Approximate Bayesian Computation (ABC) as flexible inference methods

Examples of applications from population genetics

Vítor Sousa

Population and Conservation Genetics, IGC
Centro de Biologia Ambiental, FCUL

vitorsousa@igc.gulbenkian.pt

FCUL, 19th March 2010
Sala 6.4.31, Bloco C6

FACULDADE·DE·CIÊNCIAS UNIVERSIDADE·DE·LISBOA

FUNDAÇÃO CALOUSTE GULBENKIAN
Instituto Gulbenkian de Ciência

Outline

- Motivation: inference in population genetics
- Principles of ABC
- ABC performance under "toy" examples
- ABC under "real" population genetics model-based inference problems
- Conclusions and future directions

Genetic Data

Allele frequencies in different points in space (populations)

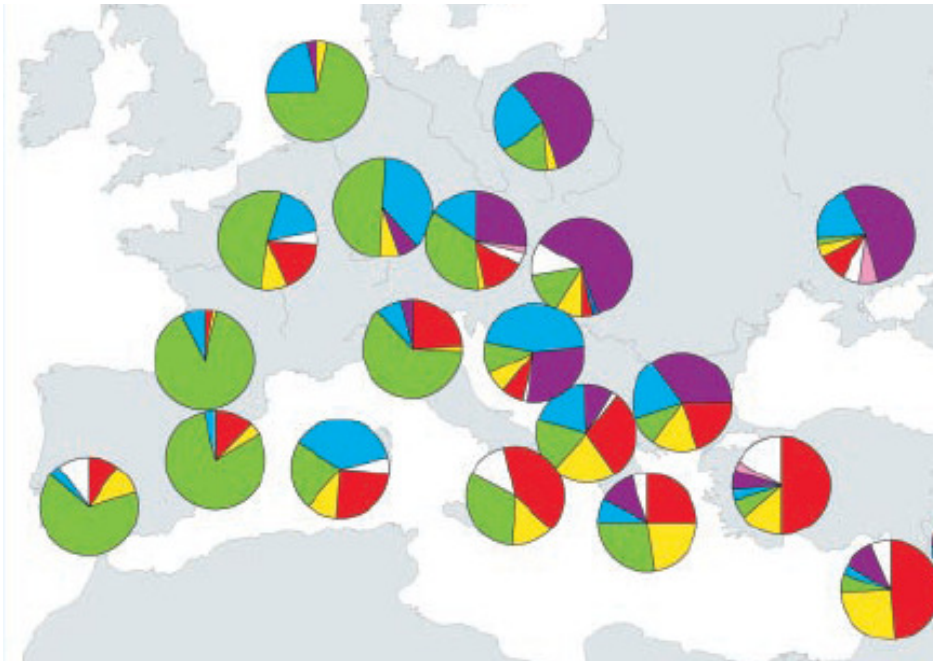
Different regions on the genome (loci)

NEUTRAL markers

(DNA sequences, microsatellites, SNPs, etc.)

Can we infer the demographic history?

- Population sizes
- Population growth rates
- Migration rates
- Time of past events
- ...



Semino et al. (2000) *Science*

Demographic models in population genetics

We focus on demographic events as:

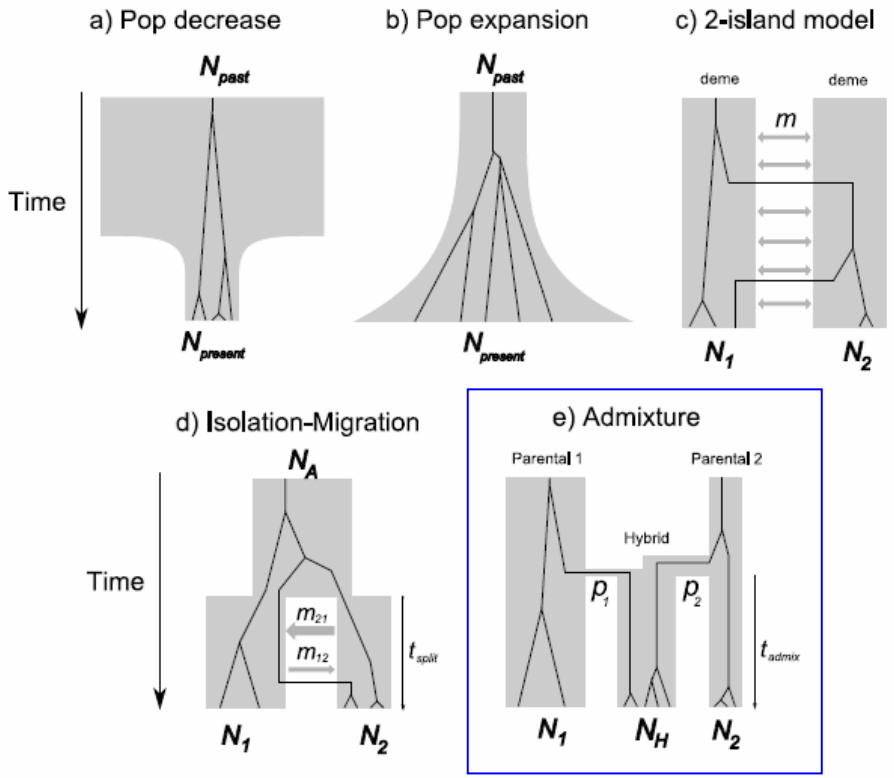
- Population size changes
- Admixture events
- Migration

Problems:

- Likelihood very complex, intractable in many models
- Many parameters, many nuisance parameters

$$P(D|\theta) = \int_{\Omega} P(D|G)P(G|\theta)dG$$

D - observed data (allele frequencies)
θ - demographic parameters (e.g. population size, migration rates, time events, etc.)
G - genealogy of the sample



Inference of demographic history based on genetic data

- **Moment-based methods** (until 1990s)
  - Based on summary statistics
  - Very simple models and difficult to obtain credibility intervals

- **Full-likelihood** methods (since mid 1990s)
  - Based on allele frequency data
  - Computationally intensive – **MCMC and/or Importance sampling**
  - Only used for simple models, and limited datasets

- **Approximate** methods (since late 1990s)
  - Based on summary statistics of the allele frequencies
  - Computationally fast – **ABC rejection**
  - Flexible and easy to apply to complex models, and analysis of large datasets

Marjoram and Tavaré (2006) *Nat Rev Genet*

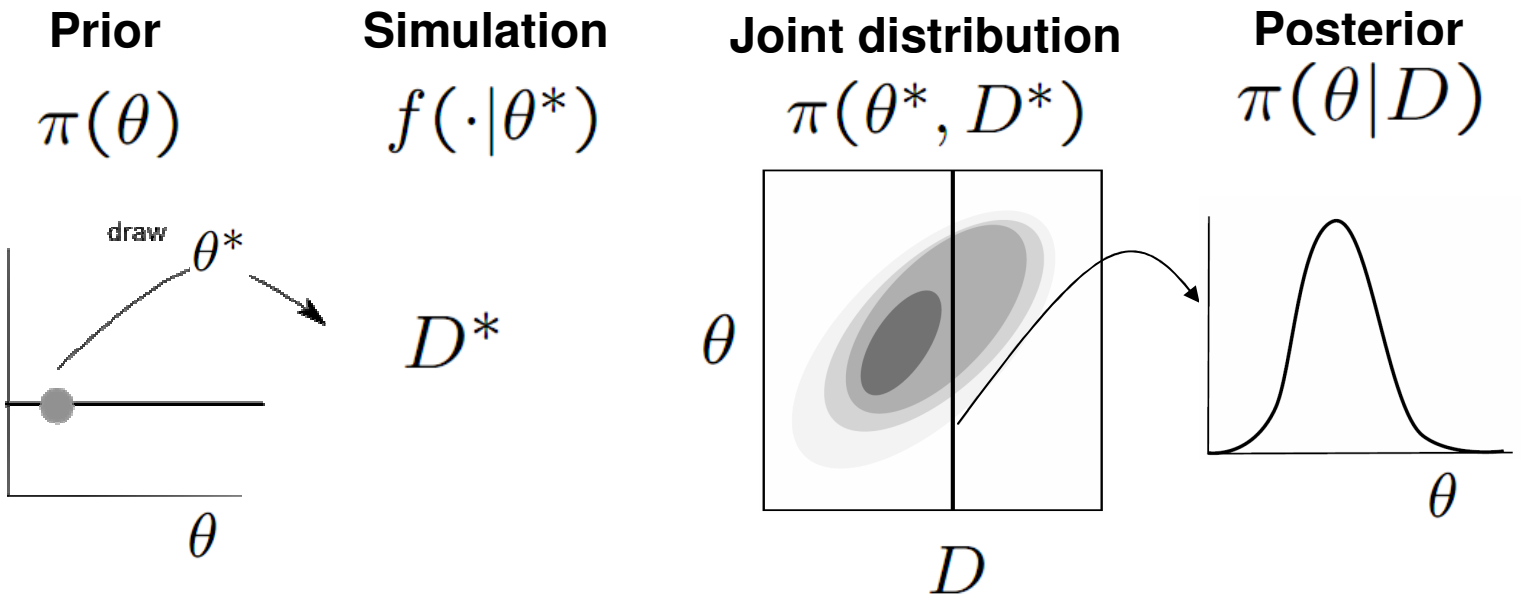
ABC as free-likelihood methods

Exact Rejection Algorithm

Do NOT require explicit likelihood

BUT, we need to be able to simulate data from the model *M*

B1. Generate θ^* from $\pi(\theta)$
B2. Generate D^* from $f(\cdot|\theta^*)$
B3. Accept θ^* if $D^* = D$; return to B1.



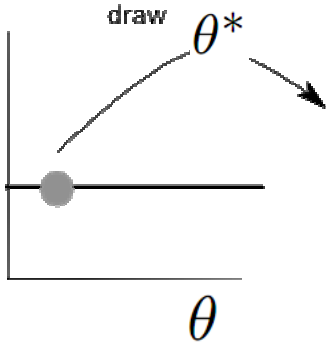
ABC as free-likelihood methods
Tolerance Rejection Algorithm

First approximation:

- Accept parameters only if distance is smaller than a given tolerance treshold

C1. Generate θ^* from $\pi(\theta)$
C2. Generate D^* from $f(\cdot|\theta^*)$
C3. Accept θ^* if $d(D^*, D) < \delta$; return to C1.

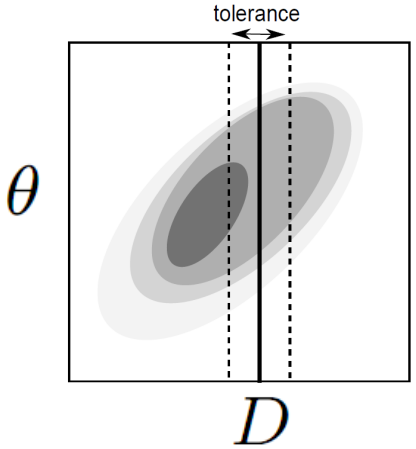
Prior $\pi(\theta)$



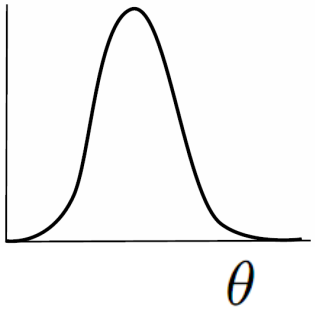
Simulation $f(\cdot|\theta^*)$

D^*

Joint distribution $\pi(\theta^*, D^*)$



Posterior $\pi(\theta|d(D^*, D) < \delta)$



ABC as free-likelihood methods

ABC Rejection Algorithm

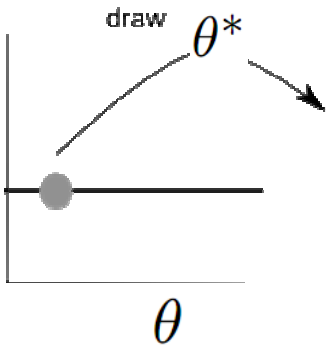
Second approximation:

- Replace data *D* by summary statistics *S*

D1. Generate θ^* from $\pi(\theta)$
D2. Generate D^* from $f(\cdot|\theta^*)$
D3. Compute summary statistics S^* from D^*
D3. Accept θ^* if $d(S^*, S) < \delta$; return to D1.

Prior

$\pi(\theta)$



Simulation

$f(\cdot|\theta^*)$

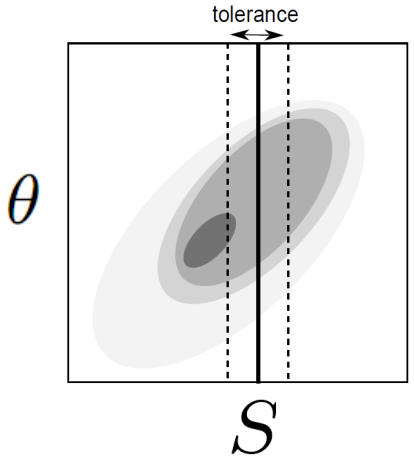
D^*

↓

S^*

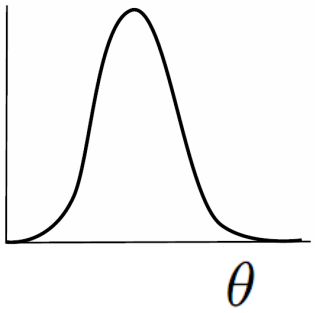
Joint distribution

$\pi(\theta^*, S^*)$



Posterior

$\pi(\theta|d(S^*, S) < \delta)$



ABC Summary

ABC provides independent samples from:

$$\pi(\theta|d(S^*, S) < \delta) \propto f(d(S^*, S) < \delta|\theta)\pi(\theta)$$

If S is a sufficient statistics, as $\delta \to 0$,
and number of simulations $\to \infty$,
$\pi(\theta|d(S^*, S) < \delta) = \pi(\theta|D)$

Efficiency of ABC methods depends on:

- Characterization of the joint distribution (number of simulations)
- Summary statistics selected (sufficient?)
- Tolerance level
- Distance metric chosen
- Dimensionality – number of parameters and number of statistics

Approximate Bayesian Computation (ABC) as an exact sampling algorithm

Wilkinson (2008) showed that if the tolerance follows a distribution, instead of taking a fixed value, ABC rejection samples from the exact posterior.

$$\pi(\theta|d(S^*, S) < \delta, \delta \sim \pi(\delta))$$

R. D. Wilkinson. Approximate Bayesian computation (ABC) gives exact results under the asumption of model error. *Arxiv:0811.3355*, 2008.

Post-Adjustment methods
Regression as conditional density estimation

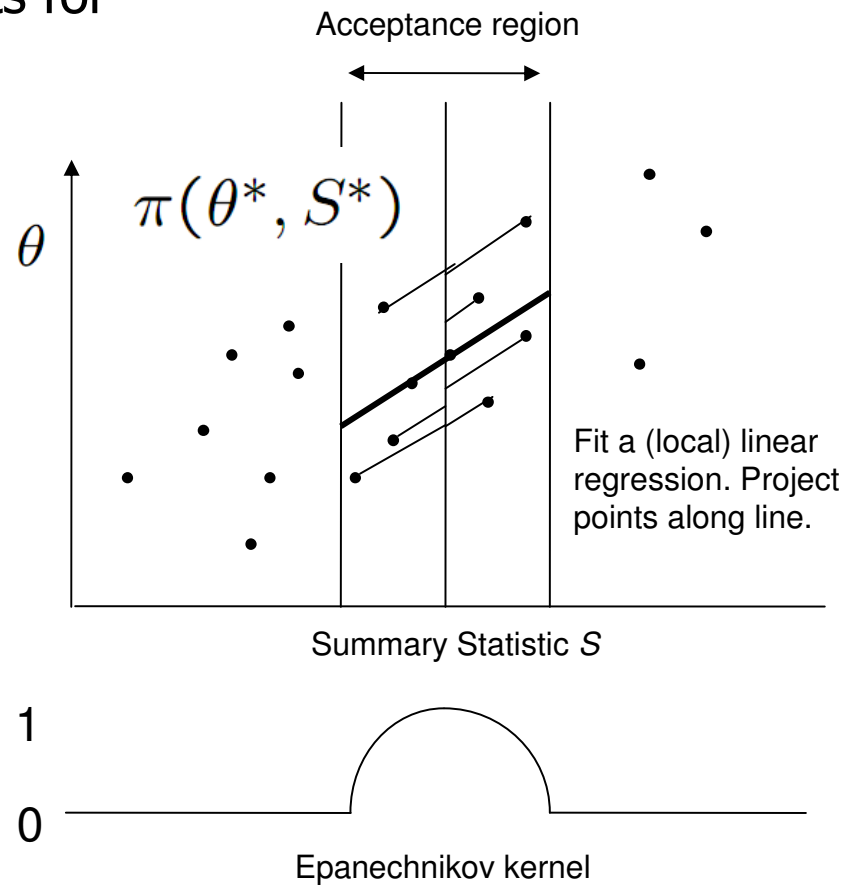
Weighted local linear regression corrects for the discrepancy between *S* and *S**

For each accepted point, we have:

- Parameter value $\boldsymbol{\theta}^i$
- Summary statistic S_i
- Distance measure $(S_i - S(\mathbf{x}^0)$

$$\boldsymbol{\theta}^i = \alpha + (S_i - S(\mathbf{x}^0))'\beta .$$

Least-squares procedure



Beaumont et al. (2002) Genetics

ABC in “toy” examples

Estimate the probability of sucess of a Binomial sampling distribution

Observed data is $x = 204$, $n = 1000$

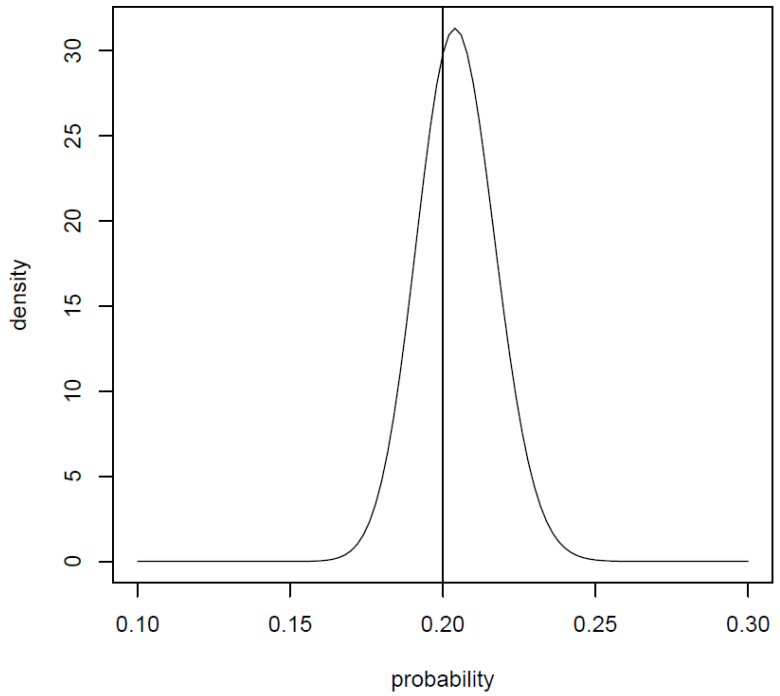
$X \sim Bin(1000, \theta)$

We want to estimate $P(\theta|X = x)$.

Prior $h(\theta) \sim Unif(0, 1) \sim Beta(1, 1)$

In this case, there is an analytical solution

$\pi(\theta|x) \sim Beta(x + 1, n - x + 1)$



ABC in “toy” examples

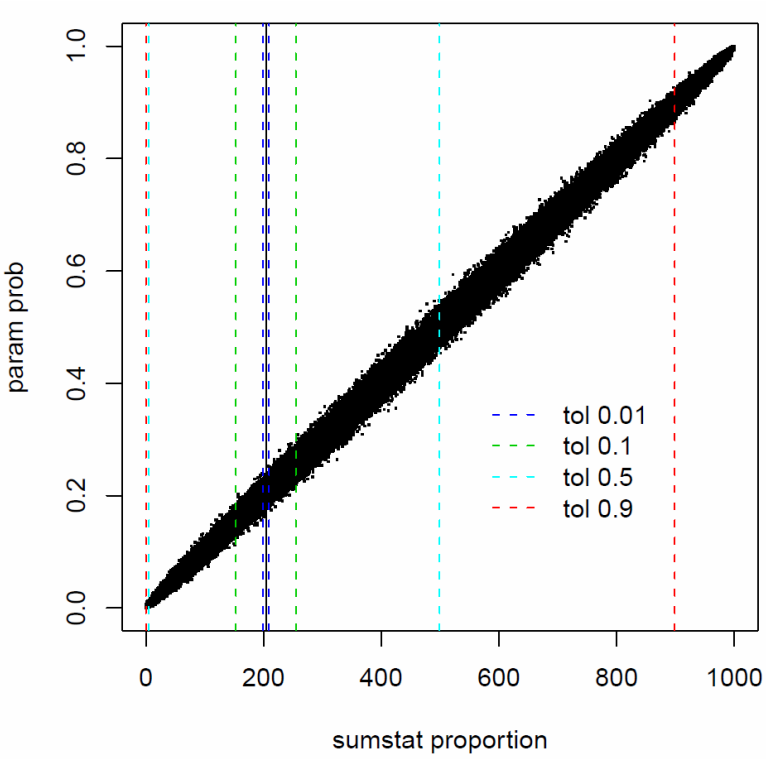
Estimate the probability of sucess of a Binomial sampling distribution

ABC rejection

1. Sample param p^* values from prior
2. Simulate data x^* with param
3. Compute a distance metric $d(x^*,x)=abs(x^*-x)$
4. Accept param p^* that generated datasets with distance smaller than a given tolerance level

Tolerance level defined as a quantile of the distance distribution

Joint distribution obtained with 10^5 simulations



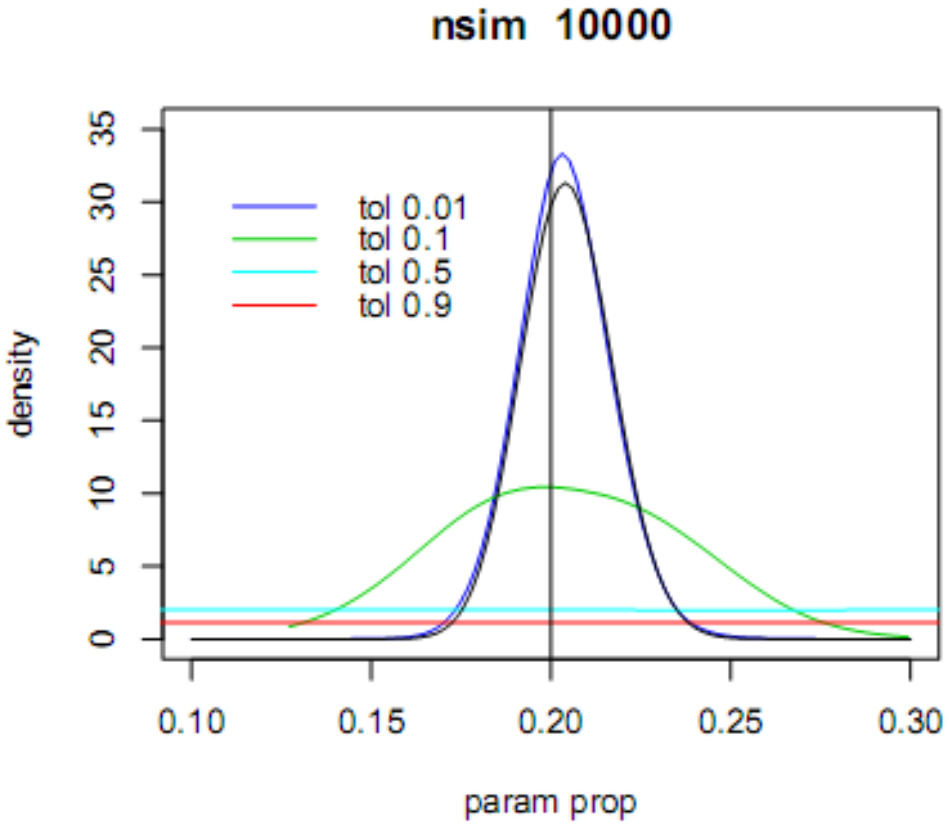
Observed x=204

ABC in “toy” examples

Estimate the probability of sucess of a Binomial sampling distribution

ABC rejection

- Effect of tolerance level
  - As the tolerance level tends to 1, we sample from the prior
  - Decreasing tolerance increases the quality of the approximation to the correct posterior

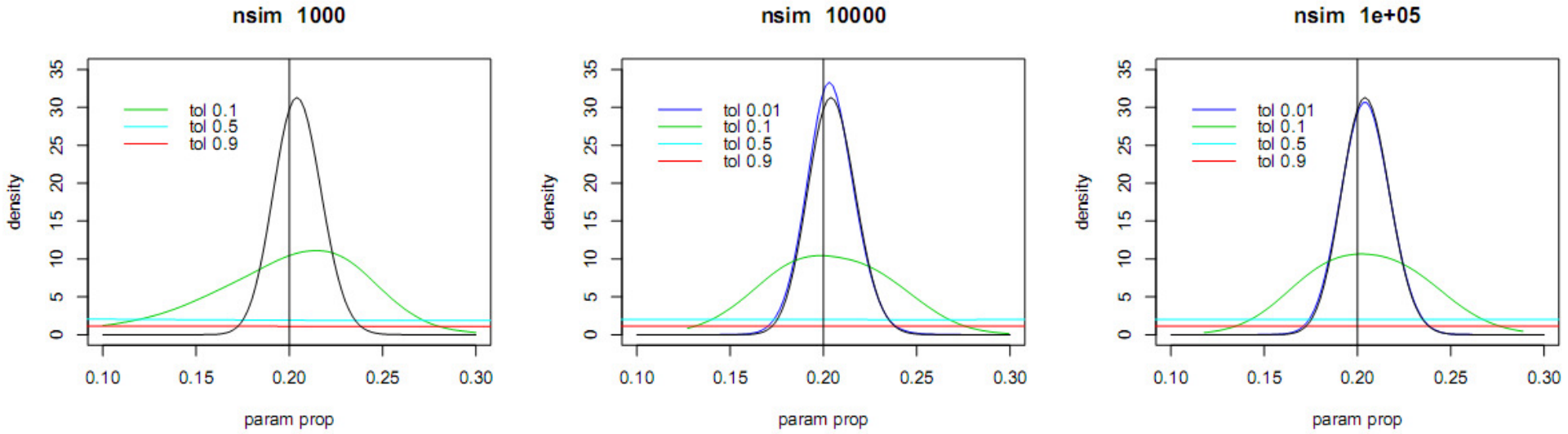


ABC in “toy” examples

Estimate the probability of sucess of a Binomial sampling distribution

ABC rejection

- Effect of tolerance level and number of simulations



- Increasing the number of simulations increases the quality of the approximation

ABC in “toy” examples

Estimate the probability of sucess of a Binomial sampling distribution

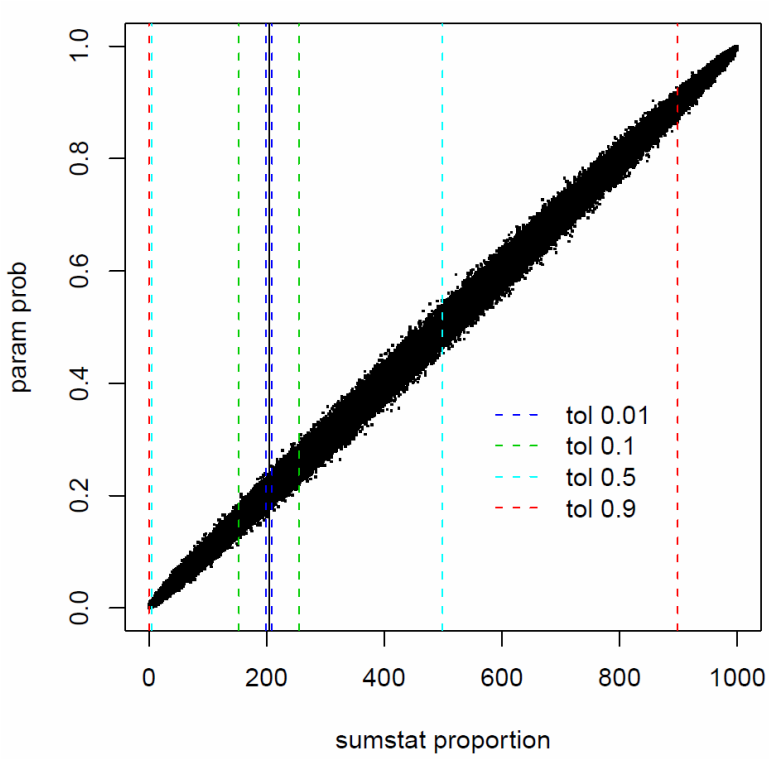
ABC regression

1. Weight the points according to the distance
2. Least-square estimation of **a**, **b**

   $p_{param} = \mathbf{a} + \mathbf{b}^{*}x_{sumstat}$

3. Project points along line of observed data (zero distance)

Joint distribution obtained with 10^5 simulations

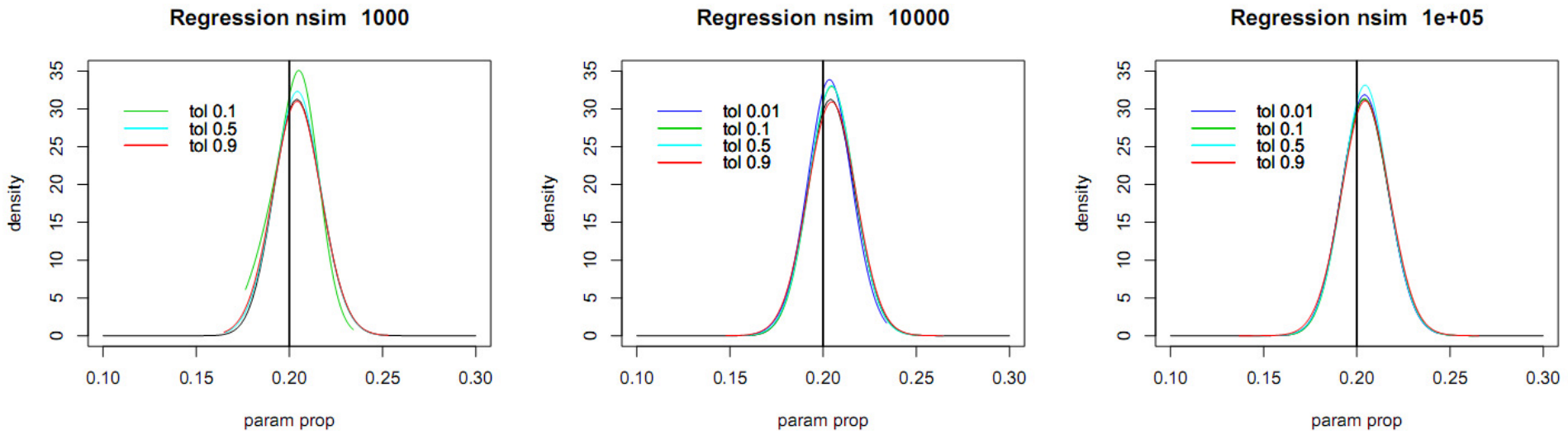


Observed x=204

ABC in “toy” examples

Estimate the probability of sucess of a Binomial sampling distribution

ABC regression



- The regression adjustment decreases the dependency on the number of simulations and tolerance level

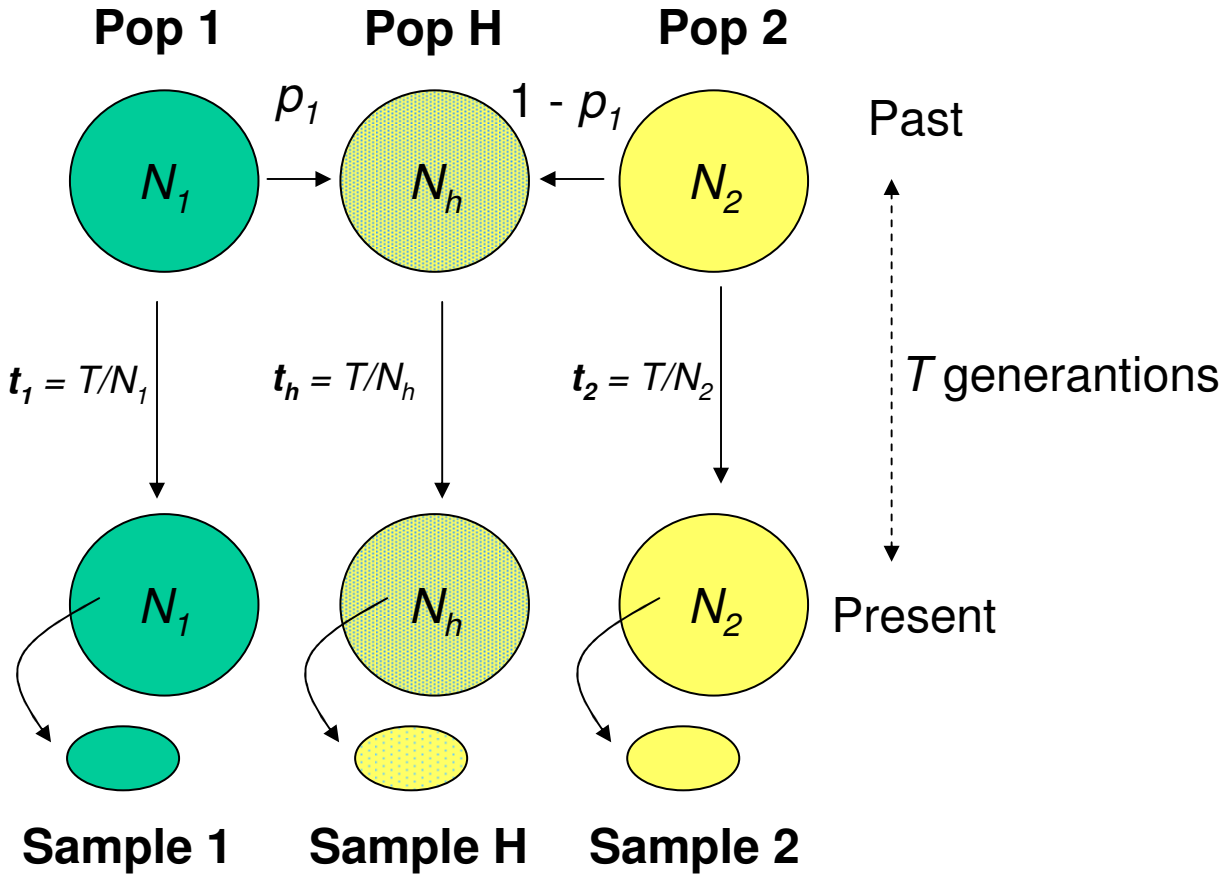
Summary

In practice, in "toy" examples…

- With reasonable number of simulations good approximations
- Sufficient statistics leads to good approximations
- Applying the regression reduces dependence on the tolerance

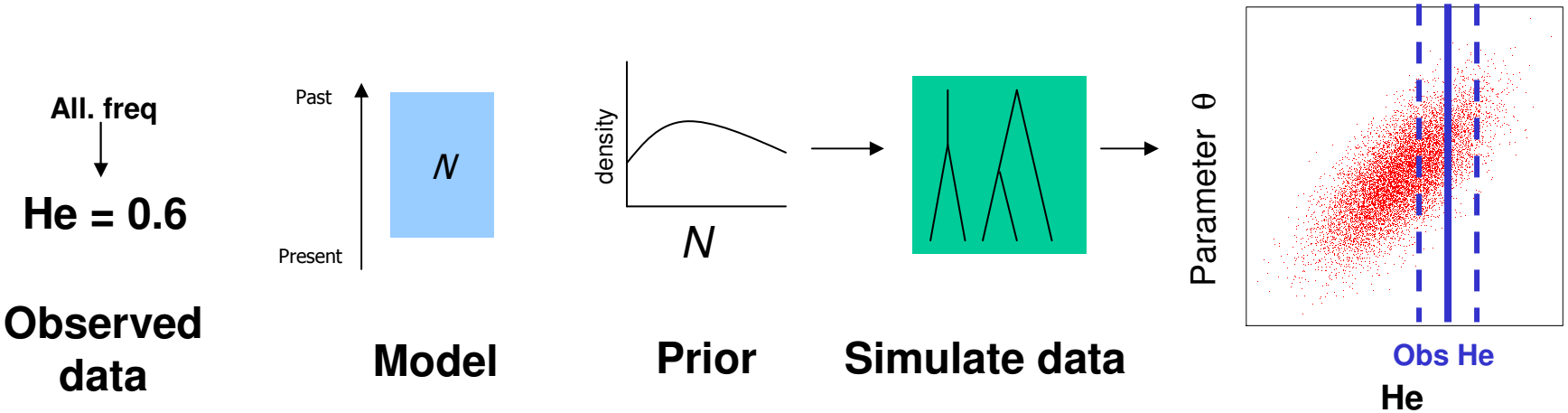
What happens when dealing with complex models?

Estimate Admixture with genetic data



allele A	40	30	35
allele B	10	20	15

Approximate Bayesian Computation (ABC)



Sample θ^* from the **prior**
Simulate data with parameter θ^*
Compute the summary statistics of simulated data $\boldsymbol{S_{sim}}$
Compute a distance between $\mathbf{d} = \boldsymbol{d}(\boldsymbol{S_{sim}}, \boldsymbol{S_{obs}})$
Accept the parameter values if $\mathbf{d} < \delta$

Beaumont et al. (2002) *Genetics*, Marjoram et al. (2003) *PNAS*

ABC in population genetics

How to choose summary statistics?

"The choice of summary statistics is crucial. There is scope for research on practical methods for identifying approximately sufficient statistics"

Marjoram et al. 2003

How informative (or redundant) are the summary statistics?

"Because of the curse of dimensionality there are limitations to the number of summary statistics that can be handled with a reasonable number of simulations"

Beaumont et al. 2002

What is the relative performance compared with a full-likelihood method?

Full-likelihood	ABC sumstat
$P(\theta \mid D)$	$P(\theta \mid d(S_{obs}, S_{sim}) < \delta)$
D = allele frequency data	S = summary statistics

Full-allelic distribution using ABC framework

Full-likelihood	**ABC allele freq**
$\mathbf{P}(\theta \mid \boldsymbol{D})$	$\mathbf{P}(\theta \mid \boldsymbol{d}(\boldsymbol{D}_{\mathbf{obs}}, \boldsymbol{D}_{\mathbf{sim}}) < \delta)$
D = allele frequency data	D = allele frequency data

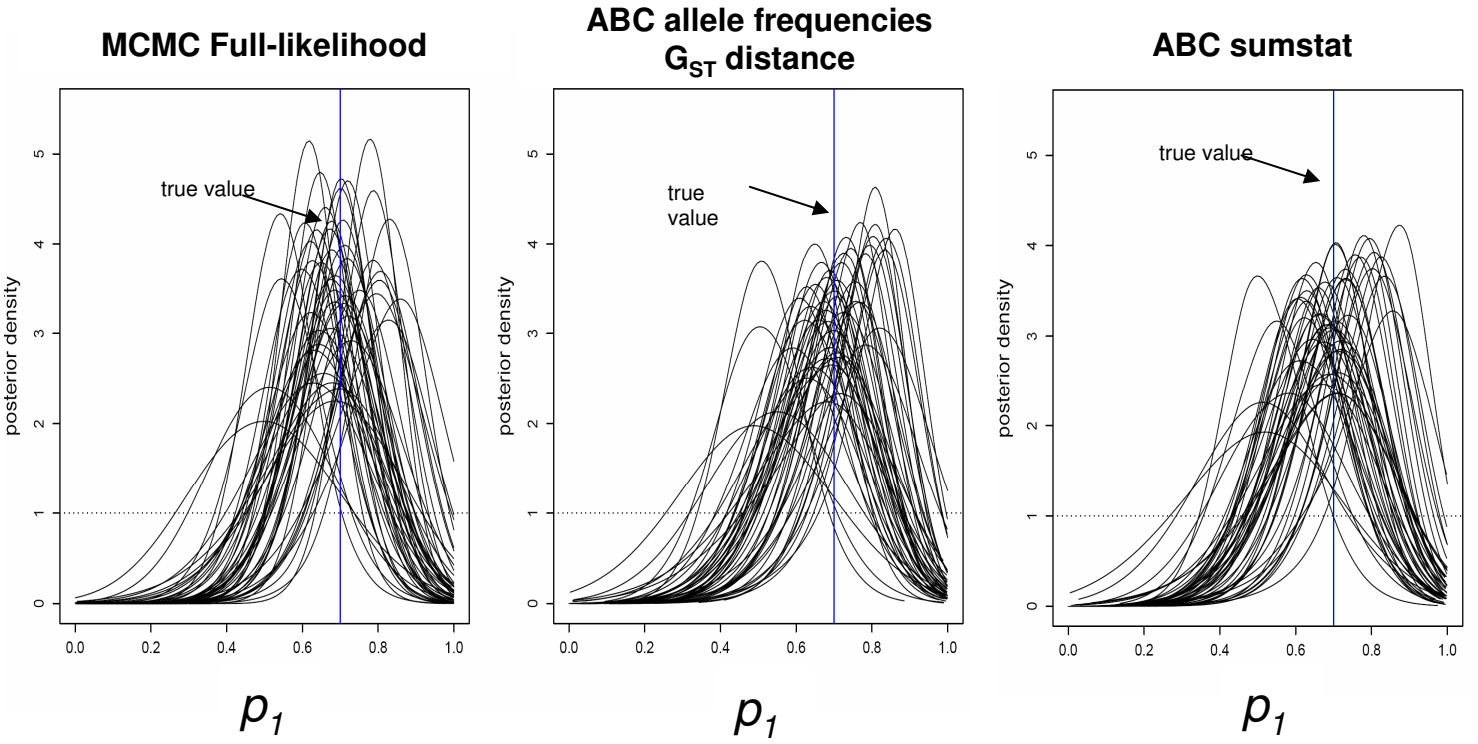
If **Nsim → infinity** and $\delta \rightarrow$ **0**, then the ABC posterior → full-likelihood

However, the problem may become highly dimensional

Sousa et al. (2009) Genetics

Full-likelihood *vs* ABC

10 biallelic loci, t_i =0.01



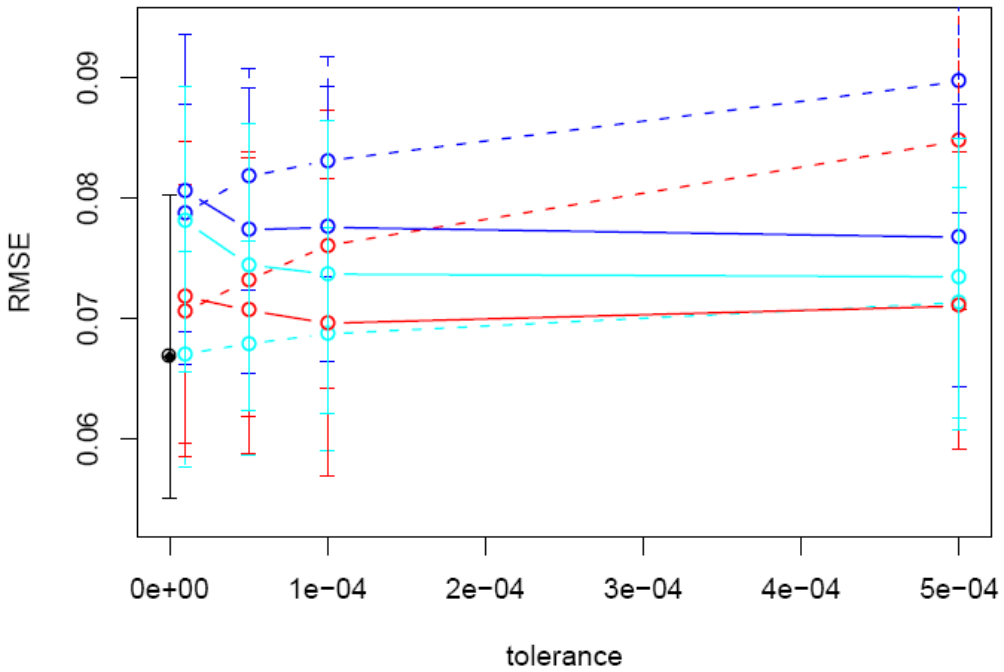
100x10^6 simulations (tolerance level=10^{-5})

- Posterior distributions of ABC approximate the results obtained with LEA
- Posterior distributions of **ABC allele freq** similar to **ABC sumstat**

Sousa et al. (2009) Genetics

Effect of tolerance and Regression

Relative Root Mean Square Error (RRMSE) for point estimates p_1



10 loci drift=0.001
100x10^6 simulations

- Rejection
- Regression
- LEA
- ABC sumstat
- ABC **all freq EUCLIDEAN**
- ABC **all freq GST**

- **ABC returns point estimates similar to LEA**
  - Rejection: error decreases with tolerance
  - Regression: error does not depend on tolerance

Summary

- **ABC provide reasonable estimates, although posteriors with higher variance than with full-likelihood**
- **ABC computationaly faster than MCMC-based methods**
- **Difficult to assess "informative" summary statistics**

ABC *vs* Full-likelihood (LEA)

ABC approximate the results obtained with LEA

Rejection step returns good point estimates

ABC with allele frequencies *vs* ABC sumstat

ABC without sumstat and ABC sumstat provide similar results

Future Directions

Improve the Rejection step:

- Sequential Approaches
- MCMC-ABC without likelihoods
- PCA on the allele frequencies to reduce dimensionality – independent sufficient statistics?

Improve the Regression step:

- Non-linear models for the regression

Conclusions

- ADVANTAGES:
  - ABC provides independent approximate samples from the posterior, instead of correlated samples as in MCMC methods
  - Easy to create databases with the simulated data and corresponding parameters
    - allow analysis of multiple observations, useful to perform simulation studies
  - Possibility to use statistics found in the literature without the need to have the original dataset: Meta-analysis
- DISADVANTAGES:
  - Difficult to select the summary statistics
  - In some cases difficult to “fine tune” the number of simulations, tolerance level, etc.

Other ABC approaches

ABC-MCMC with Metropolis-Hastings algorithm

Algorithm 6. ABC-MCMC algorithm:

1. ***Initialisation:*** *Pick* $\boldsymbol{\theta}^{(0)}$ *from an arbitrary distribution.*

2. ***At iteration*** $i > 1$

   Generate $\boldsymbol{\theta}'$ *from the proposal* $q(.|\boldsymbol{\theta}^{(i)})$ *and* $\mathbf{x}'$ *from* $f(\cdot|\boldsymbol{\theta}')$,

   Calculate the ratio $r(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}') = \min\left(1, \frac{q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i)})} \frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{(i)})} \mathbb{I}_{\{\rho(S(\mathbf{x}^0), S(\mathbf{x}')) < \epsilon\}}\right)$

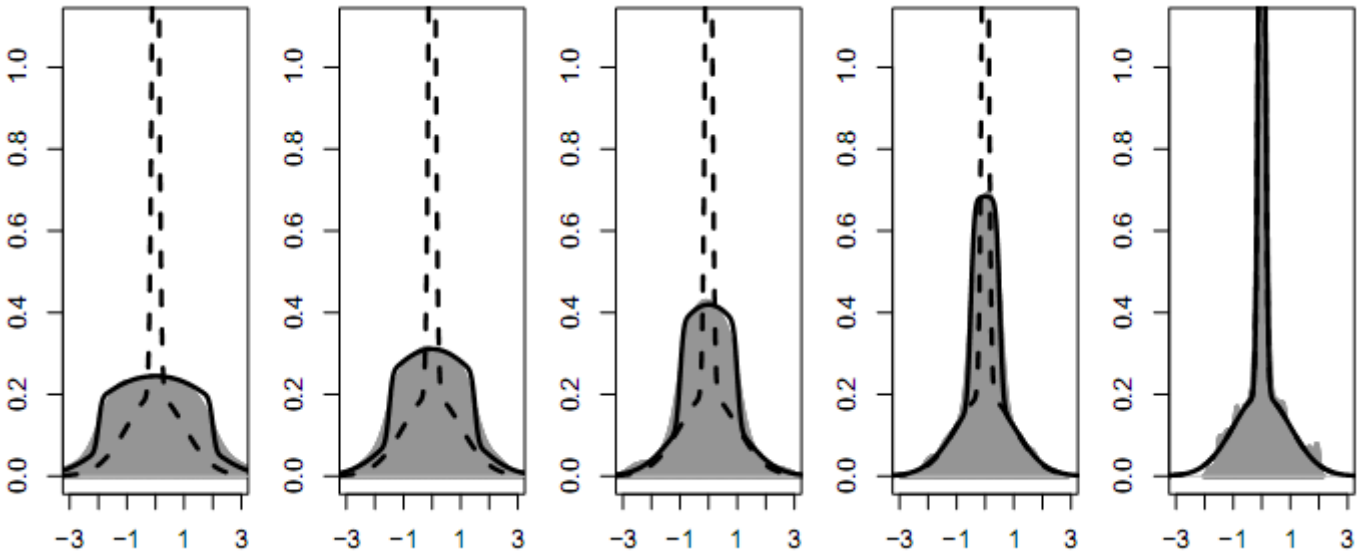
   Accept $\boldsymbol{\theta}'$ *with probability* $r(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}')$, *otherwise stay at* $\boldsymbol{\theta}^{(i)}$. *Go to 2.*

Marjoram et al. (2003) PNAS; Bortot et al. (2008)

P. Bortot, S. Coles, and S. Sisson. Inference for stereological extremes. *arXiv:0811.3355*, 2008.

Sequential ABC approaches

Population Monte Carlo ABC



Beaumont et al. (2010) Biometrika; Sisson et al. (2007) PNAS

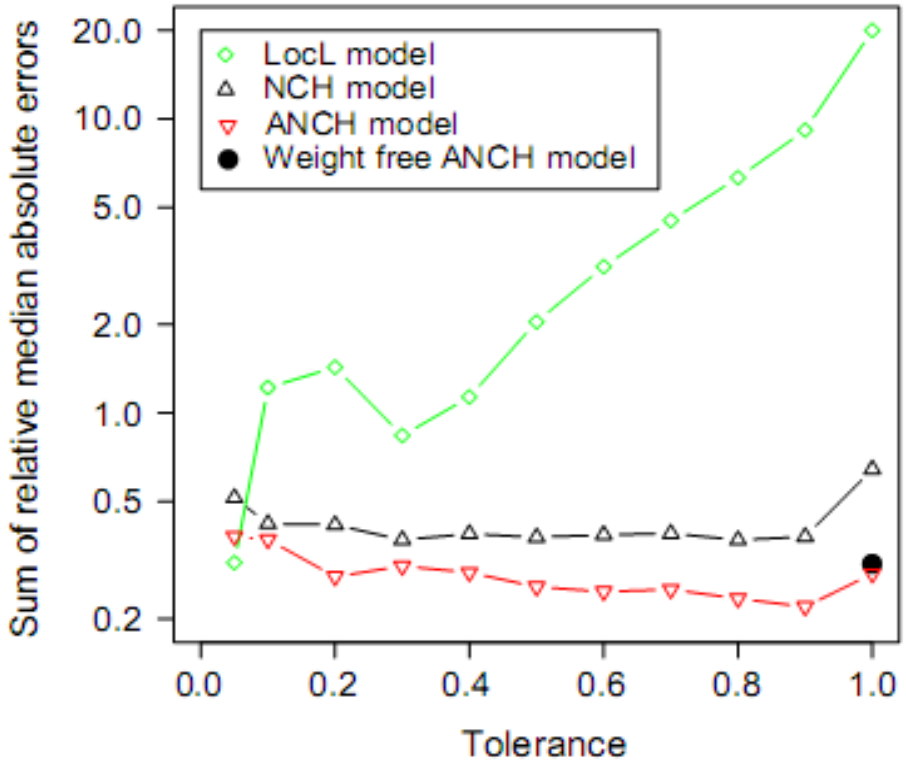
Post-adjustment approaches

Non-linear regressions

LocL – Local linear regression

NCH – Non-linear conditional heteroscedastic model

ANCH – Adaptive version of NCH



- Non-linear regression decrease the dependence in the tolerance
- Decresase computational effort

Blum and François (2008) ArXiv

Acknowledgements

- Lounès Chikhi (PCG Group/CNRS)
- Manuela Coelho (FCUL, Univ. Lisbon)

- Mark Beaumont (Reading Univ., UK)
- Nuno Sepúlveda (IGC)
- Pedro Fernandes (High Performance Computing Centre, IGC)
- Bárbara Parreira (PCG group, IGC)
- João Lopes (Univ. Reading, UK)
- Franck Jabot (Univ. Paul Sabatier, France)
- Aude Grelaud (Rutgers Univ, USA)

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

CENTRO DE BIOLOGIA AMBIENTAL

The University of Reading

FACULDADE·DE·CIÊNCIAS UNIVERSIDADE·DE·LISBOA

FUNDAÇÃO CALOUSTE GULBENKIAN

& Evolution Diversité Biologique EDB

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Population and Conservation Genetics @ IGC

"We are interested in inference methods for complex problems, and we welcome and invite everyone who wants to discuss these problems with us…"

Lounès Chikhi (PI)
chikhi@cict.fr
www.igc.gulbenkian.pt

FUNDAÇÃO CALOUSTE GULBENKIAN
Instituto Gulbenkian de Ciência

References

ABC principles:
Marjoram et al. (2003) MCMC without likelihoods **PNAS** 100: 15324-15328

ABC regression step:
Beaumont et al. (2002) **Genetics** 162: 2025-2035

ABC exact approximation:
Wilkinson (2008) ABC gives exact results under the assumption of sampling error. ArXiv

ABC-MCMC
Marjoram et al. (2003) MCMC without likelihoods **PNAS** 100: 15324-15328

Other
Sisson et al. (2007) Sequential Monte Carlo without likelihoods **PNAS** 104: 1760-1765
Bortot et al. (2008) Inference for stereological extremes. **ArXiv** 0811.3355
Blum (2009) Approximate Bayesian computation: a non parametric perspective. **ArXiv**: 0904.0635
Beaumont et al. (2010) Adptative ABC. **Biometrika** doi:10.1093/biomet/asp052
Blum and François (2010) Non-linear regression models for ABC. **Statistics and Computing** 20: 63-73

R script to estimate probability of a binomial

```
# Estimate the posterior distribution of a proportion
library(locfit)
real_param <- 0.2
n <-1000
# Observed sumstat
obs <- rbinom(1,n,real_param)

par(mfcol=c(2,3))
# Repeat varying the number of simulations performed
nsim <- c(1000,10000,100000)

for(j in 1:length(nsim)) {

        # ABC rejection scheme
        if(nsim[j] <= 1000) {
            tol <- c(0.1, 0.5,0.9)
            col.tol <- c(3,5,2)
        }else {
            tol <- c(0.01, 0.1, 0.5,0.9)
            col.tol <- c(4,3,5,2)
        }

        # Uniform prior on theta
        param <- runif(nsim[j], 0, 1)

        # plot the uniform distribution
        #hist(param)

        # ABC rejection scheme
        sumstat <- rbinom(nsim[j], n, param)

        # Compute the distance between observed and simulated data
        dst <- abs(sumstat-obs)

        # obtain the tolerance distance
        tol_dst <- quantile(dst, tol)

        # Plot the posterior
        #hist(param[dst<tol_dst], prob=T, xlim=c(0.1,0.3))

        # Posterior distribution of interest (obtained using beta function)
        plot(locfit(~param[dst<=tol_dst[1]], xlim=c(min(param[dst<=tol_dst[1]]),max(param[dst<=tol_dst[1]])), alpha=0.99), xlim=c(0.1,0.3), col=col.tol[1], lty=1, ylim=c(0,35),
        xlab="param prop", main=paste("nsim ",nsim[j]))
        for(i in 2:length(tol_dst)) {
            lines(locfit(~param[dst<=tol_dst[i]], xlim=c(min(param[dst<=tol_dst[i]]),max(param[dst<=tol_dst[i]])), alpha=0.99), col=col.tol[i], lty=1)
        }
        curve((1/beta(obs+1,n-obs+1))*(x^obs)*((1-x)^(n-obs)), add=T)
        abline(v=real_param)
        leg.text <- c(paste("tol", tol[1]))
        for(i in 2:length(tol)) leg.text <- c(leg.text, paste("tol", tol[i]))
        legend(0.1, 32, leg.text, lty=c(rep.int(1,length(tol))), col=col.tol, bty="n")

        # Regression step (perform a regression of the param over the sumstat)
        curve((1/beta(obs+1,n-obs+1))*(x^obs)*((1-x)^(n-obs)), xlim=c(0.1,0.3), ylim=c(0,35), xlab="param prop", main=paste("Regression nsim ",nsim[j]), ylab="density")
        for(i in 1:length(tol_dst)) {

            # weights according to the distance
            regwt <- 1-dst[dst<=tol_dst[i]]^2/tol_dst[i]^2
            if(sum(is.nan(regwt))==sum(dst<=tol_dst[i])) regwt <- rep.int(1,sum(dst<=tol_dst[i]))

            # perform the regression step
            fit <- lm(param[dst<=tol_dst[i]]~sumstat[dst<=tol_dst[i]], weights=regwt)

            # obtain the predicted param value for the observed data
            predmean <- sum(coef(fit) * c(1, obs), na.rm=T)

            # add the residuals to the expected values (transpose the points)
            post <- predmean + residuals(fit)

            # Plot
            lines(locfit(~post, xlim=c(min(post),max(post)), alpha=0.99), col=col.tol[i], lty=1)
            abline(v=real_param)
            legend(0.1, 32, leg.text, lty=c(rep.int(1,length(tol))), col=col.tol, bty="n")
        }
        #text(0.15, 25, paste("obs=", obs, "\nout of ", n))
}
```