sim Q_{\text{mcmc}}(\theta^{(t-1)}, y_s^{(t-1)}, d\theta dy_s), \quad \text{if } \omega_i^{(t)} > 0$$

  where Q_{mcmc} is the ABC-MCMC transition kernel
- Return to Step 1.

Remarks

- Particular instance of SMC for state-space models (Del Moral et al 2006)
- Relies on ABC-MCMC kernels to move the particles
- Adaptive tolerance levels (prevent the collapse of the SMC approximation)
- Complexity $O(N \times T \times M)$

Concluding messages

- ABC is a simulation-based method to make inference in complex models where the likelihood is hard to compute
- Regression adjustments are essential tricks in practice (R package available)
- SMC-ABC is a promising approach
- ABC is far from being 'easy as 123', but it can be a powerful tool to make inferences with complex models if the steps of Bayesian analysis are carefully applied.