

# APPROXIMATE BAYESIAN COMPUTATION

Sarah Filippi

Imperial College London

23/07/2019

# SUMMARY OF MY RESEARCH

## Computational statistics for biomedical problems



### Stochastic models for biomedical problems

Understand

- causes of disease,
- mechanisms of their development,
- new treatment

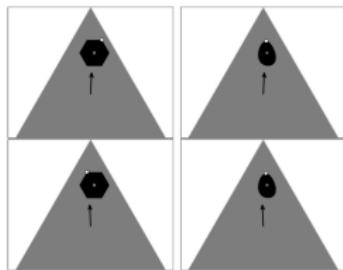
from (potentially large) datasets in  
'omics and clinical studies.

### Statistical machine learning methodological development

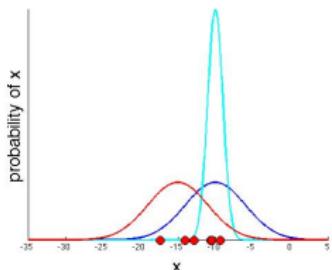
Develop methods that are

- scalable,
- robust
- flexible,
- take into account uncertainty in data-generating process.

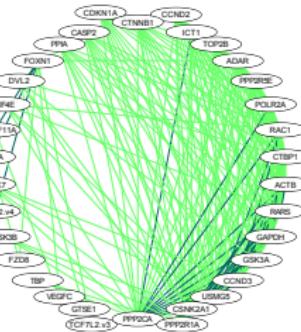
# SOME OF MY RESEARCH INTERESTS



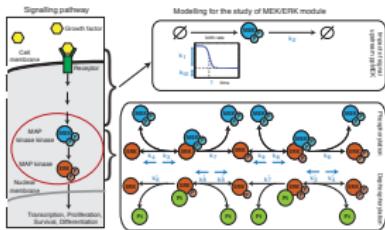
Decision process under uncertainty



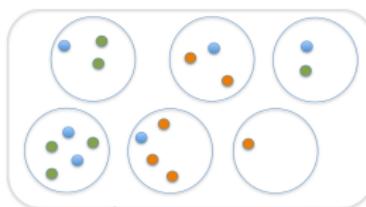
Bayesian (non-parametric) inference procedure



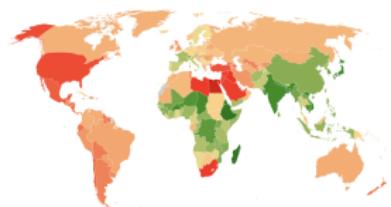
Measures of association and causal inference



Systems biology for biomedicine

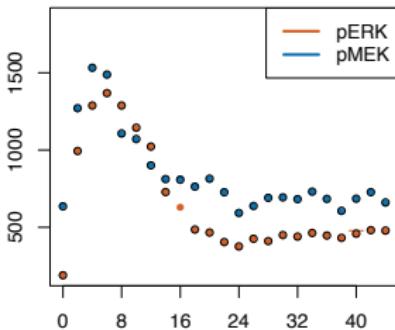
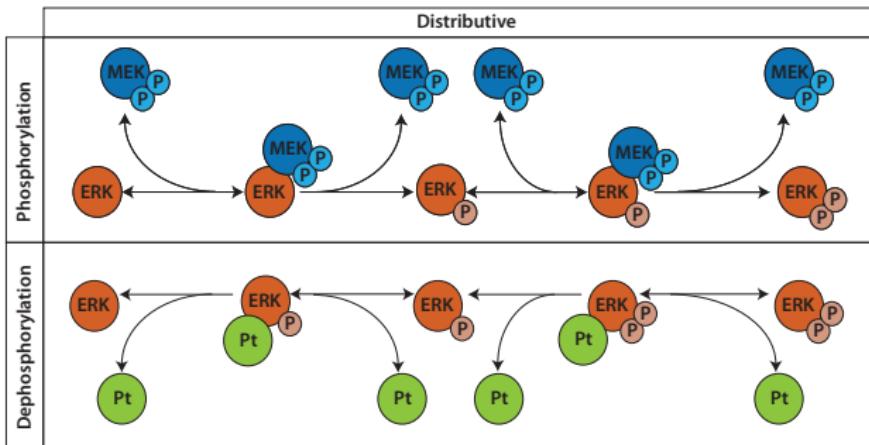
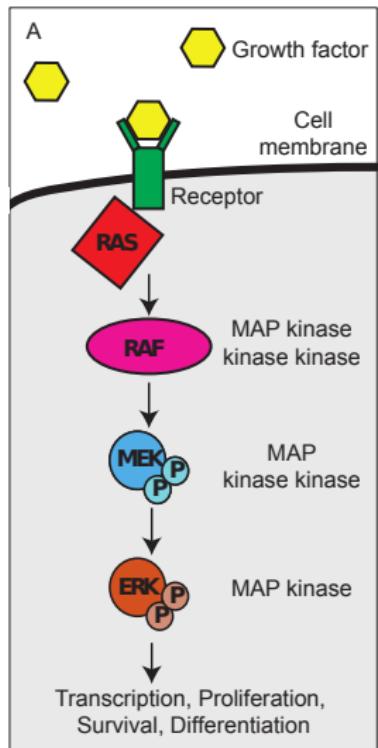


Single-cell genomics and proteomics



Global health

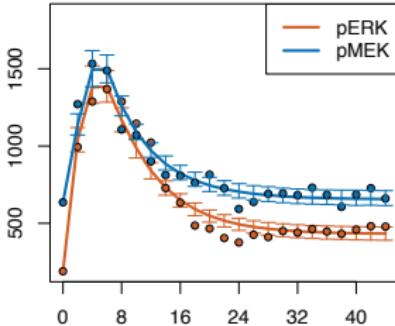
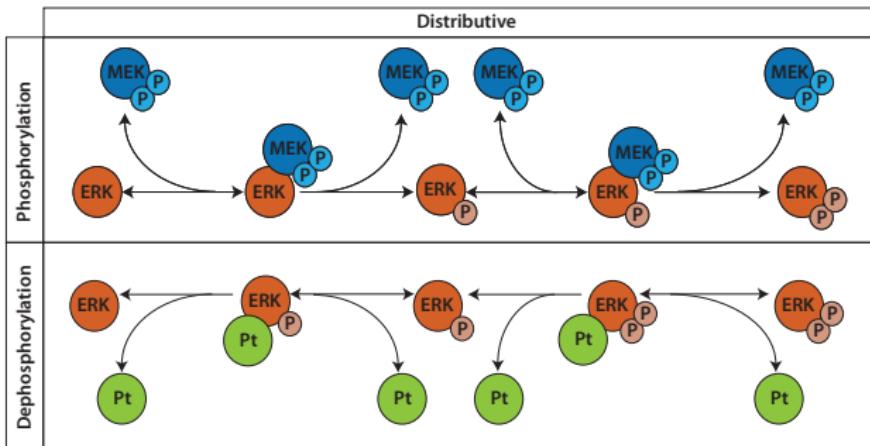
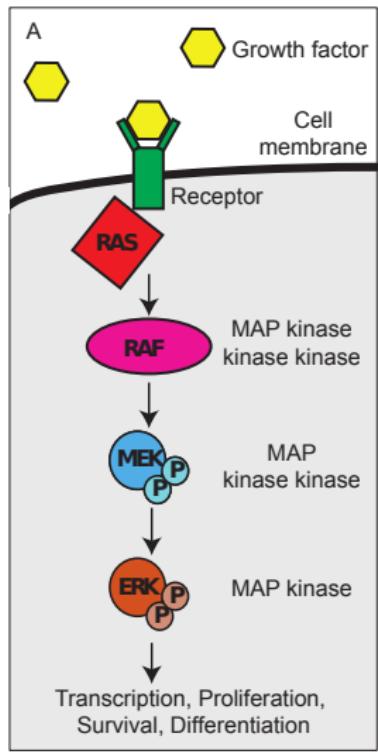
# PARAMETER INFERENCE IN A SIGNALLING PATHWAY



Model via differential equations

Species	9
Parameters	16
Initial conditions	2

# PARAMETER INFERENCE IN A SIGNALLING PATHWAY



Model via differential equations

Species	9
Parameters	16
Initial conditions	2

# WHAT IS APPROXIMATE BAYESIAN COMPUTATION?

- Approximate Bayesian Computation (ABC) is a statistical framework for bayesian parameter inference and model selection.
- It only requires to be able to simulate data from a model.
- It is also referred to likelihood-free inference.

# NOTATIONS

We have:

- an observed dataset, denoted by  $x^*$
- a parametric model, either deterministic or stochastic.

Model = data-generating process:

given a value of parameter  $\theta \in \mathbb{R}^d$ , the output of the system is

$$X \sim f(\cdot | \theta)$$

The likelihood function:  $\theta \mapsto f(x^* | \theta)$

Aim:

- **parameter inference:** determine how likely it is for each value of  $\theta$  to explain the data
- **model selection:** rank candidate models in terms of their ability to explain the data

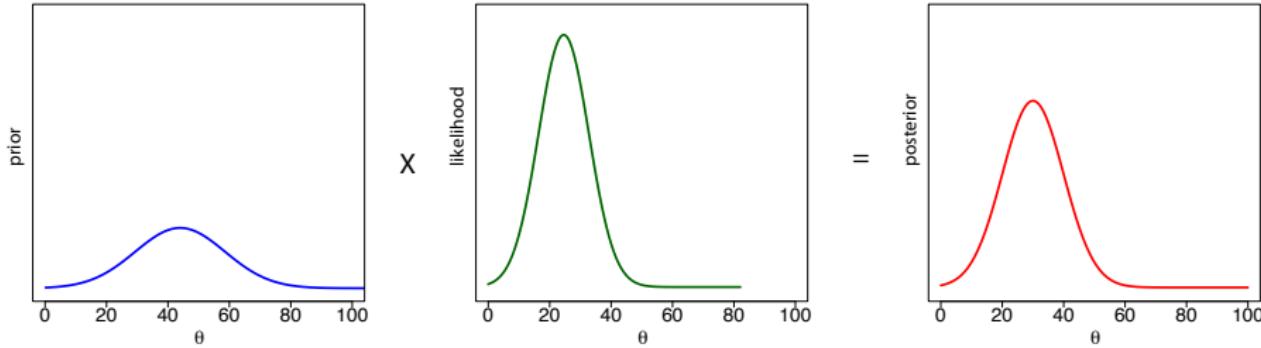
# BAYESIAN METHODS

In the Bayesian framework, we combine

- the information brought by the data  $x^*$ , via the likelihood  $f(x^*|\theta)$
- some a priori information, specified in a prior distribution  $\pi(\theta)$

These informations are summarized in the posterior distribution, which is derived using the Bayes formula:

$$\pi(\theta|x^*) = \frac{f(x^*|\theta)\pi(\theta)}{\int f(x^*|\theta')\pi(\theta')d\theta'}$$



## SAMPLE FROM THE POSTERIOR DISTRIBUTION

- In general, no closed-form solution of the posterior distribution
- Use computer simulation and Monte-Carlo techniques to sample from the posterior distribution
- Typical likelihood-based approaches:
  - Importance sampling
  - Markov Chain Monte-Carlo (MCMC)
  - Population Monte-Carlo
  - Sequential Monte-Carlo (SMC) sampler
  - ...

# WHAT IF WE CAN NOT COMPUTE THE LIKELIHOOD?

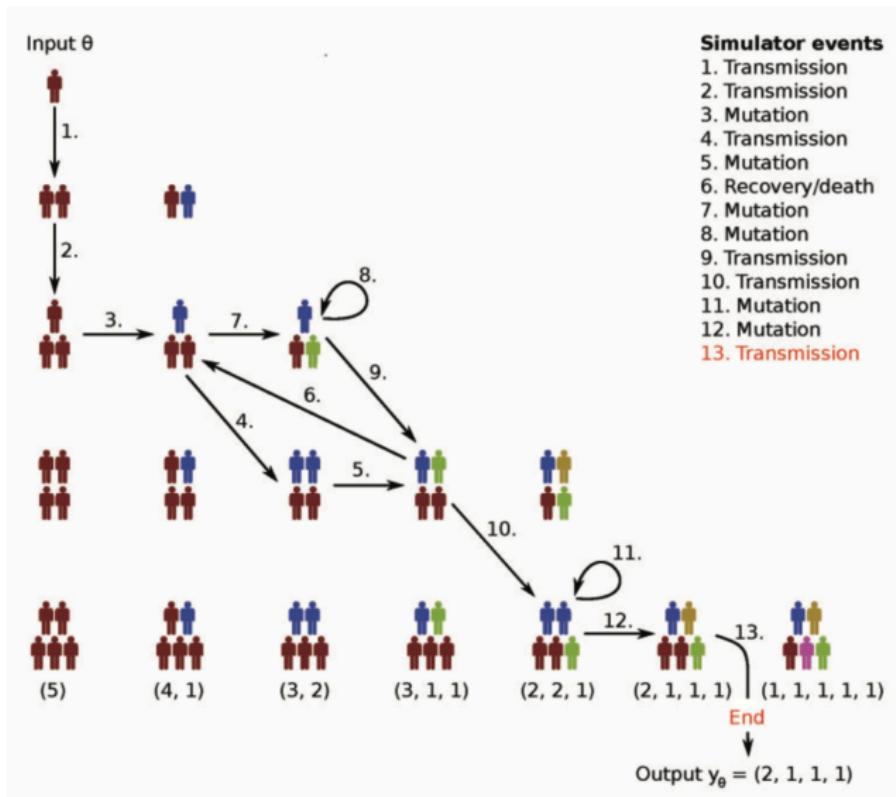
In many applications of interest: either impossible or computationally too expensive to evaluate the likelihood.

Domains of application include

- population genetics and evolutionary biology
- ecology
- infectious disease epidemiology
- astrophysics
- (biological) processes described by stochastic differential equations
- ...

Use so called **likelihood-free** method such as  
**Approximate Bayesian Computation (ABC)**.

# EXAMPLE: THE SPREAD OF TUBERCULOSIS



[Tanaka et al, 2006; Figure credit: Lintusaari et al, 2017]

# WHY IS ABC SO POPULAR?

ABC methods are popular in many areas, particularly in biological disciplines,

Approximate Bayesian Computation is the main tool for parameter inference for a complex simulator for which we can not calculate the likelihood function.

Other advantages:

- Simple to implement
- Intuitive
- Parallelizable

# OUTLINE

## 1 THE ABC METHOD

- The basic principle of ABC
- ABC with summary statistics
- Post-processing of ABC output
- Theoretical results on ABC

## 2 EFFICIENT ABC ALGORITHMS

## 3 ABC FOR MODEL CHOICE

# BASIC COMPONENTS OF ABC

Approximate Bayesian Computation is a statistical framework for bayesian parameter and model selection that only requires:

- Observed data,  $x^*$
- A prior distribution,  $\pi(\theta)$
- A simulator, typically stochastic, to simulate new data  $z \sim f(\cdot|\theta)$
- A distance to compare simulated data and observed data
- A threshold for determining what is similar enough
- Computational resources

# THE APPROXIMATE BAYESIAN COMPUTATION METHOD

## An exact rejection algorithm

Given an observation  $x^*$ , keep jointly simulating

- parameter  $\theta$  from the prior distribution :  $\theta \sim \pi(\theta)$
- auxiliary variables  $z$  from the simulator:  $z \sim f(z|\theta)$

until the auxiliary variable  $z$  is equal to the observed value  $z = x^*$ .

[Tavaré et al, 1997]

The outcome  $\theta^{(1)}, \dots, \theta^{(N)}$  resulting from this algorithm is an i.i.d. sample from the posterior distribution since

$$f(\theta^{(i)}) \propto \int \pi(\theta^{(i)}) f(z|\theta^{(i)}) \mathbb{1}_{\{z=x^*\}} dz = \pi(\theta^{(i)}) f(x^*|\theta^{(i)}) \propto \pi(\theta^{(i)}|x^*)$$

## AN APPROXIMATE METHOD

When  $x^*$  is a continuous random variable, equality  $z = x^*$  is replaced with a **tolerance** condition

$$d(x^*, z) \leq \epsilon$$

where  $d$  is a distance.

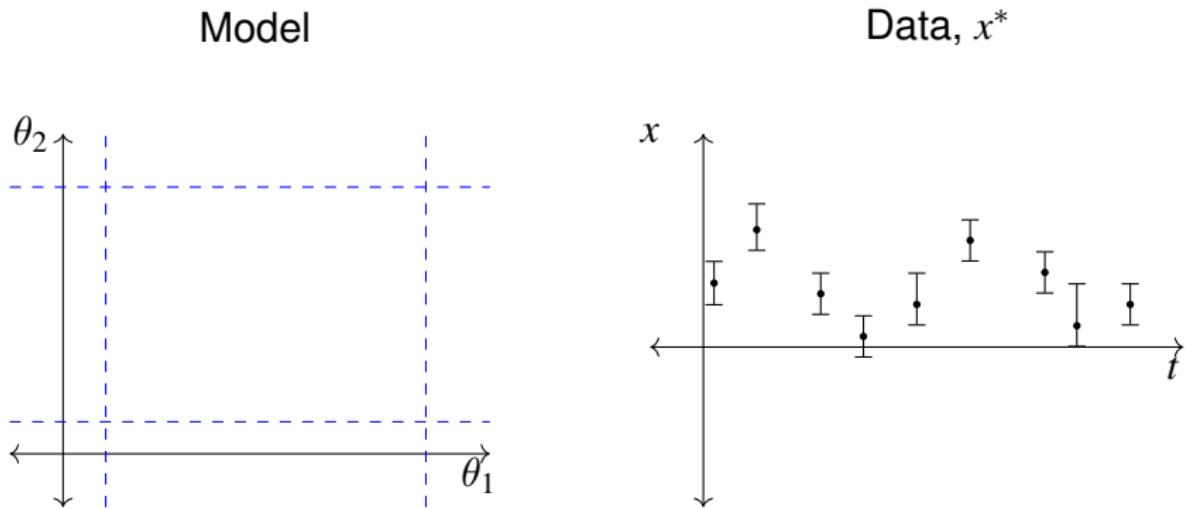
The output is not anymore distributed from the posterior distribution but from  $\pi(\theta|d(x^*, z) \leq \epsilon)$ .

[Pritchard et al, 1999]

# ABC ALGORITHM

```
1: Input: observation  $x^*$ , number of particles  $N$ , threshold  $\epsilon$ 
2: for  $i = 1, \dots, N$  do
3:   repeat
4:     generate  $\theta$  from the prior distribution  $\pi(\cdot)$ 
5:     generate  $z$  from the simulator  $f(\cdot|\theta)$ 
6:   until  $d(z, x^*) \leq \epsilon$ 
7:   set  $\theta^{(i)} = \theta$ 
8: end for
```

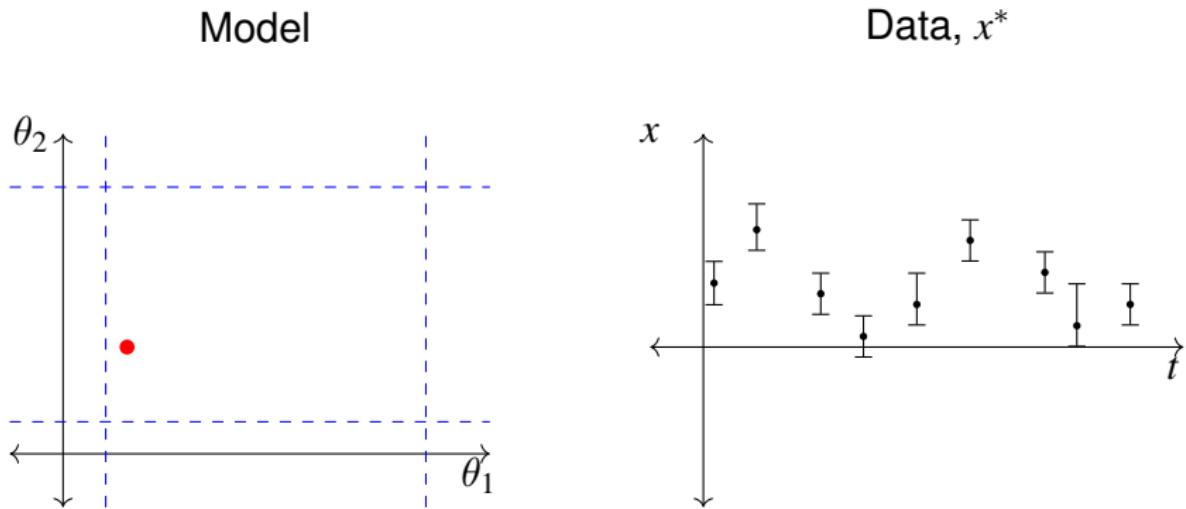
# ILLUSTRATION OF THE ABC REJECTION ALGORITHM



Acknowledgement to Prof. Michael Stumpf (Theoretical System Biology group, Imperial College London) for the slides.

Toni & Stumpf, Bioinformatics (2010)

# ILLUSTRATION OF THE ABC REJECTION ALGORITHM

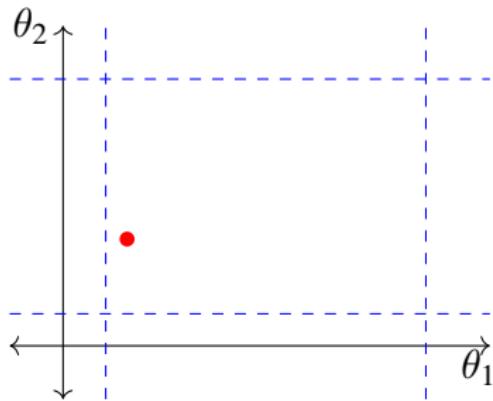


Acknowledgement to Prof. Michael Stumpf (Theoretical System Biology group, Imperial College London) for the slides.

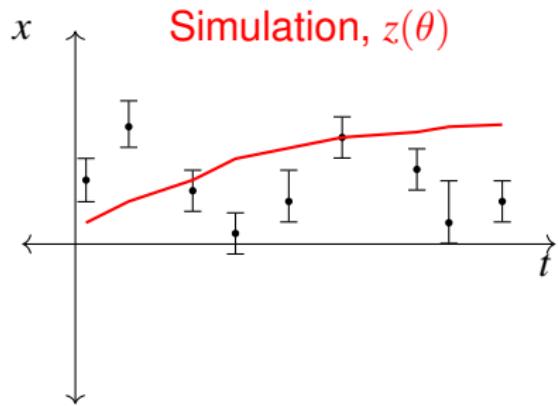
Toni & Stumpf, Bioinformatics (2010)

# ILLUSTRATION OF THE ABC REJECTION ALGORITHM

Model



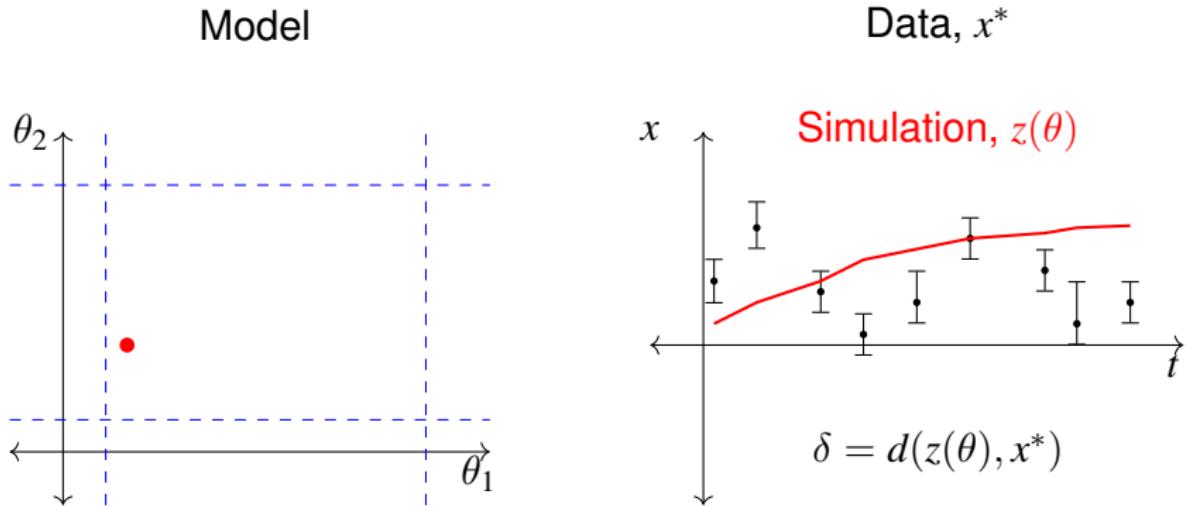
Data,  $x^*$



Acknowledgement to Prof. Michael Stumpf (Theoretical System Biology group, Imperial College London) for the slides.

Toni & Stumpf, Bioinformatics (2010)

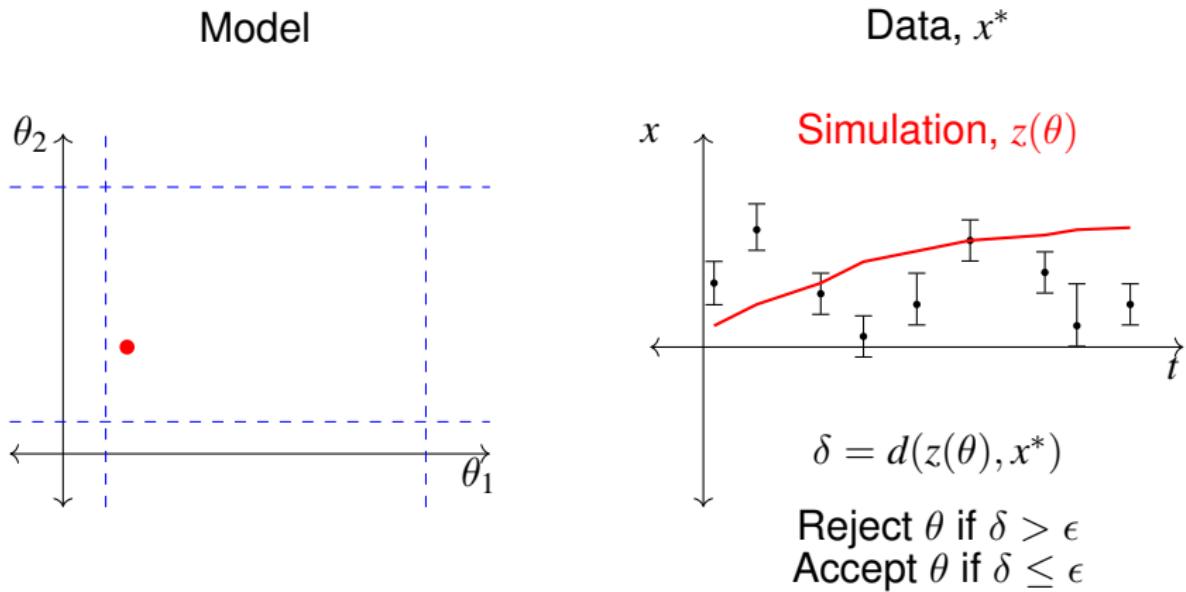
# ILLUSTRATION OF THE ABC REJECTION ALGORITHM



Acknowledgement to Prof. Michael Stumpf (Theoretical System Biology group, Imperial College London) for the slides.

Toni & Stumpf, Bioinformatics (2010)

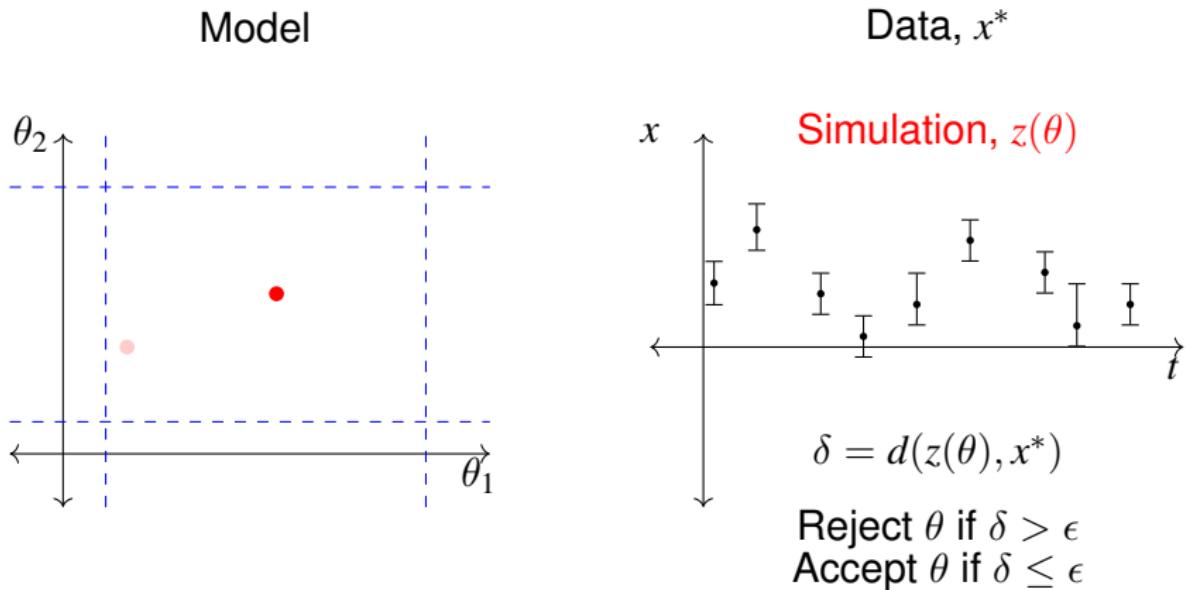
# ILLUSTRATION OF THE ABC REJECTION ALGORITHM



Acknowledgement to Prof. Michael Stumpf (Theoretical System Biology group, Imperial College London) for the slides.

Toni & Stumpf, Bioinformatics (2010)

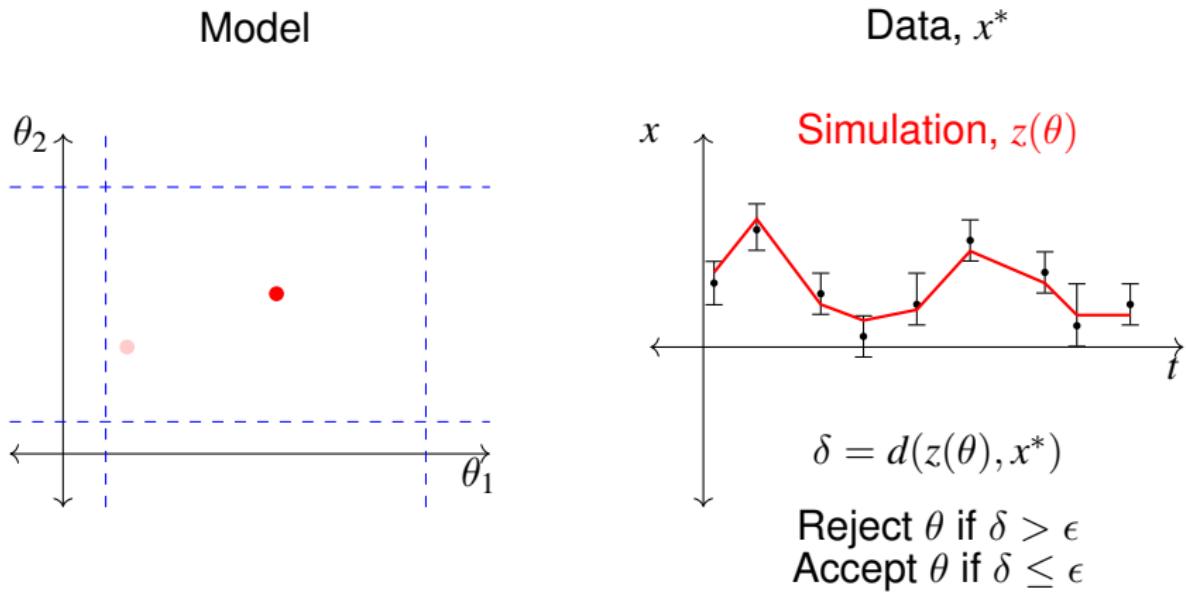
# ILLUSTRATION OF THE ABC REJECTION ALGORITHM



Acknowledgement to Prof. Michael Stumpf (Theoretical System Biology group, Imperial College London) for the slides.

Toni & Stumpf, Bioinformatics (2010)

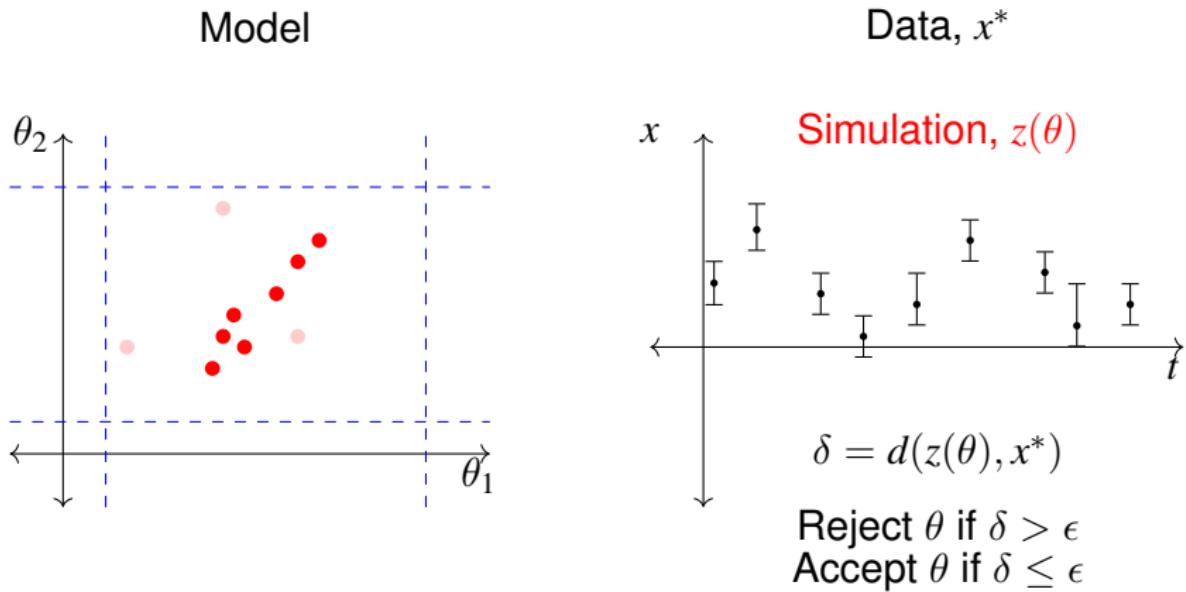
# ILLUSTRATION OF THE ABC REJECTION ALGORITHM



Acknowledgement to Prof. Michael Stumpf (Theoretical System Biology group, Imperial College London) for the slides.

Toni & Stumpf, Bioinformatics (2010)

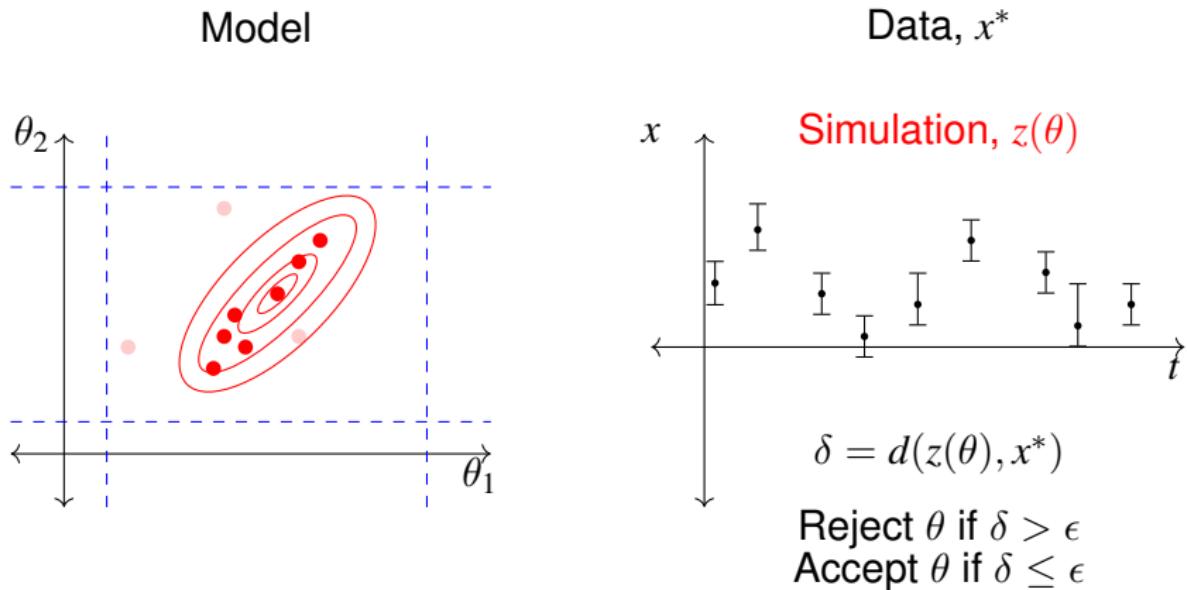
# ILLUSTRATION OF THE ABC REJECTION ALGORITHM



Acknowledgement to Prof. Michael Stumpf (Theoretical System Biology group, Imperial College London) for the slides.

Toni & Stumpf, Bioinformatics (2010)

# ILLUSTRATION OF THE ABC REJECTION ALGORITHM



Acknowledgement to Prof. Michael Stumpf (Theoretical System Biology group, Imperial College London) for the slides.

Toni & Stumpf, Bioinformatics (2010)

## OUTPUT OF THE ALGORITHM

The ABC rejection algorithm samples from the marginal in  $z$  of the joint distribution

$$\pi_\epsilon(\theta, z|x^*) = \frac{\pi(\theta)f(z|\theta)\mathbb{1}_{A_{\epsilon,x^*}}(z)}{\int \pi(\theta)f(z|\theta)\mathbb{1}_{A_{\epsilon,x^*}}(z)d\theta dz}$$

where  $A_{\epsilon,x^*} = \{z \mid d(z, x^*) \leq \epsilon\}$ .

The basic idea behind ABC is that using a **small enough tolerance threshold  $\epsilon$**  produce a **good enough approximation of the posterior distribution**, i.e.

$$\pi_\epsilon(\theta|x^*) = \int \pi_\epsilon(\theta, z|x^*)dz \approx \pi(\theta|x^*)$$

[Marin et al, 2012]

## CHOICE OF THE THRESHOLD

- $\epsilon$  should be close to 0 to obtain samples from a distribution close to the posterior distribution

$$\lim_{\epsilon \rightarrow 0} \pi_\epsilon(\theta|x^*) = \pi(\theta|x^*)$$

- choice of  $\epsilon$  is mostly a matter of computational costs:
  - small  $\epsilon$  are associated with high computational costs
  - Practice of ABC: select  $\epsilon$  as a small percentile of the simulated distance  $d(\eta(z), \eta(x^*))$  [Beaumont et al, 2009]

$$\epsilon_N = q_\alpha(d_1, \dots, d_N)$$

- other possible approach: including  $\epsilon$  in the inferential framework [Ratmann et al, 2009]

# ABC ALGORITHM WITH QUANTILE DISTANCE

```
1: Input: observation  $x^*$ , number of particles  $N$ , parameter  $\alpha$ 
2: for  $i = 1, \dots, N$  do
3:   generate  $\theta^{(i)}$  from the prior distribution  $\pi(\cdot)$ 
4:   generate  $z^{(i)}$  from the simulator  $f(\cdot | \theta)$ 
5: end for
6: compute  $\epsilon = \alpha\text{-th quantile of } \{d(\eta(z^{(i)}), \eta(x^*))\}_i$ 
7: Return:  $\{\theta^{(i)} \text{ s.t. } d(\eta(z^{(i)}), \eta(x^*)) \leq \epsilon\}$ 
```

The returned sample consists of  $\alpha \times N$  particles.

## EXAMPLE: THE MA PROCESS

The MA( $q$ ) process is a stochastic process  $(y_k)_{k \in \mathbb{N}^*}$  defined by

$$y_k = u_k + \sum_{i=1}^q \theta_i u_{k-i}$$

where  $(u_k)_{k \in \mathbb{Z}}$  is an i.i.d. sequence of standard Gaussians  $\mathcal{N}(0, 1)$ .

Simple prior: uniform over the inverse roots in

$$\mathcal{Q}(x) = 1 - \sum_{i=1}^q \theta_i x^i$$

under the identifiability condition.

For  $q = 2$ , this leads to a uniform prior over a triangle.

[Marin et al, 2012]

## EXAMPLE: THE MA PROCESS

Why using ABC ?

- The likelihood associated with  $(y_k)_{1 \leq k \leq n}$  is complex because of the need to integrate out  $u_{-q+1}, \dots, u_{-1}, u_0$ .
- MCMC algorithm feasible mainly for small values of  $n$

Running one iteration of ABC for this problem simply requires:

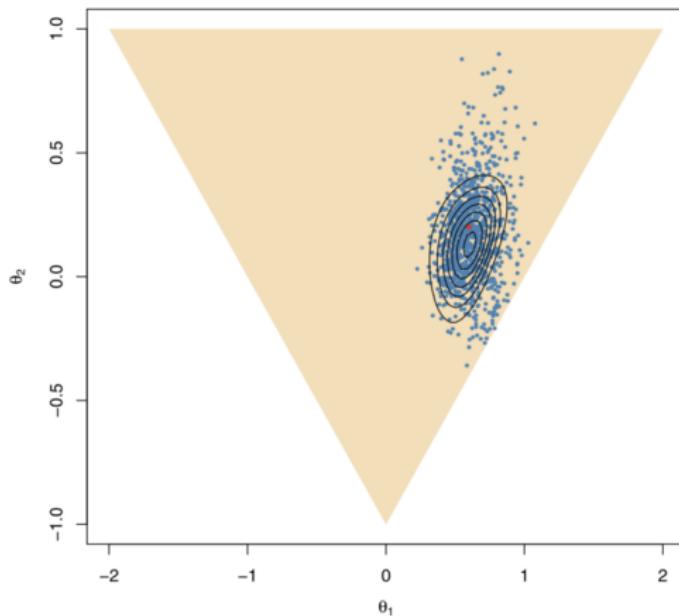
- ① Simulating the MA( $q$ ) coefficients  $\theta$  uniformly over the acceptable range,
- ② generating an i.i.d sequence  $(u_k)_{-q \leq k \leq n}$ ,
- ③ producing a simulated series  $(y_k)_{1 \leq k \leq n}$ .

Two distances:

- raw distance:  $d^2((z_k)_{1 \leq k \leq n}, (x_k^*)_{1 \leq k \leq n}) = \sum_{k=1}^n (y_k - x_k^*)^2$
- quadratic distance between  $q$  autocovariances  $\tau_j = \sum_{k=j+1}^n y_k y_{k-j}$

[Marin et al, 2012]

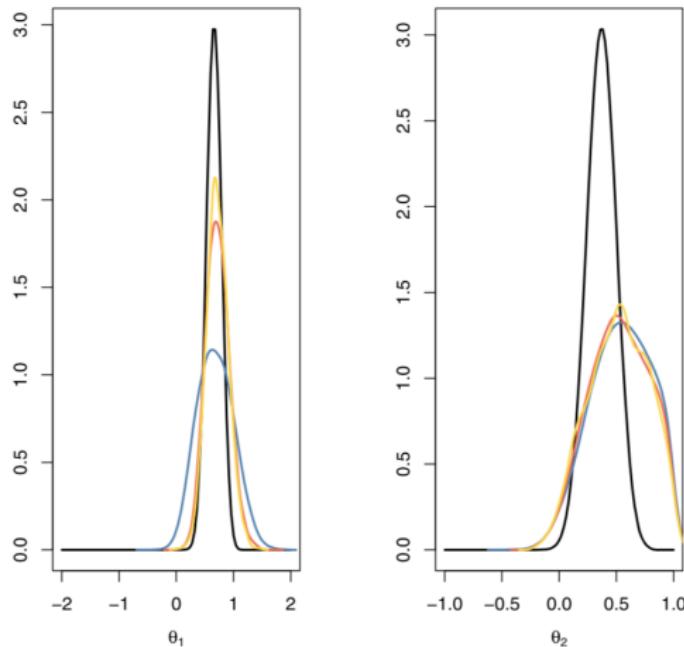
## EXAMPLE: THE MA PROCESS



Model: MA(2) with  $n = 100$  and  $\theta = (0.6, 0.2)$ ; autocovariance distance;  
 $\epsilon$  chosen so that 0.1% of  $N = 10^6$  simulated data are accepted

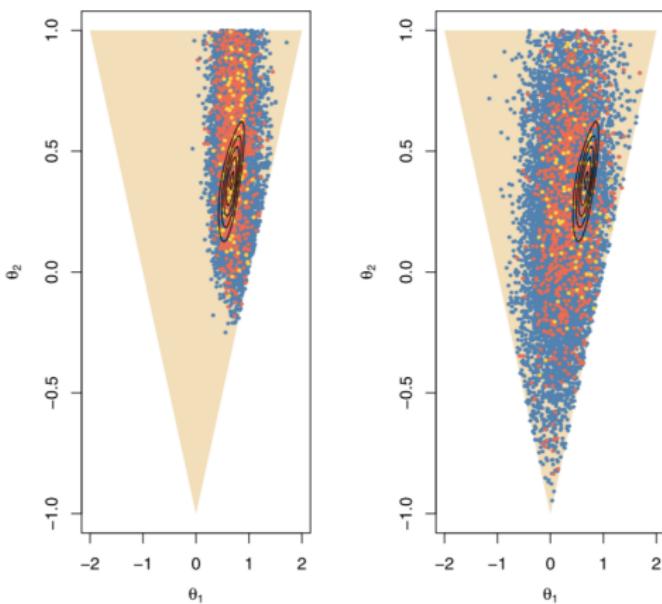
[Figure credit: Marin et al, 2012]

## EXAMPLE: THE MA PROCESS – THRESHOLD CHOICE



Evolution of the distribution of ABC samples using different quantiles for  $\epsilon = 10\%$  (blue),  $1\%$  (red) or  $0.1\%$  (yellow) compared to the true marginal density (black).

## EXAMPLE: THE MA PROCESS – DISTANCE CHOICE



Distance: autocovariance (left), raw distance (right);  $\epsilon$  chosen based on different quantile (1%, 0.1% and 0.01%).

[Figure credit: Marin et al, 2012]

# ABC FOR HIGH DIMENSIONAL DATA

## ABC algorithm

```
1: Input: observation  $x^*$ , number of particles  $N$ , threshold  $\epsilon$ 
2: for  $i = 1, \dots, N$  do
3:   repeat
4:     generate  $\theta$  from the prior distribution  $\pi(\cdot)$ 
5:     generate  $z$  from the simulator  $f(\cdot|\theta)$ 
6:   until  $d(z, x^*) \leq \epsilon$ 
7:   set  $\theta^{(i)} = \theta$ 
8: end for
```

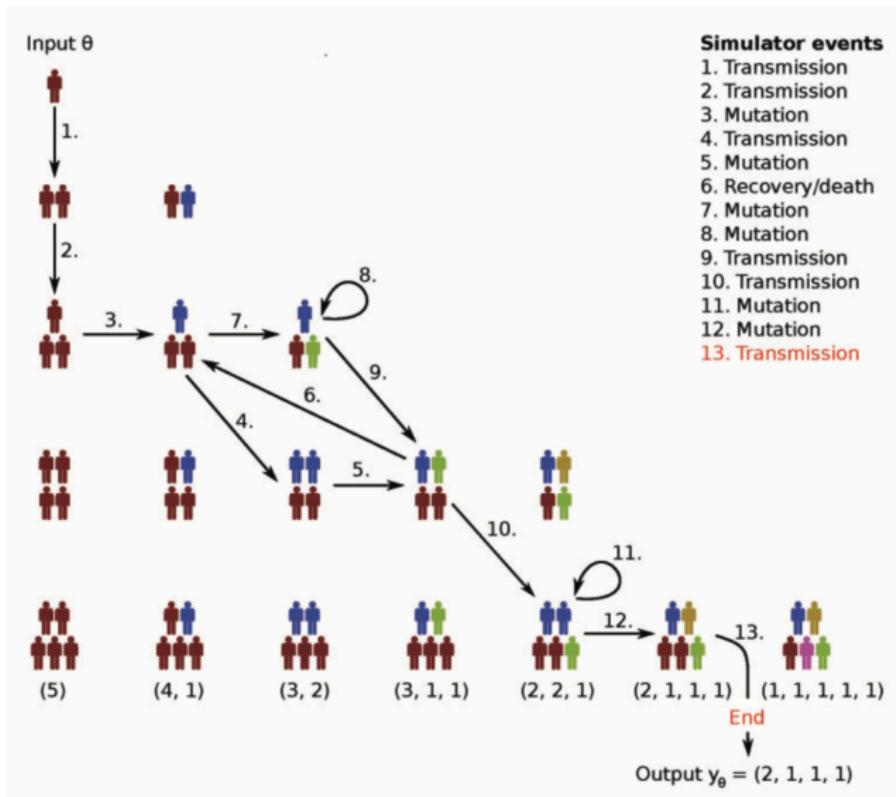
- **Curse of dimensionality:** If the data are high dimensional, we might never observe simulations close to the observed data.
- Solution: Reduce the dimension using summary statistics.

# ABC WITH SUMMARY STATISTICS

- **Curse of dimensionality:** If the data are high dimensional, we might never observe simulations close to the observed data.
- Solution: Reduce the dimension using summary statistics.
- Algorithm equivalent to the previous one i.f.f. sufficient statistic.

```
1: Input: observation  $x^*$ , number of particles  $N$ , threshold  $\epsilon$ 
2: for  $i = 1, \dots, N$  do
3:   repeat
4:     generate  $\theta$  from the prior distribution  $\pi(\cdot)$ 
5:     generate  $z$  from the simulator  $f(\cdot|\theta)$ 
6:   until  $d(\eta(z), \eta(x^*)) \leq \epsilon$ 
7:   set  $\theta^{(i)} = \theta$ 
8: end for
```

# EXAMPLE: THE SPREAD OF TUBERCULOSIS



[Tanaka et al, 2006; Figure credit: Lintusaari et al, 2017]

# EXAMPLE: THE SPREAD OF TUBERCULOSIS

Parameters of the model:

- transmission rate,  $\alpha$
- probability that a host stops being infectious,  $\delta$
- mutation rate,  $\tau$

**Observations:** Number of people infected with different pathogens in a subsample  $n$  of the population size  $m$ .

Example:  $x^* = (6, 4, 2, 2, 1, 1, 1)$

Why ABC?

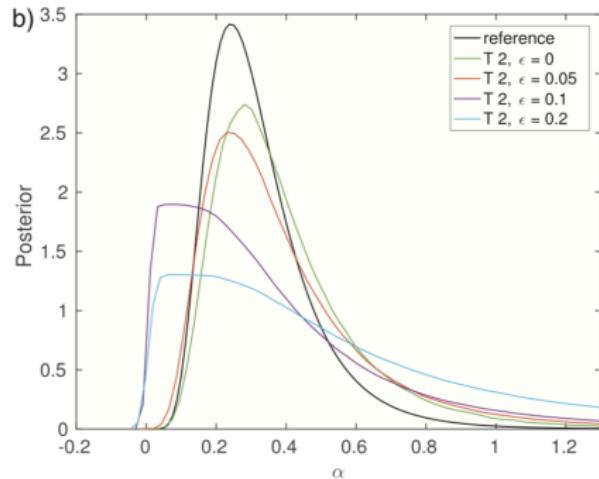
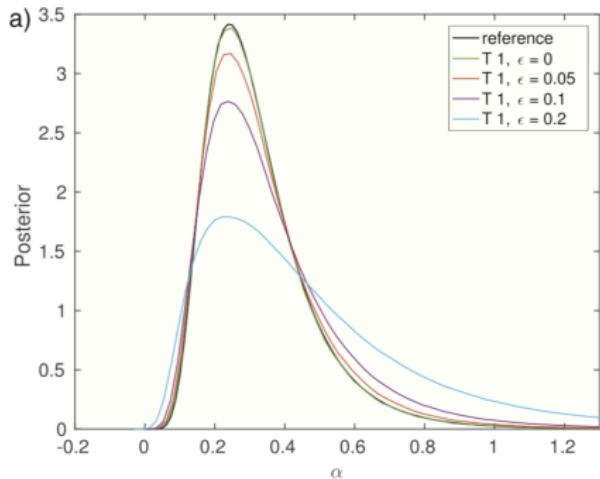
Intractable likelihood if population size  $m$  large or death rate,  $\delta > 0$ .

# EXAMPLE: THE SPREAD OF TUBERCULOSIS

Inference of the transmission rate,  $\alpha$ , for  $x^* = (6, 3, 2, 2, 1, 1, 1, 1, 1, 1)$  i.e.  $n = 20$

Two summary statistics:  $T_1(z) = \text{nb of clusters}/\text{sample size}$  and

$$T_2(z) = \text{genetic diversity} = 1 - \sum_i \frac{n_i}{n}$$

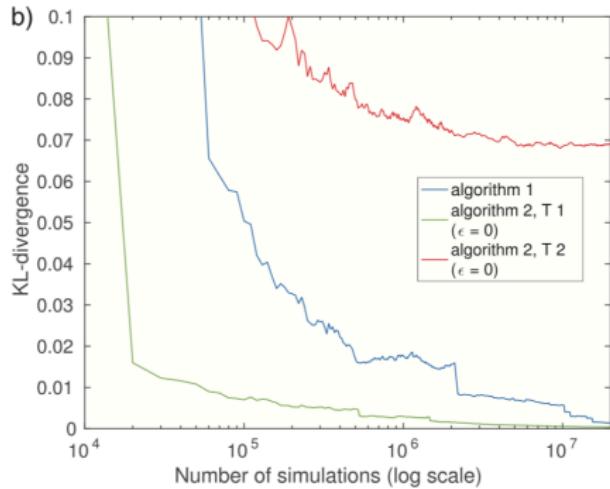
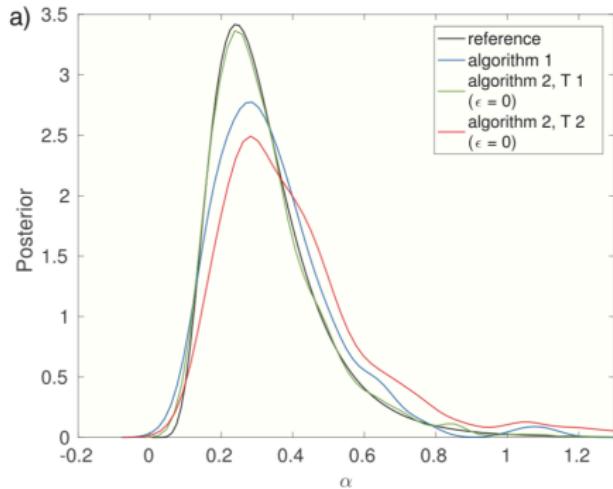


Posterior distribution obtained using the ABC rejection algorithm with 20 million simulated datasets.

[Figure credit: Lintusaari et al, 2017]

# EXAMPLE: THE SPREAD OF TUBERCULOSIS

Comparison of the efficiency of (1) exact rejection and (2) ABC rejection samplers



Left: Posterior distribution for a computational budget of 100.000 iterations

Right: Accuracy versus computational cost.

Smaller KL divergence means more accurate inference of the posterior distribution.

[Figure credit: Lintusaari et al, 2017]

# ERROR IN ABC

ABC contains two levels of approximation:

- ① the use of a threshold:  $\pi(\theta|\eta(x^*)) \approx \pi_\epsilon(\theta|\eta(x^*))$
- ② the use of a summary statistic:  $\pi(\theta|x^*) \approx \pi(\theta|\eta(x^*)).$   
If  $\eta : \mathcal{X} \rightarrow \mathbb{R}^k$  is a sufficient statistics:  $\pi(\theta|x^*) = \pi(\theta|\eta(x^*)).$

In general, trade-off:

- If  $k$  is small,  $\pi(\theta|\eta(x^*)) \approx \pi_\epsilon(\theta|\eta(x^*))$  but  $\pi(\theta|x^*) \neq \pi(\theta|\eta(x^*))$
- If  $k$  is large,  $\pi(\theta|x^*) \approx \pi(\theta|\eta(x^*))$  but  $\pi(\theta|\eta(x^*)) \neq \pi_\epsilon(\theta|\eta(x^*))$

Ideal situation:  $k$  small and  $\eta(\cdot)$  informative like a sufficient statistic.

## SUMMARY STATISTICS IN ABC

- Summary statistics are particularly useful if the observed data sets contain many exchangeable elements
- E.g. data is a  $n$ -sample from a model
- Distance metrics, such as Euclidian, on the raw data are inefficient as they do not take into account exchangeability [Sousa et al, 2009]

How can we choose good **low dimensional** summary statistics?

# SUMMARY STATISTICS IN ABC

Suppose we are given a candidate set  $\mathcal{S} = \{s_1, \dots, s_p\}$  of summary statistics from which to choose.

Two main approaches:

- Identify best subset selection  
[Joyce & Marjoram, 2008; Nunes & Balding, 2010]
- Find an optimal projection of  $\mathcal{S}$  onto a lower dimension  
[Wegmann et al, 2009; Fearnhead and Prangle, 2012]

**Warning:** Automated methods are a poor replacement for expert knowledge.

# POST-PROCESSING OF ABC OUTPUT

## Abc algorithm

```
1: Input: observation  $x^*$ , number of particles  $N$ , threshold  $\epsilon$ 
2: for  $i = 1, \dots, N$  do
3:   repeat
4:     generate  $\theta$  from the prior distribution  $\pi(\cdot)$ 
5:     generate  $z$  from the simulator  $f(\cdot | \theta)$ 
6:   until  $d(\eta(z), \eta(x^*)) \leq \epsilon$ 
7:   set  $\theta^{(i)} = \theta$ 
8: end for
```

Could we use all the simulated data instead of throwing away a lot of them?

## POST-PROCESSING OF ABC OUTPUT

ABC algorithm provides a sample  $(\theta^{(i)}, z^{(i)})_i \sim \pi_\epsilon(\theta, z | \eta(x^*))$ .

Two post-processing ideas:

- ① weighting the  $\theta^{(i)}$  according to  $d(\eta(z^{(i)}), x^*)$
- ② adjust the value of  $\theta^{(i)}$  using local regression to weaken the effect of the discrepancy between  $\eta(z^{(i)})$  and  $\eta(x^*)$ .

# REGRESSION ADJUSTMENTS

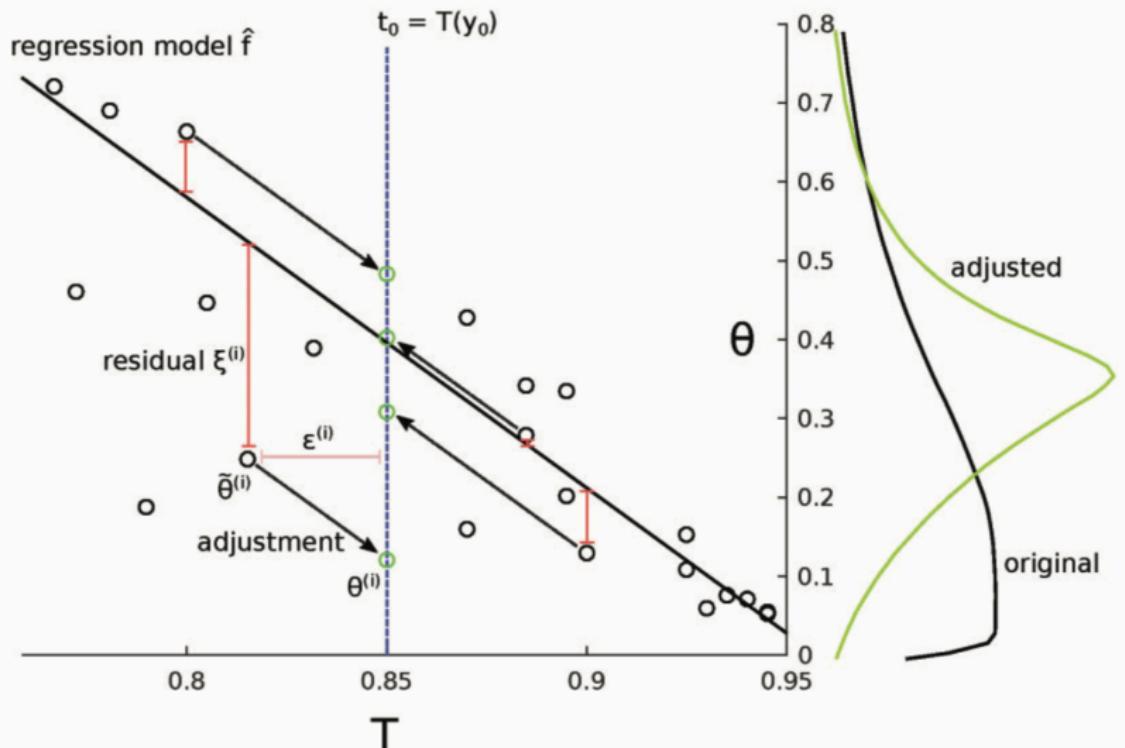
Regression adjustments have been shown to often give improved convergence of  $\pi_\epsilon(\theta|\eta(x^*))$  to the ideal target  $\pi(\theta|\eta(x^*))$ .

- ① Instead of throwing away  $\theta$  values such that  $d(\eta(z), \eta(x^*)) \geq \epsilon$ , adjust each sample  $\theta^{(i)}$  as

$$\tilde{\theta}^{(i)} = \theta^{(i)} - [\hat{E}(\theta|\eta(z^{(i)})) - \hat{E}(\theta|\eta(x^*))]$$

- ② Use regression to estimate  $\hat{E}(\theta|\eta(z))$

# ILLUSTRATION



Here,  $T$  denotes the summary statistic and  $y_0$  is the observed data

[Figure credit: Lintusaari et al, 2017]

# POST-PROCESSING OF ABC OUTPUT

**Local linear regression** [Beaumont et al, 2002]:

$$\tilde{\theta}^{(i)} = \theta^{(i)} - (\eta(z) - \eta(x^*))^T \hat{\beta}$$

where  $\hat{\beta}$  is obtained by a weighted least square regression on  $(\eta(z(\theta)) - \eta(x^*))$ , using weights of the form  $K_\delta(d(\eta(z), \eta(x^*)))$  for a non-parametric kernel  $K_\delta$  with bandwidth  $\delta$ .

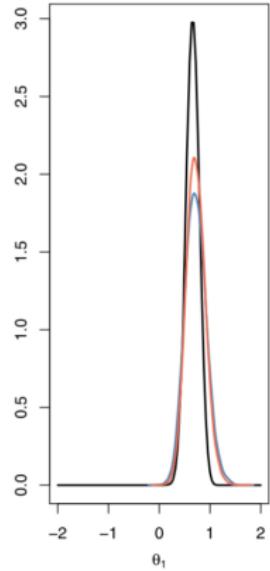
**Non-linear regression approaches** include

- one-layer neural network [Blum & Francois, 2010]
- kernel-ridge regression [Nakagome et al, 2013]

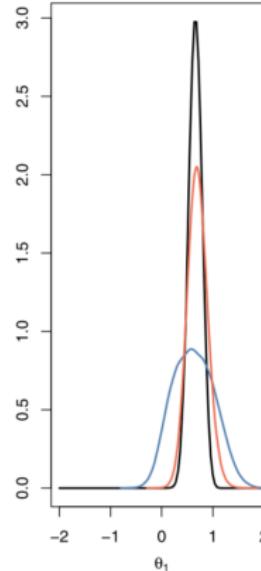
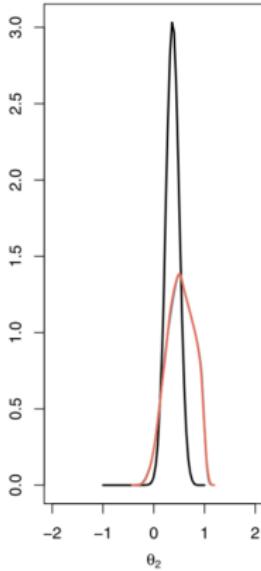
**Possible limitations:** Regression-adjustment can yield values of  $\tilde{\theta}$  that are outside of the range of the prior.

[Leuenberger & Wegmann, 2010; Fen et al, 2013]

# EXAMPLE: THE MA PROCESS



$\epsilon = 0.1\%$  quantile



$\epsilon = 20\%$  quantile

Black: Exact posterior; Blue: ABC posterior; Red: Beaumont correction

[Figure credit: Marin et al, 2012]

## Theoretical results on ABC

# POSTERIOR CONSISTENCY

As more data accumulates, we would like the posterior to stabilise around some true value.

**Bayesian consistency:**

- Assuming the dataset  $x^*(n)$  of size  $n$  is generated from the model with parameter value  $\theta_0$
- A posterior density  $\pi(\theta|x^*(n))$  is **Bayesian consistent** if for any  $\delta > 0$

$$\Pi(\|\theta - \theta_0\| > \delta \mid x^*(n)) \xrightarrow[n \rightarrow \infty]{P} 0$$

# POSTERIOR CONSISTENCY AND ABC

Bayesian consistency of posterior densities obtained from ABC

- requires  $n \rightarrow \infty$  and  $\epsilon \rightarrow 0$
- depends on the choice of  $\eta(\cdot) : \mathcal{X} \rightarrow \mathbb{R}^k$  with  $k \geq p$

## Bayesian consistency for ABC

Assuming

- the dataset  $x^*(n)$  of size  $n$  is generated from the model with parameter value  $\theta_0 \in \Theta \subset \mathbb{R}^p$
- concentration of  $\eta(\cdot)$

then the ABC posterior density  $\pi_\epsilon(\theta | \eta(x^*(n)))$  is Bayesian consistent, i.e. for any  $\delta > 0$

$$\Pi_\epsilon(\|\theta - \theta_0\| > \delta \mid x^*(n)) \xrightarrow[n \rightarrow \infty \text{ and } \epsilon \rightarrow 0]{P} 0$$

[Frazier et al, 2015]

# POSTERIOR CONSISTENCY AND ABC

## Bayesian consistency for ABC

Assuming

- the dataset  $x^*(n)$  of size  $n$  is generated from the model with parameter value  $\theta_0 \in \Theta \subset \mathbb{R}^p$
- concentration of  $\eta(\cdot)$

then the ABC posterior density  $\pi_\epsilon(\theta | \eta(x^*(n)))$  is Bayesian consistent, i.e. for any  $\delta > 0$

$$\Pi_\epsilon(\|\theta - \theta_0\| > \delta \mid x^*(n)) \xrightarrow[n \rightarrow \infty \text{ and } \epsilon \rightarrow 0]{P} 0$$

## Concentration of $\eta(\cdot)$ :

Consider a deterministic function  $\theta \mapsto b(\theta)$  such that

- $\eta(x^*(n)) \xrightarrow[n \rightarrow \infty]{} b(\theta_0)$
- and for all  $\theta$ ,  $\eta(z(\theta, n)) \xrightarrow[n \rightarrow \infty]{} b(\theta)$

[Frazier et al, 2015]

## ABC POSTERIOR SHAPE

- Posterior consistency guarantees that when the number of data increases and the threshold tends to 0, the ABC posterior stabilises around the true value.
- It does not indicate precisely how this mass accumulates.

This depends on the rate of convergence  $\sigma_n$  of the summary statistics  $\eta(z(\theta, n))$  towards  $b(\theta)$ :

- if  $\lim_{n \rightarrow \infty} \sigma_n / \epsilon_n = 0$ , the posterior distribution converges towards a uniform distribution.
- if  $\sigma_n \gg \epsilon_n$ , the posterior distribution converges towards a **Normal distribution**.

[Frazier et al, 2018]

### PRACTICAL CONSEQUENCE

No need to choose  $\epsilon_n$  arbitrarily small: just need  $\sigma_n \gg \epsilon_n$ .  
No gain in being more precise!

# MODEL MISSPECIFICATION

"All models are wrong but some are useful."

George Box

ABC under model misspecification:

- Different versions of ABC can yield substantially different results.
- Under regularity conditions, the ABC rejection approach concentrates posterior mass on an appropriately defined pseudo-true parameter value.
- Posterior obtained from regression-adjustments are potentially more misleading than the use of a standard rejection.

[van der Vart et al, 2015; Frazier et al, 2017; Beaumont, 2019]

# SUMMARY OF THE FIRST PART

## WHEN TO USE ABC?

To perform parameter inference for a complex model for which the likelihood can not be calculated.

## ABC ALGORITHM

- 1: **Input:** observation  $x^*$ , number of particles  $N$ , threshold  $\epsilon$
- 2: **for**  $i = 1, \dots, N$  **do**
- 3:     **repeat**
- 4:         generate  $\theta \sim \pi(\cdot)$
- 5:         generate  $z \sim f(\cdot|\theta)$
- 6:         **until**  $d(\eta(z), \eta(x^*)) \leq \epsilon$
- 7:     set  $\theta^{(i)} = \theta$
- 8: **end for**

## What we covered:

- Choice of statistic  $\eta(\cdot)$
- Post-processing
- Posterior consistency

# OUTLINE

## 1 THE ABC METHOD

- The basic principle of ABC
- ABC with summary statistics
- Post-processing of ABC output
- Theoretical results on ABC

## 2 EFFICIENT ABC ALGORITHMS

## 3 ABC FOR MODEL CHOICE

# APPROXIMATE BAYESIAN COMPUTATION

More efficient ABC algorithms include:

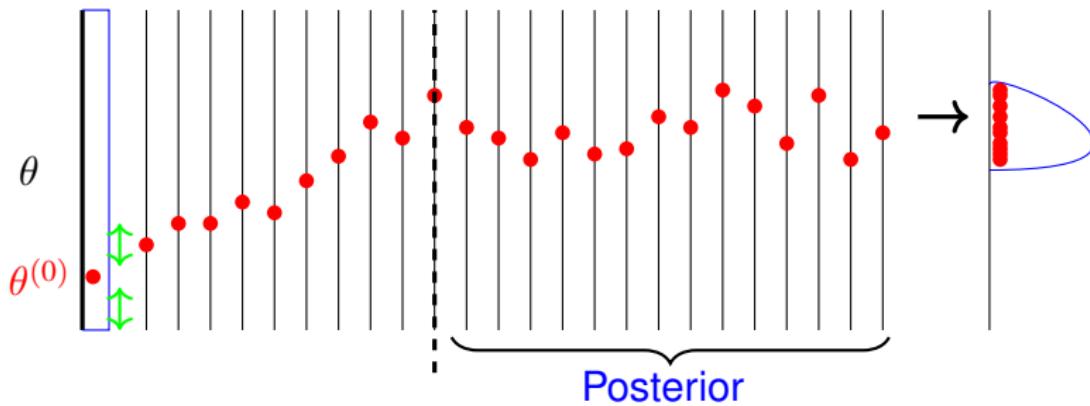
- Markov chain Monte Carlo ABC schemes  
[Marjoram and Molitor, 2003; Ratmann et al., 2007]
- ABC implementing some variant of sequential importance sampling (SIS) or sequential Monte Carlo (SMC)  
[ Sisson et al., 2007; Beaumont et al., 2009; Toni et al., 2009; Del Moral et al., 2011]

# THE METROPOLIS-HASTING ALGORITHM

- start from an arbitrary point  $\theta^{(0)}$
- for  $t = 1, \dots$ 
  - generate  $\tilde{\theta} \sim q(\tilde{\theta}|\theta^{(t)})$
  - take

$$\theta^{(t+1)} = \begin{cases} \tilde{\theta} & \text{with probability } p = \min\left(1, \frac{f(x^*|\tilde{\theta})\pi(\tilde{\theta})q(\theta^{(t)}|\tilde{\theta})}{f(x^*|\theta^{(t)})\pi(\theta^{(t)})q(\tilde{\theta}|\theta^{(t)})}\right) \\ \theta^{(t)} & \text{with probability } 1 - p \end{cases}$$

Target = stationnary distribution =  $\pi(\theta|x^*) \propto f(x^*|\theta)\pi(\theta)$



## ABC-MCMC METHOD

- MCMC approaches require to be able to evaluate the likelihood  
The acceptance ratio is

$$p = \min \left( 1, \frac{f(x^* | \tilde{\theta}) \pi(\tilde{\theta}) q(\theta^{(t)} | \tilde{\theta})}{f(x^* | \theta^{(t)}) \pi(\theta^{(t)}) q(\tilde{\theta} | \theta^{(t)})} \right)$$

- How can we construct a Markov-Chain Monte Carlo algorithm in the context of ABC?
- Idea: produce an MCMC algorithm which targets  $\pi_\epsilon(\theta, z | x^*)$  as its stationary distribution [Marjoram et al, 2003]

# ABC-MCMC METHOD

- 1: Use ABC rejection algorithm to get a realization  $(\theta^{(0)}, z^{(0)})$  from the target  $\pi_\epsilon(\theta, z|x^*)$
- 2: **for**  $t = 1, \dots, N$  **do**
- 3:   generate  $\theta$  from a Markov kernel  $q(\cdot|\theta^{(t-1)})$
- 4:   generate  $z$  from the simulator  $f(\cdot|\theta)$
- 5:   generate  $u$  from  $\mathcal{U}(0, 1)$
- 6:   **if**  $u \leq \frac{\pi(\theta)q(\theta^{(t-1)}|\theta)}{\pi(\theta^{(t-1)})q(\theta|\theta^{(t-1)})}$  and  $d(\eta(z), \eta(x^*)) \leq \epsilon$  **then**
- 7:     set  $(\theta^{(t)}, z^{(t)}) = (\theta, z)$
- 8:   **else** set  $(\theta^{(t)}, z^{(t)}) = (\theta^{(t-1)}, z^{(t-1)})$
- 9:   **end if**
- 10: **end for**

# ABC-MCMC – WHY DOES IT WORK?

- Target distribution:  $\pi_\epsilon(\theta, z|x^*) = \pi(\theta) f(z|\theta) \mathbb{1}_{A_{\epsilon,x^*}}(z)$
- Markov kernel from  $(\theta^{(t-1)}, z^{(t-1)})$  to  $(\theta, z)$ :  $q(\theta|\theta^{(t-1)})f(z|\theta)$

The acceptance ratio is therefore

$$\begin{aligned} & \frac{\pi_\epsilon(\theta, z|x^*)}{\pi_\epsilon(\theta^{(t-1)}, z^{(t-1)}|x^*)} \times \frac{q(\theta^{(t-1)}|\theta) f(z^{(t-1)}|\theta^{(t-1)})}{q(\theta|\theta^{(t-1)}) f(z|\theta)} \\ &= \frac{\pi(\theta) \textcolor{blue}{f(z|\theta)} \mathbb{1}_{A_{\epsilon,x^*}}(z)}{\pi(\theta^{(t-1)}) \textcolor{green}{f(z^{(t-1)}|\theta^{(t-1)})} \mathbb{1}_{A_{\epsilon,x^*}}(z^{(t-1)})} \times \frac{q(\theta^{(t-1)}|\theta) \textcolor{green}{f(z^{(t-1)}|\theta^{(t-1)})}}{q(\theta|\theta^{(t-1)}) \textcolor{blue}{f(z|\theta)}} \\ &= \frac{\pi(\theta) q(\theta^{(t-1)}|\theta)}{\pi(\theta^{(t-1)}) q(\theta|\theta^{(t-1)})} \mathbb{1}_{A_{\epsilon,x^*}}(z) \end{aligned}$$

and does not involve calculating the likelihood.

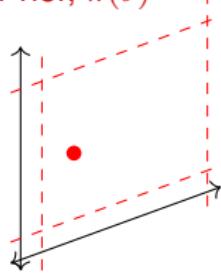
## SEQUENTIAL ABC

- Sequential approaches to ABC proposed by [Beaumont et al., 2009; Toni et al., 2009]
- Aim: computational efficiency
- Instead of directly producing a sample from  $\pi_\epsilon(\theta|\eta(x^*))$  for a small threshold  $\epsilon$
- Produce samples from  $\pi_{\epsilon_t}(\theta|\eta(x^*))$  for a decreasing sequence of thresholds  $(\epsilon_t)_t$

# SEQUENTIAL ABC – A SCHEMATIC REPRESENTATION

Population 1

Prior,  $\pi(\theta)$

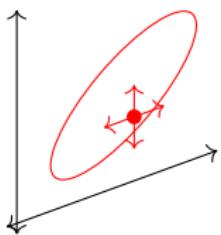


Define set of intermediate distributions,  $\pi_t, t = 1, \dots, T$

$$\epsilon_1 > \epsilon_2 > \dots > \epsilon_T$$

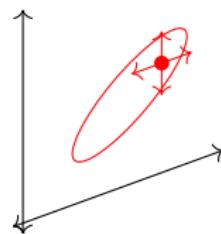
Population  $t - 1$

$$\pi_{t-1}(\theta | d(x^*, x(\theta)) < \epsilon_{t-1})$$



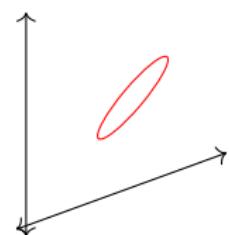
Population  $t$

$$\pi_t(\theta | d(x^*, x(\theta)) < \epsilon_t)$$



Population  $T$

$$\pi_T(\theta | d(x^*, x(\theta)) < \epsilon_T)$$



Acknowledgement to Prof. Michael Stumpf (Theoretical System Biology group, Imperial College London) for the slides.  
Toni et al., J.Roy.Soc. Interface (2009); Toni & Stumpf, Bioinformatics (2010).

## POPULATION MONTE-CARLO

The sequential ABC algorithms are related to the **population Monte-Carlo algorithm** [Cappé *et al*, 2004] and invoke **importance sampling** arguments:

- if a sample  $(\theta^{(1,t)}, \dots, \theta^{(N,t)})$  is produced by simulating each  $\theta^{(i,t)} \sim q_{it}$  independently of one another conditional on the past samples
- and if each point  $\theta^{(i,t)}$  is associated to an important weight

$$\omega^{(i,t)} \propto \frac{p(\theta^{(i,t)})}{q_{it}}$$

then  $\frac{1}{N} \sum_{i=1}^N \omega^{(i,t)} h(\theta^{(i,t)})$  is an unbiased estimator of  $\int h(\theta) p(\theta) d\theta$ .

The population Monte-Carlo algorithm produces a weighted sample  $(\theta^{(i,t)}, \omega^{(i,t)})_{i=1}^N$  from the target distribution  $p(\theta)$ .

# POPULATION MONTE-CARLO FOR ABC

- In ABC, the weight should not involve calculating the likelihood
- To do so, same trick as in ABC-MCMC: consider the joint target

$$\pi_\epsilon(\theta, z|x^*) = \pi(\theta) f(z|\theta) \mathbb{1}_{A_{\epsilon,x^*}}(z)$$

- At each iteration  $t$ , the sequential ABC algorithm produces a weighted sample  $(\theta^{(i,t)}, z^{(i,t)}, \omega^{(i,t)})_{i=1}^N$  from the target distribution.

## SEQUENTIAL ABC – THE APPROACH (1)

At iteration  $t$ , for every  $1 \leq i \leq N$ ,  $(\theta^{(i,t)}, z^{(i,t)})$  is sampled as follows:

- $\theta$  is sampled from the previous population  $\{\theta^{(i,t-1)}, \omega^{(i,t-1)}\}_{1 \leq i \leq N}$ ,
- then it is perturbed using a perturbation kernel  $\tilde{\theta} \sim K_t(\cdot | \theta)$
- a simulated data  $z$  is generated from the simulator  $f(\cdot | \tilde{\theta})$
- and  $(\theta^{(i,t)}, z^{(i,t)}) = (\tilde{\theta}, z)$  if  $d(\eta(z), \eta(x^*)) \leq \epsilon_t$

therefore the proposal distribution is

$$q_t(\theta^{(i,t)}, z^{(i,t)}) = \sum_{j=1}^n \omega^{(j,t-1)} K_t(\theta^{(i,t)} | \theta^{(j,t-1)}) f(z^{(i,t)} | \theta^{(i,t)}) \mathbb{1}_{A_{\epsilon_t, x^*}}(z^{(i,t)})$$

## SEQUENTIAL ABC – THE APPROACH (2)

At iteration  $t$ , for every  $1 \leq i \leq N$ ,  $(\theta^{(i,t)}, z^{(i,t)})$  is sampled from

$$q_t(\theta^{(i,t)}, z^{(i,t)}) = \sum_{j=1}^n \omega^{(j,t-1)} K_t(\theta^{(i,t)} | \theta^{(j,t-1)}) f(z^{(i,t)} | \theta^{(i,t)}) \mathbb{1}_{A_{\epsilon_t,x^*}}(z^{(i,t)}) .$$

The associated importance weight is

$$\begin{aligned}\omega(i, t) &\propto \frac{\text{Target}(\theta^{(i,t)}, z^{(i,t)})}{\text{Proposal}(\theta^{(i,t)}, z^{(i,t)})} \\&= \frac{\pi(\theta^{(i,t)}) f(z^{(i,t)} | \theta^{(i,t)}) \mathbb{1}_{A_{\epsilon_t,x^*}}(z^{(i,t)})}{\sum_{j=1}^n \omega^{(j,t-1)} K_t(\theta^{(i,t)} | \theta^{(j,t-1)}) f(z^{(i,t)} | \theta^{(i,t)}) \mathbb{1}_{A_{\epsilon_t,x^*}}(z^{(i,t)})} \\&= \frac{\pi(\theta^{(i,t)})}{\sum_{j=1}^n \omega^{(j,t-1)} K_t(\theta^{(i,t)} | \theta^{(j,t-1)})}\end{aligned}$$

and does not involve calculating the likelihood !

# SEQUENTIAL ABC – THE ALGORITHM

```
1: for all  $t$  do
2:    $i \leftarrow 1$ 
3:   repeat
4:     if  $t=1$  then
5:       sample  $\tilde{\theta}$  from  $\pi(\theta)$ 
6:     else
7:       sample  $\theta$  from the previous population  $\{\theta^{(i,t-1)}, \omega^{(i,t-1)}\}_{1 \leq i \leq N}$ 
8:       perturb  $\tilde{\theta}$  from  $K_t(\cdot|\theta)$  so that  $\pi(\tilde{\theta}) > 0$ 
9:     end if
10:    sample  $z$  from  $f(\cdot|\tilde{\theta})$ 
11:    if  $d(\eta(z), \eta(x^*)) \leq \epsilon_t$  then
12:       $\theta^{(i,t)} \leftarrow \tilde{\theta}; \quad i \leftarrow i + 1$ 
13:    end if
14:    until  $i = N + 1$ 
15:    calculate the weights:  $\omega^{(i,t)} \propto \frac{\pi(\theta^{(i,t)})}{\sum_{j=1}^n \omega^{(j,t-1)} K_t(\theta^{(i,t)} | \theta^{(j,t-1)})}; \quad \omega^{(i,1)} = 1/N$ 
16: end for
```

[ Toni *et al*, 2009; Beaumont *et al*,  
2009]

# SEQUENTIAL ABC – THE ALGORITHM

```
1: for all  $t$  do
2:    $i \leftarrow 1$ 
3:   repeat
4:     if  $t=1$  then
5:       sample  $\tilde{\theta}$  from  $\pi(\theta)$ 
6:     else
7:       sample  $\theta$  from the previous population  $\{\theta^{(i,t-1)}, \omega^{(i,t-1)}\}_{1 \leq i \leq N}$ 
8:       perturb  $\tilde{\theta}$  from  $K_t(\cdot|\theta)$  so that  $\pi(\tilde{\theta}) > 0$ 
9:     end if
10:    sample  $z$  from  $f(\cdot|\tilde{\theta})$ 
11:    if  $d(\eta(z), \eta(x^*)) \leq \epsilon_t$  then
12:       $\theta^{(i,t)} \leftarrow \tilde{\theta}; \quad i \leftarrow i + 1$ 
13:    end if
14:  until  $i = N + 1$ 
15:  calculate the weights:  $\omega^{(i,t)} \propto \frac{\pi(\theta^{(i,t)})}{\sum_{j=1}^n \omega^{(j,t-1)} K_t(\theta^{(i,t)} | \theta^{(j,t-1)})}; \quad \omega^{(i,1)} = 1/N$ 
16: end for
```

Impact on computational efficiency:

- perturbation kernels  $\{K_t\}_t$
- threshold schedule  $\{\epsilon_t\}_t$

Toni *et al.*, 2009; Beaumont *et al.*, 2009

# CHOICE OF THE PERTURBATION KERNEL

Use adaptive perturbation kernel

- Local perturbation kernel:
  - hardly moves particles
  - high acceptance rate if successive values of  $\epsilon$  close enough
- Widely spread kernel:
  - exploration of the parameter space
  - low acceptance rate

**Balance between exploring the parameter space and ensuring a high acceptance rate**

[Filippi et al, 2013]

# PROPERTIES OF OPTIMAL KERNEL

## ① From sequential importance sampling theory:

similarity between two joint distributions of  $(\theta^{(t-1)}, \theta^{(t)})$  where  $\theta^{(t-1)} \sim \pi_{\epsilon_{t-1}}$  and

- $\theta^{(t)}$  either constructed by perturbing  $\theta^{(t-1)}$  using kernel  $K$  and accepting according to threshold  $\epsilon_t$
- or  $\theta^{(t)} \sim \pi_{\epsilon_t}$

## ② Computational efficiency: high acceptance rate

## ③ Theoretical requirements for convergence:

- kernel with larger support than the target distribution  
⇒ guarantee asymptotic unbiasedness of the empirical mean
- vanish slowly enough in the tails of the target  
⇒ guarantee finite variance of the estimator

[Filippi et al, 2013]

# DERIVATION OF OPTIMAL KERNEL

**Criteria 1:** Resemblance between the two distributions

$$q_{\epsilon_{t-1}, \epsilon_t}(\theta^{(t-1)}, \theta^{(t)}) = \frac{\pi_{\epsilon_{t-1}}(\theta^{(t-1)} | x^*) K_t(\theta^{(t)} | \theta^{(t-1)}) \int f(z | \theta^{(t)}) \mathbb{1}_{A_{\epsilon_t, x^*}}(z) dz}{\alpha(K_t, \epsilon_{t-1}, \epsilon_t, x^*)}$$

and

$$q_{\epsilon_{t-1}, \epsilon_t}^*(\theta^{(t-1)}, \theta^{(t)}) = \pi_{\epsilon_{t-1}}(\theta^{(t-1)} | x^*) \pi_{\epsilon_t}(\theta^{(t)} | x^*)$$

in terms of the KL divergence [Douc *et al*, 2007; Cappé *et al*, 2008; Beaumont *et al*, 2009)]

$$KL(q_{\epsilon_{t-1}, \epsilon_t}; q_{\epsilon_{t-1}, \epsilon_t}^*) = -Q(K_t, \epsilon_{t-1}, \epsilon_t, x^*) + \log \alpha(K_t, \epsilon_{t-1}, \epsilon_t, x^*) + C(\epsilon_{t-1}, \epsilon_t, x^*)$$

where

$$Q(K_t, \epsilon_{t-1}, \epsilon_t, x^*) = \iint \pi_{\epsilon_{t-1}}(\theta^{(t-1)} | x^*) \pi_{\epsilon_t}(\theta^{(t)} | x^*) \log K_t(\theta^{(t)} | \theta^{(t-1)}) d\theta^{(t-1)} d\theta^{(t)}$$

[Filippi *et al*, 2013]

# DERIVATION OF OPTIMAL KERNEL

**Criteria 1:** minimise the KL divergence:

$$KL(q_{\epsilon_{t-1}, \epsilon_t}; q_{\epsilon_{t-1}, \epsilon_t}^*) = -Q(K_t, \epsilon_{t-1}, \epsilon_t, x) + \log \alpha(K_t, \epsilon_{t-1}, \epsilon_t, x^*) + C(\epsilon_{t-1}, \epsilon_t, x^*)$$

[Filippi et al, 2013]

# DERIVATION OF OPTIMAL KERNEL

**Criteria 1:** minimise the KL divergence:

$$KL(q_{\epsilon_{t-1}, \epsilon_t}; q_{\epsilon_{t-1}, \epsilon_t}^*) = -Q(K_t, \epsilon_{t-1}, \epsilon_t, x) + \log \alpha(K_t, \epsilon_{t-1}, \epsilon_t, x^*) + C(\epsilon_{t-1}, \epsilon_t, x^*)$$

**Criteria 2:** maximise the acceptance rate  $\alpha(K_t, \epsilon_{t-1}, \epsilon_t, x^*)$

## METHOD

Select the kernel  $K_t$  which maximises  $Q(K_t, \epsilon_{t-1}, \epsilon_t, x^*)$  which is equivalent to maximise  $-KL(q_{\epsilon_{t-1}, \epsilon_t}; q_{\epsilon_{t-1}, \epsilon_t}^*) + \log \alpha(K_t, \epsilon_{t-1}, \epsilon_t, x^*)$ .

## Remark:

$$Q(K_t, \epsilon_{t-1}, \epsilon_t, x^*) = \iint \pi_{\epsilon_{t-1}}(\theta^{(t-1)} | x^*) \pi_{\epsilon_t}(\theta^{(t)} | x^*) \log K_t(\theta^{(t)} | \theta^{(t-1)}) d\theta^{(t-1)} d\theta^{(t)}$$

can be maximized easily for some families of kernels. [Filippi et al, 2013]

# GAUSSIAN RANDOM WALK KERNELS

- **Component-wise kernel:**

diagonal covariance matrix  $\Sigma^{(t)}$  Beaumont et al, 2009

- **Multi-variate normal kernel:**

$$\Sigma^{(t)} \approx \sum_{i=1}^N \sum_{k=1}^{N_0} \omega^{(i,t-1)} \tilde{\omega}^{(k)} (\tilde{\theta}^{(k)} - \theta^{(i,t-1)}) (\tilde{\theta}^{(k)} - \theta^{(i,t-1)})^T$$

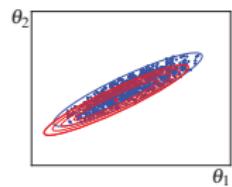
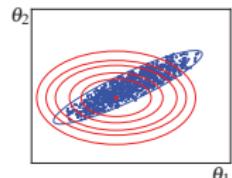
where  $\left\{ \tilde{\theta}^{(k)} \right\}_{1 \leq k \leq N_0} = \left\{ \theta^{(i,t-1)} \text{ s.t. } z^{(i,t-1)} \in \mathbb{1}_{A_{\epsilon_t, x^*}} \right\}$

- **Local multi-variate normal kernels:**

use a different kernel for each particle  $\theta^{(t-1)}$ .

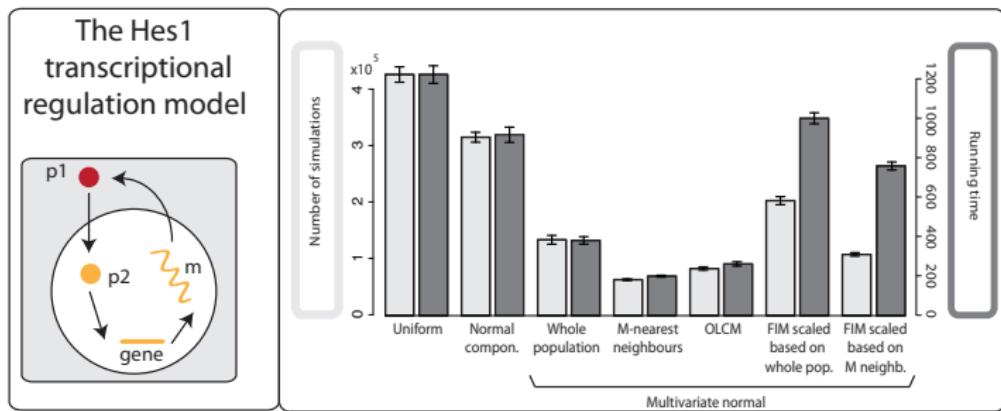
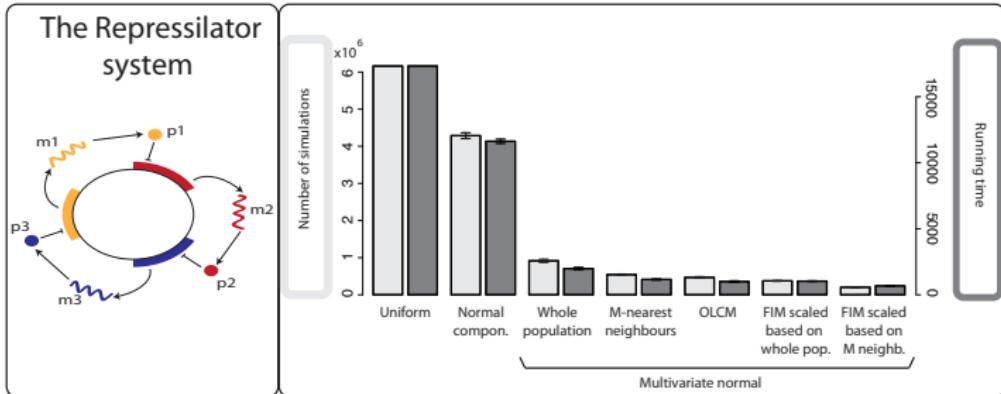
- Nearest neighbours:  $\Sigma_{\theta^{(t-1)}, M}^{(t)}$  based on the  $M$  nearest neighbours
- Optimal local covariance matrix:

$$\Sigma_{\theta^{(t-1)}}^{(t)} \approx \sum_{k=1}^{N_0} \tilde{\omega}^{(k)} (\tilde{\theta}^{(k)} - \theta^{(t-1)}) (\tilde{\theta}^{(k)} - \theta^{(t-1)})^T$$



[Filippi et al, 2013]

# COMPUTATIONAL EFFICIENCY AND KERNEL



# COMPUTATIONAL COST OF PERTURBATION KERNEL

- simulating the data dominates
- computational cost of perturbation kernel implementation:

Component-wise normal	$O(dN^2)$
Multivariate normal based on the whole previous population	$O(d^2N^2)$
Multivariate normal based on the $M$ nearest neighbours	$O((d + M)N^2 + d^2M^2N)$
Multivariate normal with OLCM	$O(d^2N^2)$
Multivariate normal based on the FIM (normalized with entire population)	$O(dCN + d^2N^2)$

$N$  = population size;  $d$  = parameter dimension

## RECOMMENDATION

Use of multivariate kernels with OLCM

- highest acceptance rate in our examples
- relatively easy to implement at acceptable computational cost

[Filippi et al, 2013]

# CHOICE OF THE THRESHOLD SCHEDULE

```
1: for all  $t$  do
2:    $i \leftarrow 1$ 
3:   repeat
4:     if  $t=1$  then
5:       sample  $\tilde{\theta}$  from  $\pi(\theta)$ 
6:     else
7:       sample  $\theta$  from the previous population  $\{\theta^{(i,t-1)}, \omega^{(i,t-1)}\}_{1 \leq i \leq N}$ 
8:       perturb  $\tilde{\theta}$  from  $K_t(\cdot | \theta)$  so that  $\pi(\tilde{\theta}) > 0$ 
9:     end if
10:    sample  $z$  from  $f(\cdot | \tilde{\theta})$ 
11:    if  $d(\eta(z), \eta(x^*)) \leq \epsilon_t$  then
12:       $\theta^{(i,t)} \leftarrow \tilde{\theta}; \quad i \leftarrow i + 1$ 
13:    end if
14:    until  $i = N + 1$ 
15:    calculate the weights:  $\omega^{(i,t)} \propto \frac{\pi(\theta^{(i,t)})}{\sum_{j=1}^n \omega^{(j,t-1)} K_t(\theta^{(i,t)} | \theta^{(j,t-1)})}; \quad \omega^{(i,1)} = 1/N$ 
16: end for
```

- final value of  $\epsilon$  close to 0
- minimise number of simulations:
  - minimise number of populations
  - maximise acceptance rate per pop.

# CHOICE OF THE THRESHOLD SCHEDULE

## PREVIOUS APPROACHES

- Trial and error
- Adaptive method based on quantile

Beaumont *et al*, 2009; Del Moral *et al*, 2008; Drovandi *et al*, 2011

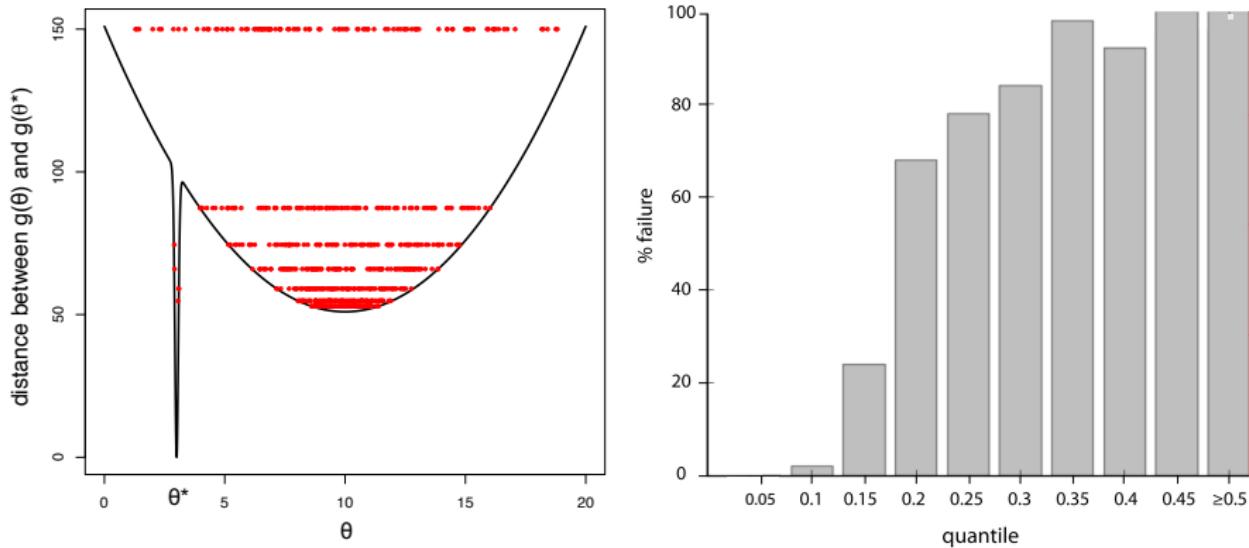
- ...

## The quantile based approach:

$$\epsilon_t = \alpha - \text{quantile of the distances } \{d(\eta(x^{(i,t-1)}), \eta(x^*))\}_{1 \leq i \leq N}$$

Which value of  $\alpha$  should we chose?

# DRAWBACK OF THE QUANTILE APPROACH



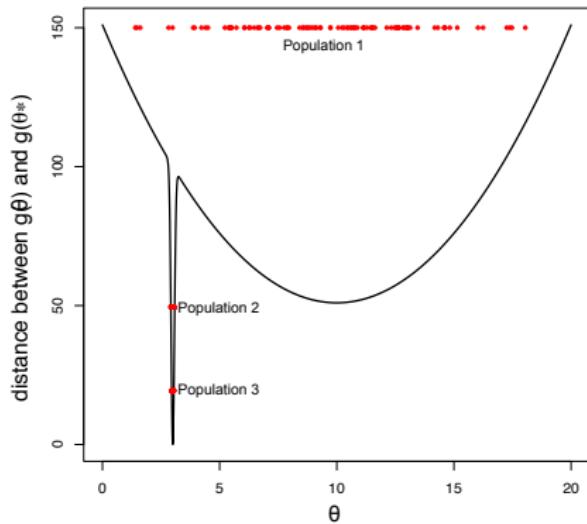
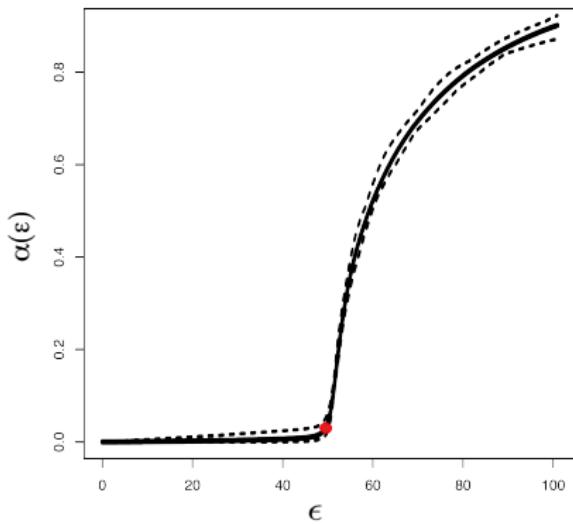
- For large  $\alpha$ , the algorithm ‘fails’.
- The optimal (or safe) choice of  $\alpha$  depends on the data, the model and the prior range.

[Silk et al, 2013]

# THE THRESHOLD-ACCEPTANCE RATE CURVE

Idea:

- estimate the threshold-acceptance rate curve
- avoid area with excessively high acceptance rate

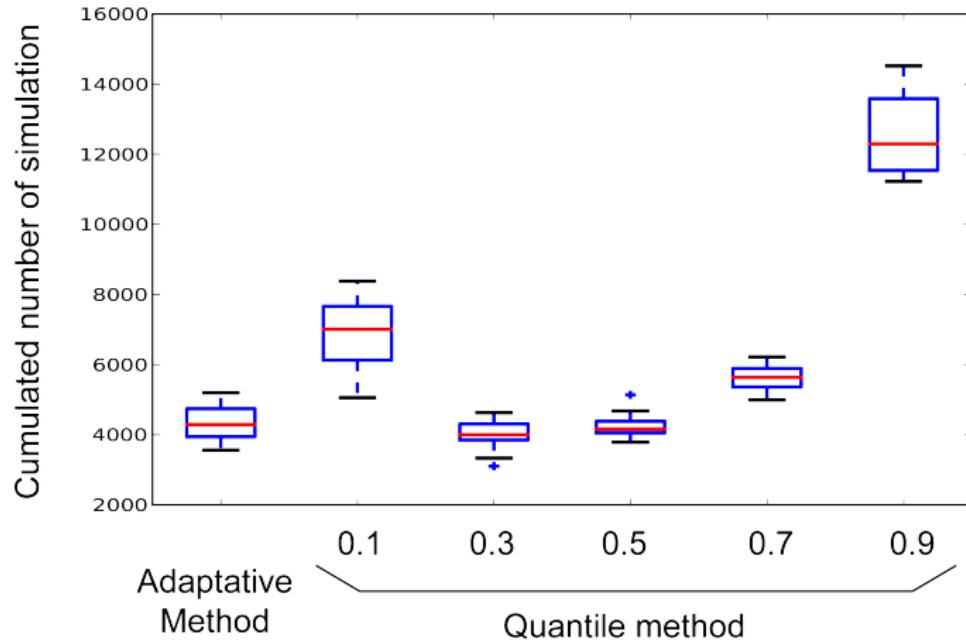


Possible solution: use Unscented Transform to estimate the threshold-acceptance rate curve.

[Silk et al, 2013]

# COMPUTATIONAL EFFICIENCY

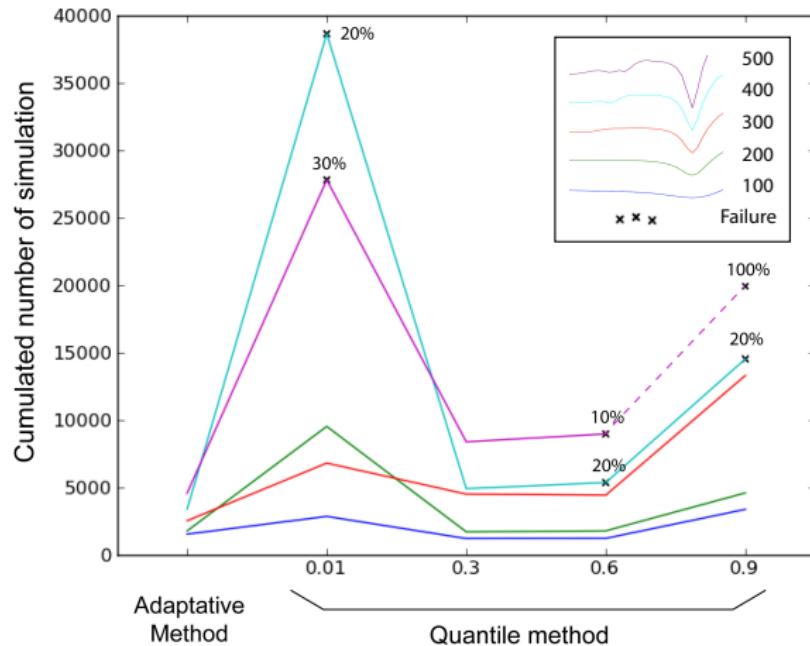
The Repressilator system:



[Silk et al, 2013]

# COMPUTATIONAL EFFICIENCY

Chemical reaction system illustrating the "local minimum problem"



[Silk et al, 2013]

# OUTLINE

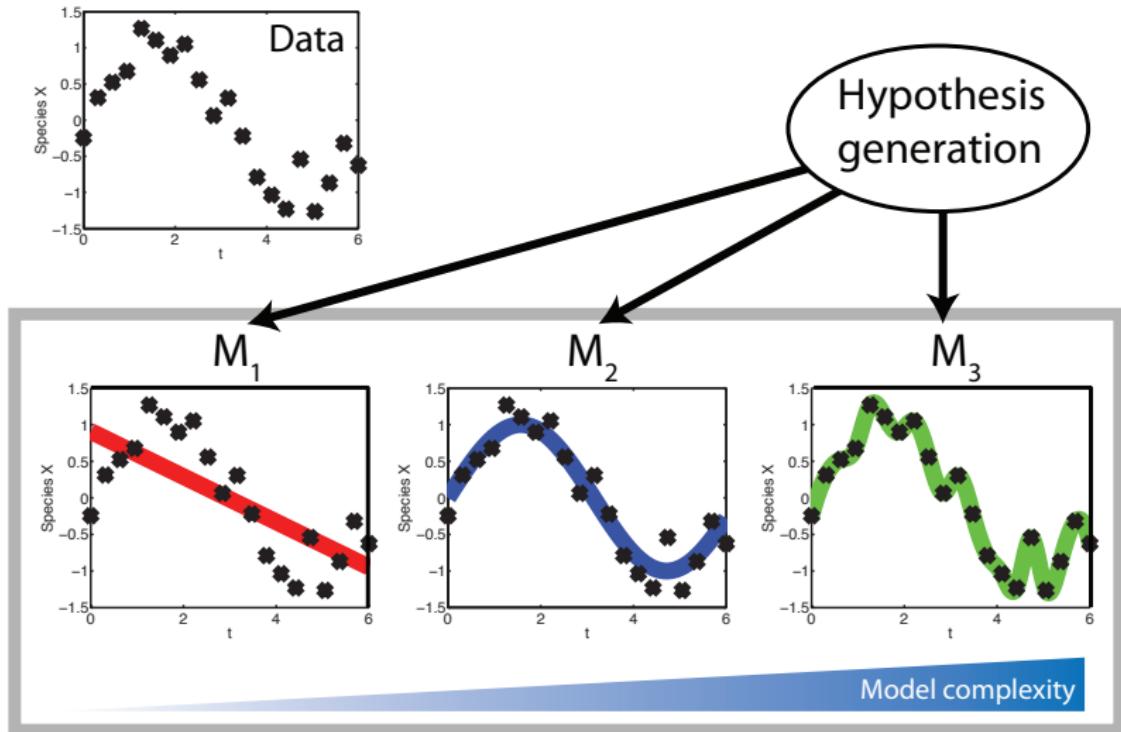
## 1 THE ABC METHOD

- The basic principle of ABC
- ABC with summary statistics
- Post-processing of ABC output
- Theoretical results on ABC

## 2 EFFICIENT ABC ALGORITHMS

## 3 ABC FOR MODEL CHOICE

# MODEL SELECTION



[ Kirk et al, 2013]

# BAYESIAN APPROACH FOR MODEL SELECTION

- How likely is a model  $\mathcal{M}$  given the observed data  $x^*$ ?

$$\pi(\mathcal{M}|x^*) \propto f(x^*|\mathcal{M})\pi(\mathcal{M})$$

- Model evidence

$$\pi(x^*|\mathcal{M}) = \int f(x^*|\theta, \mathcal{M})\pi(\theta|\mathcal{M})d\theta$$

- Bayes factor to discriminate between two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$ :

$$B_{1,2} = \frac{\pi(\mathcal{M}_1|x^*)}{\pi(\mathcal{M}_2|x^*)} = \frac{\pi(x^*|\mathcal{M}_1)\pi(\mathcal{M}_1)}{\pi(x^*|\mathcal{M}_2)\pi(\mathcal{M}_2)}$$

Use Jeffrey's scale to interpret values of Bayes factor

# COMPUTATION OF THE EVIDENCE

Model evidence:

$$\pi(x^*|\mathcal{M}) = \int f(x^*|\theta, \mathcal{M})\pi(\theta|\mathcal{M})d\theta$$

- Analytically (?)
- Using importance sampling methods
- Some parameter inference methods also allows us to compute model evidence
- Perform inference on the joint space (parameter and model) and then marginalisation

# GENERIC ABC FOR MODEL CHOICE

```
1: Input: observation  $x^*$ , threshold  $\epsilon$ , finite set of models
2: for  $i = 1, \dots, N$  do
3:   repeat
4:     generate  $m$  from the prior  $\pi(\mathcal{M} = m)$ 
5:     generate  $\theta_m$  from the prior distribution  $\pi_m(\cdot)$ 
6:     generate  $z$  from the simulator  $f_m(\cdot | \theta_m)$ 
7:   until  $d(\eta(z), \eta(x^*)) \leq \epsilon$ 
8:   set  $m^{(i)} = m$  and  $\theta^{(i)} = \theta_m$ 
9: end for
```

Notation: Here we index by  $m$  the priors and likelihood functions to highlight the fact that they are different for each model.

# ABC FOR MODEL SELECTION

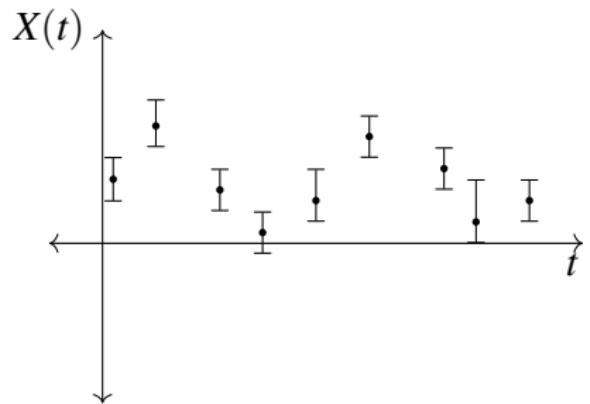
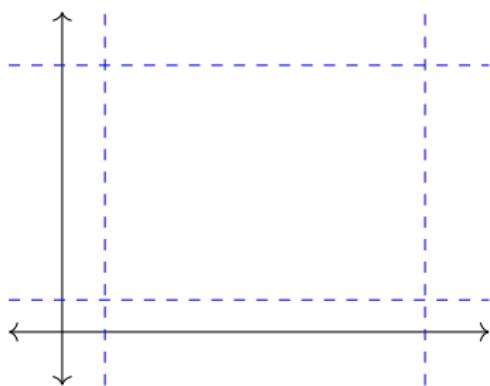
	$M_1$	$M_2$	$M_3$	$M_4$
Model accepted	-	-	-	-

# ABC FOR MODEL SELECTION

	$M_1$	$M_2$	$M_3$	$M_4$
Model accepted	-	-	-	-

# ABC FOR MODEL SELECTION

	$M_1$	$M_2$	$M_3$	$M_4$
Model accepted	-	-	-	-
Model $M_2$				

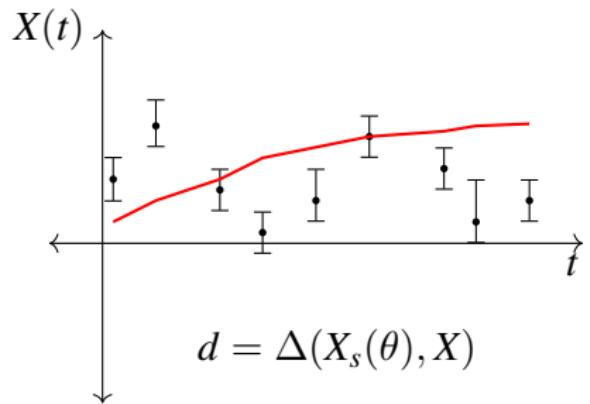
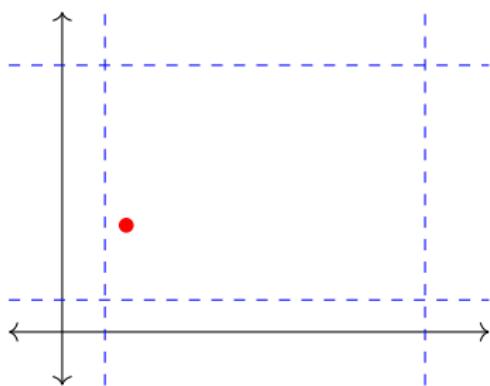


# ABC FOR MODEL SELECTION

$M_1 \quad M_2 \quad M_3 \quad M_4$

Model accepted

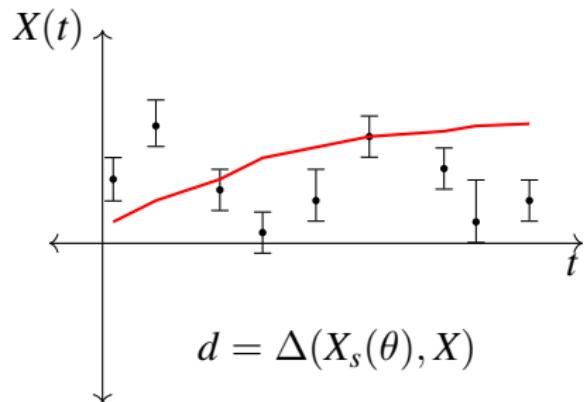
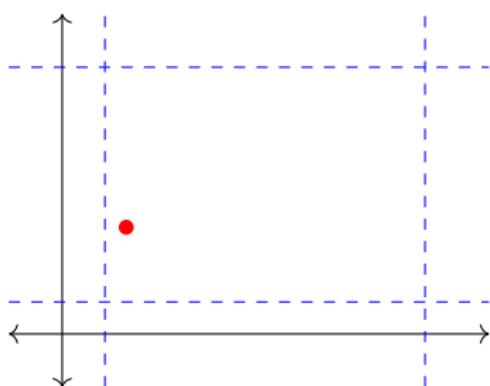
Model  $M_2$



# ABC FOR MODEL SELECTION

	$M_1$	$M_2$	$M_3$	$M_4$
Model accepted	-	0	-	-

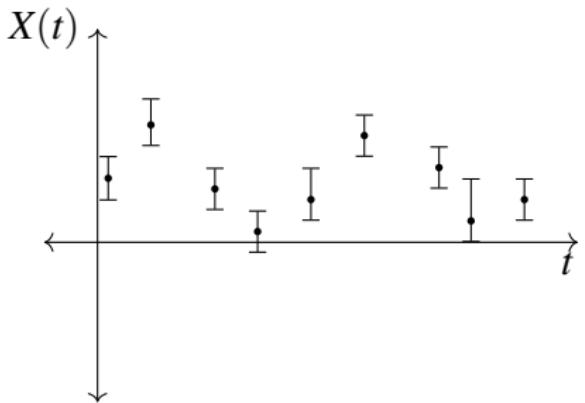
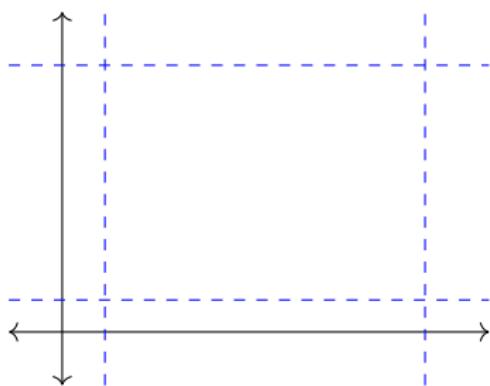
Model  $M_2$



Reject  $(M, \theta)$  if  $d > \epsilon$   
Accept  $(M, \theta)$  if  $d \leq \epsilon$

# ABC FOR MODEL SELECTION

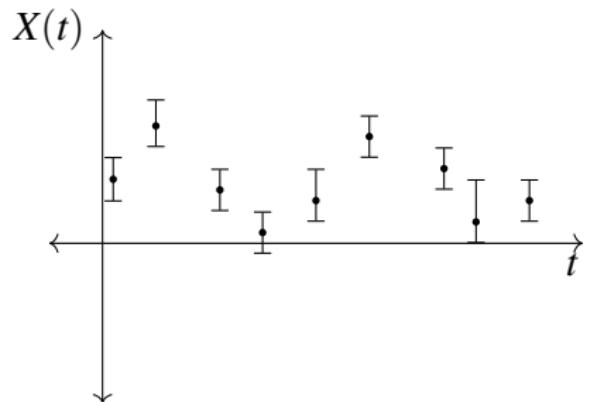
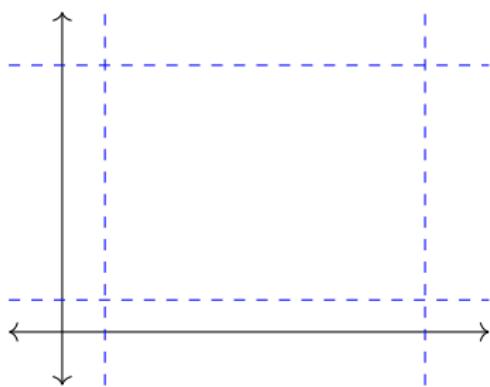
	$M_1$	$M_2$	$M_3$	$M_4$
Model accepted	-	0	-	-
Model ?				



# ABC FOR MODEL SELECTION

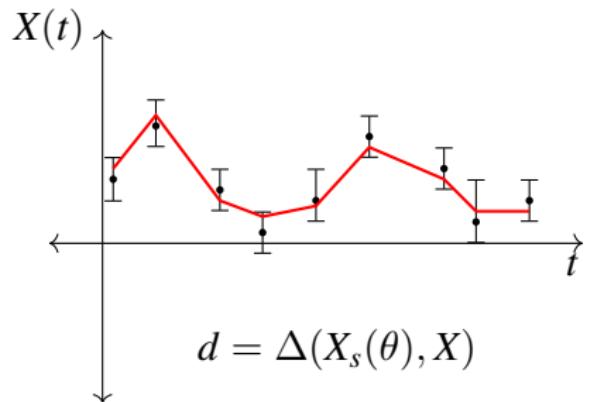
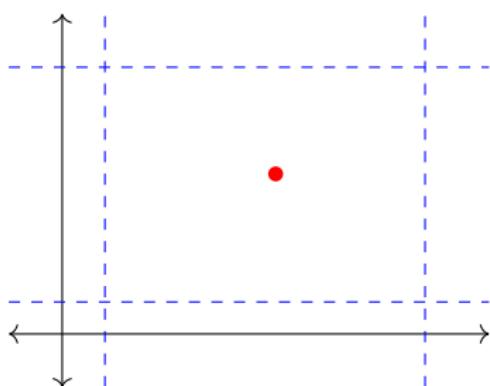
	$M_1$	$M_2$	$M_3$	$M_4$
Model accepted	-	0	-	-

Model 3



# ABC FOR MODEL SELECTION

	$M_1$	$M_2$	$M_3$	$M_4$
Model accepted	-	0	1	-



Reject  $(M, \theta)$  if  $d > \epsilon$   
Accept  $(M, \theta)$  if  $d \leq \epsilon$

## ABC FOR MODEL SELECTION

At the end, we obtain

	$M_1$	$M_2$	$M_3$	$M_4$
Model accepted	3000	70	20	800

and we infer that the first model is the best to explain the data.

## ABC FOR MODEL SELECTION

Posterior probability  $\pi(\mathcal{M} = m|x^*)$  is approximated by the frequency of acceptances from model  $m$

$$\frac{1}{N} \sum_{i=1}^N \mathbf{1}_{m^{(i)}=m} .$$

As with parameter inference,

- variants of the MCMC and SMC versions of ABC can be used for model selection
- regression-based variants of ABC can be used to estimate the posterior probability of each model using multinomial logistic regression [Beaumont, 2009] or logistic regression with neural network [Blum & Francois, 2010]

## POTENTIAL ISSUES

- Should tolerance  $\epsilon$  be the same for all models?
- Should summary statistics vary across models?
- Should the distance measure vary as well?

## ABC APPROXIMATION TO A BAYES FACTOR

Consider two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$  and an observation  $x^*$ . To discriminate between the two models, we wish to compute the Bayes factor

$$B_{1,2} = \frac{\pi(x^*|\mathcal{M}_1)\pi(\mathcal{M}_1)}{\pi(x^*|\mathcal{M}_2)\pi(\mathcal{M}_2)} = \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \frac{\int \pi(\theta_1)f_1(x^*|\theta_1)d\theta_1}{\int \pi(\theta_2)f_2(x^*|\theta_2)d\theta_2}$$

The ABC approximation to the Bayes factor resulting from the previous algorithm is

$$\hat{B}_{1,2} = \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \frac{\sum_{t=1}^T \mathbb{1}_{m^{(t)}=1} \mathbb{1}_{d(\eta(z^{(t)}), \eta(x^*)) \leq \epsilon}}{\sum_{t=1}^T \mathbb{1}_{m^{(t)}=2} \mathbb{1}_{d(\eta(z^{(t)}), \eta(x^*)) \leq \epsilon}}$$

where the pair  $(m^{(t)}, z^{(t)})$  is simulated from the joint prior and  $T$  is the number of simulations necessary for  $N$  acceptance.

[Robert et al, 2011]

## ABC APPROXIMATION TO A BAYES FACTOR

The ABC approximation to the Bayes factor resulting from the previous algorithm is

$$\hat{B}_{1,2} = \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \frac{\sum_{t=1}^T \mathbb{1}_{m^{(t)}=1} \mathbb{1}_{d(\eta(z^{(t)}), \eta(x^*)) \leq \epsilon}}{\sum_{t=1}^T \mathbb{1}_{m^{(t)}=2} \mathbb{1}_{d(\eta(z^{(t)}), \eta(x^*)) \leq \epsilon}}$$

where the pair  $(m^{(t)}, z^{(t)})$  is simulated from the joint prior and  $T$  is the number of simulations necessary for  $N$  acceptance.

Taking the limit when  $T$  goes to infinity:

$$\begin{aligned}\hat{B}_{1,2} \rightarrow B_{1,2}^\epsilon &= \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \frac{\iint \pi(\theta_1) f_1(z|\theta_1) \mathbb{1}_{d(\eta(z), \eta(x^*)) \leq \epsilon} dz d\theta_1}{\iint \pi(\theta_2) f_2(z|\theta_2) \mathbb{1}_{d(\eta(z), \eta(x^*)) \leq \epsilon} dz d\theta_2} \\ &= \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \frac{\iint \pi(\theta_1) f_1^\eta(\boldsymbol{\eta}|\theta_1) \mathbb{1}_{d(\boldsymbol{\eta}, \eta(x^*)) \leq \epsilon} d\boldsymbol{\eta} d\theta_1}{\iint \pi(\theta_2) f_2^\eta(\boldsymbol{\eta}|\theta_2) \mathbb{1}_{d(\boldsymbol{\eta}, \eta(x^*)) \leq \epsilon} d\boldsymbol{\eta} d\theta_2}\end{aligned}$$

where  $f_1^\eta(\boldsymbol{\eta}|\theta_1)$  and  $f_2^\eta(\boldsymbol{\eta}|\theta_2)$  are the density of  $\eta(z)$  when  $z \sim f_1(z|\theta_1)$  and  $z \sim f_2(z|\theta_2)$ .

[Robert et al, 2011]

## ABC APPROXIMATION TO A BAYES FACTOR

Taking the limit when  $T$  goes to infinity:

$$\begin{aligned}\hat{B}_{1,2} \rightarrow B_{1,2}^\epsilon &= \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \frac{\iint \pi(\theta_1) f_1(z|\theta_1) \mathbb{1}_{d(\eta(z), \eta(x^*)) \leq \epsilon} dz d\theta_1}{\iint \pi(\theta_2) f_2(z|\theta_2) \mathbb{1}_{d(\eta(z), \eta(x^*)) \leq \epsilon} dz d\theta_2} \\ &= \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \frac{\iint \pi(\theta_1) f_1^\eta(\eta|\theta_1) \mathbb{1}_{d(\eta, \eta(x^*)) \leq \epsilon} d\eta d\theta_1}{\iint \pi(\theta_2) f_2^\eta(\eta|\theta_2) \mathbb{1}_{d(\eta, \eta(x^*)) \leq \epsilon} d\eta d\theta_2}\end{aligned}$$

where  $f_1^\eta(\eta|\theta_1)$  and  $f_2^\eta(\eta|\theta_2)$  are the density of  $\eta(z)$  when  $z \sim f_1(z|\theta_1)$  and  $z \sim f_2(z|\theta_2)$ .

When  $\epsilon$  goes to 0,

$$B_{1,2}^\epsilon \rightarrow B_{1,2}^\eta = \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \frac{\int \pi(\theta_1) f_1^\eta(\eta(x^*)|\theta_1) d\theta_1}{\int \pi(\theta_2) f_2^\eta(\eta(x^*)|\theta_2) d\theta_2}$$

which depends only on the distribution of  $\eta$  under both models.

[Robert et al, 2011]

## SUFFICIENT STATISTICS AND MODEL SELECTION

- If a summary statistic  $\eta(\cdot)$  is sufficient for a model then  $\pi(\theta|\eta(x^*)) = \pi(\theta|x^*)$ .
- **Warning !**  $\eta(\cdot)$  being sufficient for both  $f_1(x^*|\theta_1)$  and  $f_2(x^*|\theta_2)$  does not usually implies that  $\eta(\cdot)$  is sufficient for model selection.
- Explanation: for such a statistic, there exist  $g_1(\cdot)$  and  $g_2(\cdot)$  such that

$$f_i(x^*|\theta_i) = g_i(x^*) f_i^\eta(\eta(x^*)|\theta_i)$$

and

$$\begin{aligned} B_{1,2} &= \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \frac{\int \pi(\theta_1) f_1(x^*|\theta_1) d\theta_1}{\int \pi(\theta_2) f_2(x^*|\theta_2) d\theta_2} \\ &= \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \frac{g_1(x^*) \int \pi(\theta_1) f_1^\eta(\eta(x^*)|\theta_1) d\theta_1}{g_2(x^*) \int \pi(\theta_2) f_2^\eta(\eta(x^*)|\theta_2) d\theta_2} = \frac{g_1(x^*)}{g_2(x^*)} B_{1,2}^\eta \end{aligned}$$

[Robert et al, 2011]

# HOW TO SELECT SUMMARY STATISTICS?

- Previous result implies that ABC model selection will only give accurate estimate of  $B_{1,2}$  if  $g_1(x^*)/g_2(x^*) = 1$ .
- This holds in the case where models  $\mathcal{M}_1$  and  $\mathcal{M}_2$  are nested submodels of a model  $\mathcal{M}$  for which  $\eta(\cdot)$  is sufficient.  
*[Didelot et al, 2011]*
- A necessary condition for an ABC model choice algorithm to converge to the true model:  
as the sample size increases, the mean of the posterior predictive distribution of the summary statistics converges to different values under different models. *[Marin et al, 2014]*

## EXAMPLE: LAPLACE VS NORMAL

Consider the two following models with equal prior:

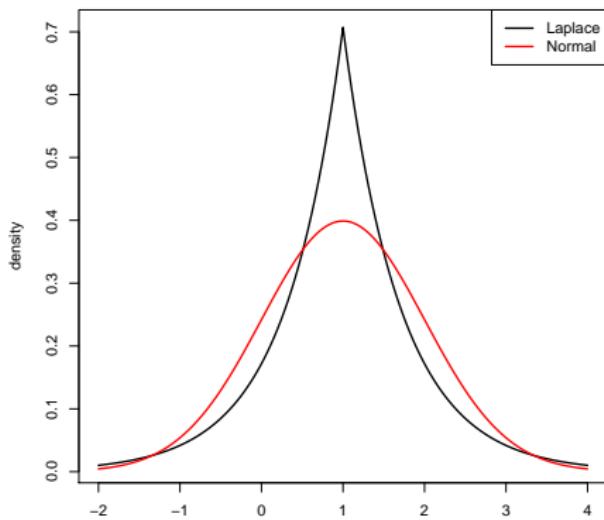
$$\mathcal{M}_1 : y \sim \mathcal{N}(\theta_1, 1) \quad \text{and} \quad \mathcal{M}_2 : y \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$$

Prior distribution:  $\mathcal{N}(0, 4)$

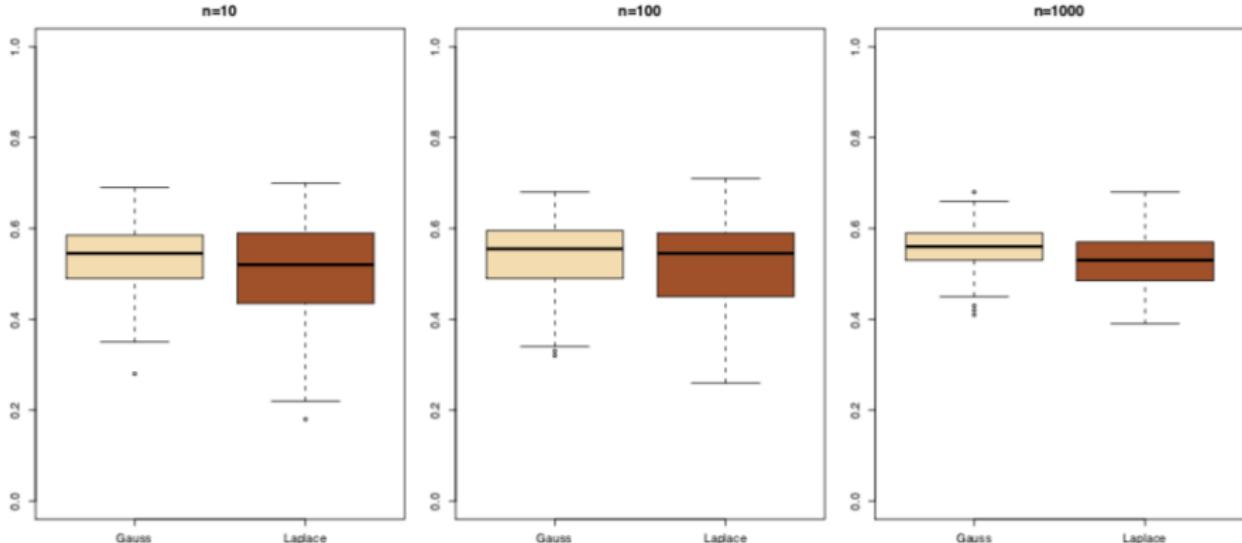
Summary statistics:

- sample mean
- sample median
- sample variance
- median absolute deviation
- sample fourth moment

sufficient statistics, ancillary statistics,  
statistic for model selection



# EXAMPLE: LAPLACE VS NORMAL



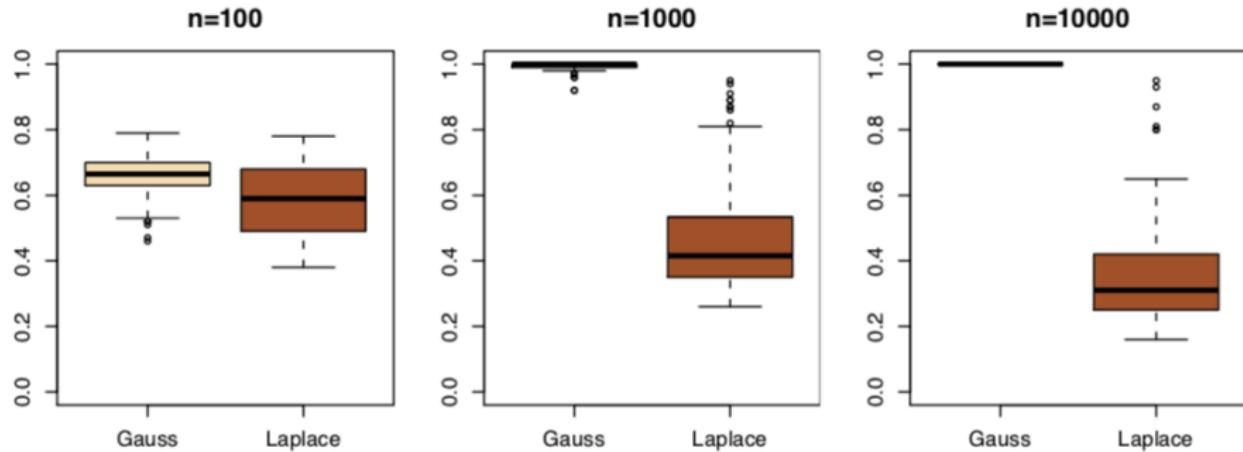
Data: sample of  $n$  observations either from a Gaussian or Laplace distribution with mean 0;

Summary statistics: sample mean, median and variance;

ABC : 5000 simulations for each model and  $\epsilon=1\%$  distance quantile.

[Marin et al, 2014]

# EXAMPLE: LAPLACE VS NORMAL



Summary statistics: empirical fourth moment.

For Normal:  $\theta^4 + 3 + 6\theta^2$  and for Laplace:  $\theta^4 + 6 + 6\theta^2$

[Marin et al, 2014]

# HOW TO SELECT SUMMARY STATISTICS?

Theoretical result from [Marin et al, 2014] motivated approaches to discriminate models.

Recent approaches include:

- Simulating samples from posterior predictive distributions of the summary statistics and select statistics with different means under different models [Marin et al, 2014]
- Logistic regression to estimate posterior probability of of summary statistics from pilot simulations [Prangle et al, 2014]
- Linear discriminant analysis to maximize separability of the models [Estoup et al, 2012]
- Random forest [Pudlo et al, 2015]

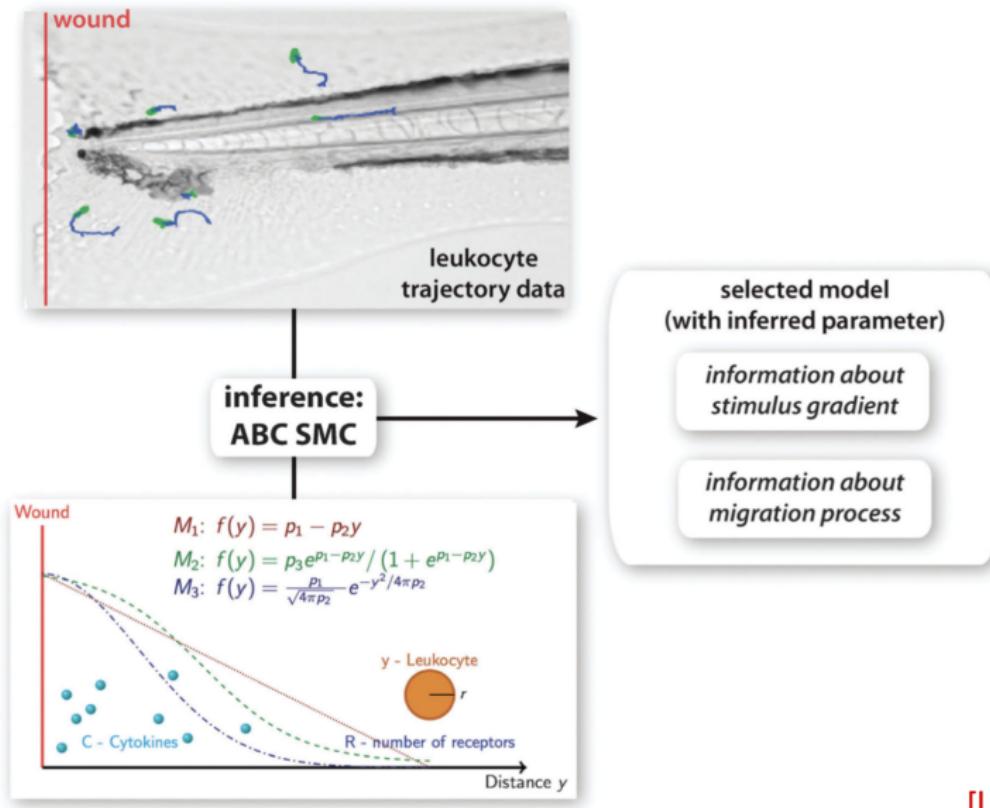
[Beaumont et al, 2019]

## EXAMPLE: LEUKOCYTE MIGRATION

- Stimulus released from injury
- Observation: movements of the leukocytes
- Models to describe the leukocyte dynamics with 3 different stimulus gradient shapes

[Liepe et al, 2012]

# EXAMPLE: LEUKOCYTE MIGRATION



[Liepe et al, 2012]

# TAKE-HOME MESSAGES

## APPROXIMATE BAYESIAN COMPUTATION

To be used for **parameter inference** and **model selection** when the likelihood can not be computed but it is possible to simulate from the model.

## SIMPLE APPROACH

Only requires:

- a simulator
- a distance
- a threshold
- computational resources

Warning ! Be careful with choice of summary statistic.

## COMPUTATIONAL EFFICIENT APPROACHES

- ABC-MCMC
- Sequential ABC – with an optimal choice of perturbation kernel

# REFERENCES

## Basics

- Rubin, *Annals of Statistics*, 1984
- Tavaré, Balding and Griths, *Genetics*, 1997
- Pritchard et al, *Molec. Biol. Evol.*, 1999.
- Beaumont, Zhang, Balding, *Genetics*, 2002
- Wilkinson, Tavaré, *Theoretical Population Biology*, 2009.
- Fearnhead and Prangle, *JRSS Ser. B*, 2012
- Wilkinson, *SAGMB*, 2013.

## Review articles

- Marin, Pudlo. Robert, and Ryder, *Statistics and Computing*, 2011.
- Lintusaari et al , *System Biology*, 2017.
- Beaumont, *Annu. Rev. Stat. Appl.*, 2019.

# REFERENCES

## Theoretical results

- Frazier et al, *Arxiv*, 2015.
- Frazier, Robert and Rousseau, *Arxiv*, 2017.
- Frazier, Martin, Robert and Rousseau, *Biometrika*, 2018.

## Post-processing

- Beaumont et al, *Genetics*, 2002.
- Blum and Francois, *Statist. Comput.*, 2010.
- Blum, *J. American Statist. Assoc.*, 2010.
- Nakagome et al, *SAGMB*, 2013.
- Leuenberger and Wegmann, *Genetics*, 2010.
- Fan et al, *Statistics*, 2013.

## Applications

- Tanaka, Francis, Luciani and Sisson, *Genetics*, 2006.
- Liepe et al, *Integr. Biol.*, 2012

## REFERENCES

### MCMC or Sequential approaches

- Marjoram, Molitor, Plagnol, and Tavaré, *PNAS*, 2003.
- Cappe et al, *J. Comput. Graph. Statist.*, 2004.
- Del Moral Doucet, and Jasra. *JRSS Series B*, 2006.
- Ratmann et al, *PLoS Comput Biol*, 2007.
- Sisson et al, *PNAS*, 2007.
- Toni et al. *JRSS Interface*, 2009.
- Toni and Stumpf, *Bioinformatics*, 2010.
- Beaumont., Cornuet, Marin and Robert, *Biometrika*, 2009.
- Drovandi et al, *Biometrics*, 2011.
- Filippi, Barnes, Cornebise, and Stumpf. *SAGMB*, 2013.
- Silk, Filippi and Stumpf. *SAGMB*, 2013.

# REFERENCES

## Summary statistics

- Didelot et al, *Bayesian analysis*, 2011.
- Fearnhead and Prangle, *JRSS: Series B, (With discussion.)*, 2012.
- Robert, Cornuet, Marin and Pillai, *PNAS*, 2011.
- Estoup et al, *Mol, Ecol, Resourc.*, 2012.
- Marin et al, *JRSS*, 2014.
- Prangle et al, *SAGMB*, 2014.
- Pudlo et al, *Bioinformatics*, 2015.