Efficient Likelihood-Free Inference

Michael Gutmann

http://homepages.inf.ed.ac.uk/mgutmann

Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh

8th November 2017

Problem statement

Likelihood-free inference:

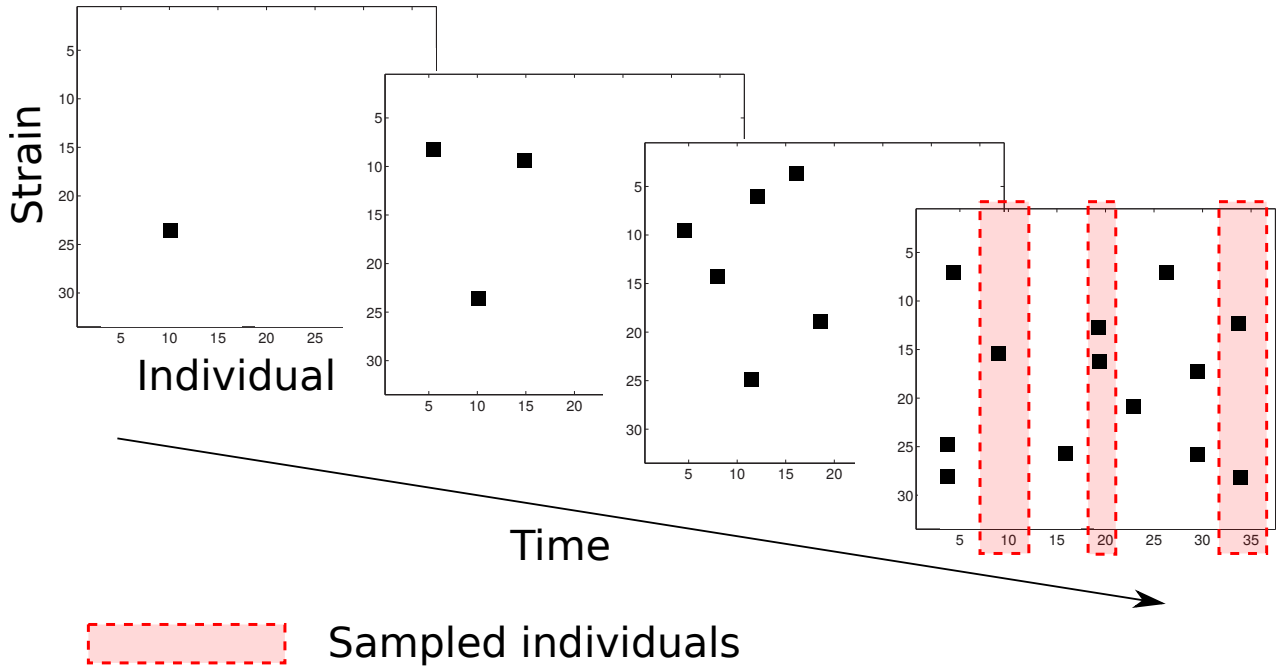
Perform statistical inference for models where

1. the likelihood function is too costly to evaluate
2. sampling – simulating data – from the model is possible

Importance

Such models and inference problems occur widely

- Neuroscience:
  Simulating neural circuits
- Evolutionary biology:
  Simulating evolution
- Computer vision:
  Simulating naturalistic scenes
- Health science:
  Simulating the spread of an infectious disease



Program

Background

Previous work

Our approach

Program

Background

Previous work

Our approach

Assumptions on the models

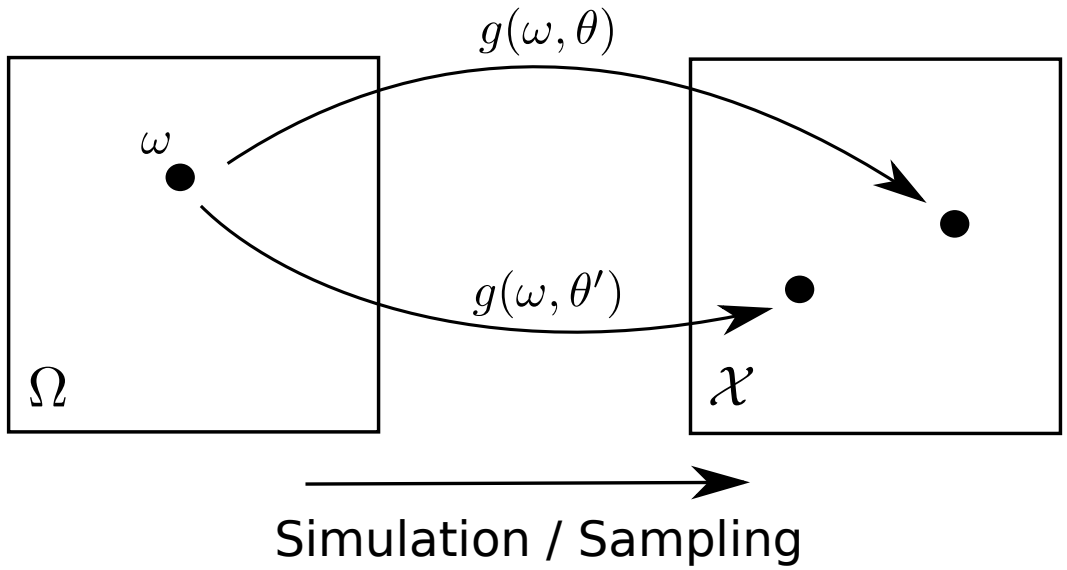
- Only assumption: sampling – simulating data – from the model is possible
- Models specified by a data generating mechanism
  - e.g. stochastic nonlinear dynamical systems
  - e.g. computer models / simulators of some complex physical or biological process
- Different communities use different names:
  - Simulator-based models
  - Stochastic simulation models
  - Implicit models
  - Generative (latent-variable) models
  - Probabilistic programs

Definition of simulator-based models (SBMs)

- Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space.
- A simulator-based model is a collection of (measurable) functions $g(., \boldsymbol{\theta})$ parametrised by $\boldsymbol{\theta}$,

$$\boldsymbol{\omega} \in \Omega \mapsto \boldsymbol{x}_{\boldsymbol{\theta}} = g(\boldsymbol{\omega}, \boldsymbol{\theta}) \in \mathcal{X} \tag{1}$$

- For any fixed $\boldsymbol{\theta}$, $\boldsymbol{x}_{\boldsymbol{\theta}} = g(., \boldsymbol{\theta})$ is a random variable.
- $g(., \boldsymbol{\theta})$ typically not available in closed form



Strengths of SBMs

- Direct implementation of hypotheses of how the observed data were generated.
- Neat interface with scientific models (e.g. from physics or biology).
- Modelling by replicating the mechanisms of nature that produced the observed/measured data. ("Analysis by synthesis")
- Possibility to perform experiments in silico.

Weaknesses of SBMs

- Generally elude analytical treatment.
- Can be easily made more complicated than necessary.
- Statistical inference is difficult.

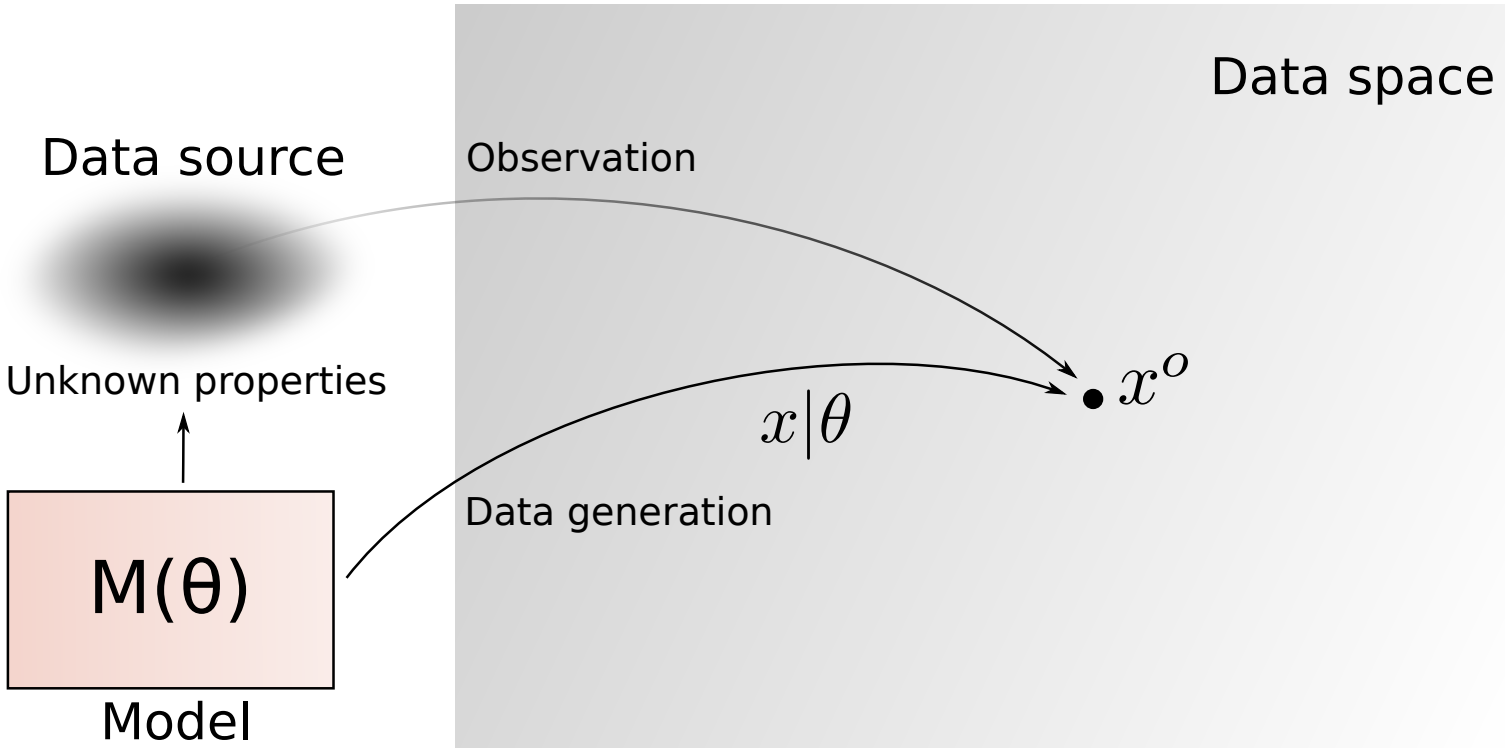
Weaknesses of SBMs

- Generally elude analytical treatment.
- Can be easily made more complicated than necessary.
- Statistical inference is difficult.

  Main reason: *Likelihood function is too expensive to evaluate*

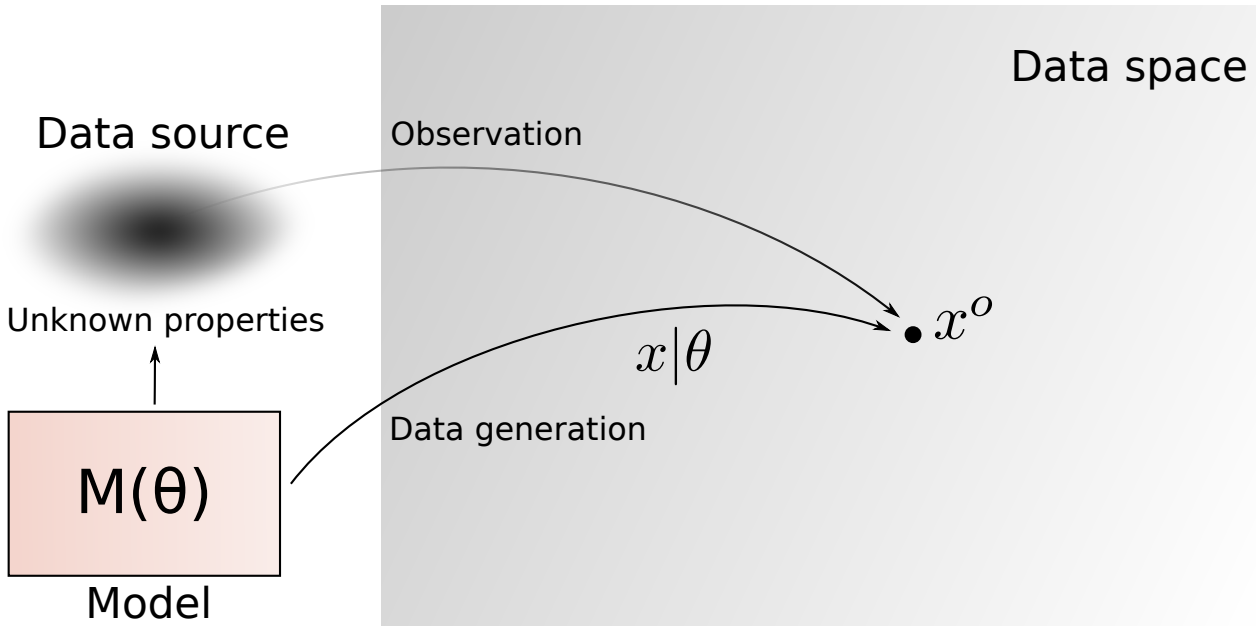
The likelihood function $L(\boldsymbol{\theta})$

- Well defined but generally intractable for SBMs
- Probability that the model generates data like $\boldsymbol{x}^o$ when using parameter value $\boldsymbol{\theta}$



Three foundational issues

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the probability of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
3. For which values of $\boldsymbol{\theta}$ should we compute it?



Likelihood: Probability that the model generates data like $\boldsymbol{x}^o$ for parameter value $\boldsymbol{\theta}$

Program

Background

Previous work

Our approach

Approximate Bayesian computation

For recent review, see: Lintusaari et al (2017) "Fundamentals and recent developments in approximate Bayesian computation", Systematic Biology

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ Check whether $||T(\boldsymbol{x}_{\boldsymbol{\theta}}) - T(\boldsymbol{x}^o)|| \leq \epsilon$
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ By counting
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   $\Rightarrow$ Sample from the prior (or other proposal distributions)

Approximate Bayesian computation

For recent review, see: Lintusaari et al (2017) "Fundamentals and recent developments in approximate Bayesian computation", Systematic Biology

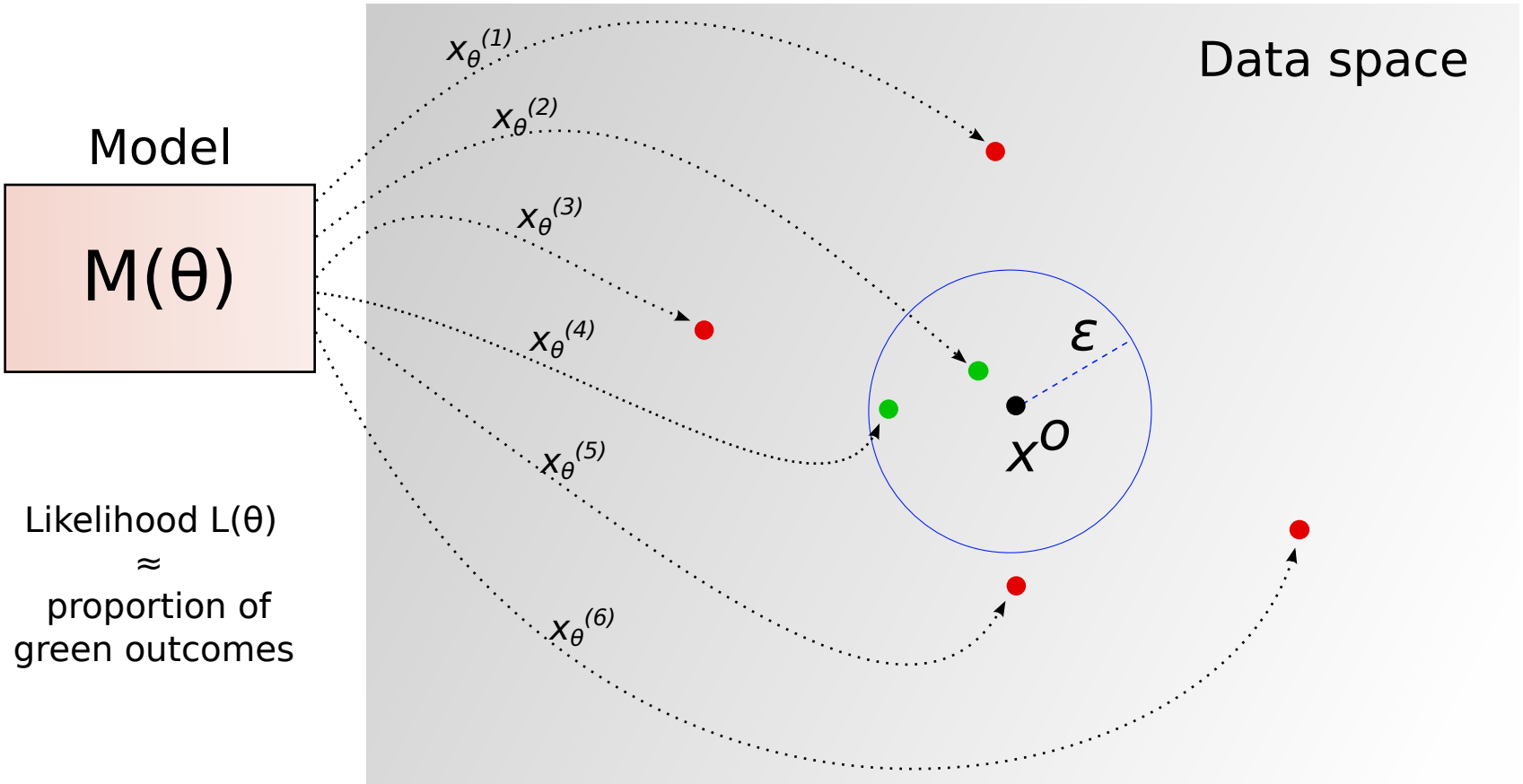
1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ Check whether $||T(\boldsymbol{x}_{\boldsymbol{\theta}}) - T(\boldsymbol{x}^o)|| \leq \epsilon$
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ By counting
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   $\Rightarrow$ Sample from the prior (or other proposal distributions)

Difficulties:

- Choice of $T()$ and ϵ
- Typically high computational cost

Implicit likelihood approximation

Likelihood: Probability to generate data like x^o for parameter value θ



$$L(\boldsymbol{\theta}) \approx \mathbb{P}^N(d(\boldsymbol{x_\theta}, \boldsymbol{x^o}) \leq \epsilon) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left(d(\boldsymbol{x_\theta^{(i)}}, \boldsymbol{x^o}) \leq \epsilon\right)$$

Synthetic likelihood

(Simon Wood, Nature, 2010)

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Compute summary statistics $\boldsymbol{t}_{\boldsymbol{\theta}} = T(\boldsymbol{x}_{\boldsymbol{\theta}})$
   - $\Rightarrow$ Model their distribution as a Gaussian
   - $\Rightarrow$ Compute likelihood function with $T(\boldsymbol{x}^o)$ as observed data
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   - $\Rightarrow$ Use obtained "synthetic" likelihood function as part of a Monte Carlo method

Synthetic likelihood

(Simon Wood, Nature, 2010)

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Compute summary statistics $\boldsymbol{t}_{\boldsymbol{\theta}} = T(\boldsymbol{x}_{\boldsymbol{\theta}})$
   - $\Rightarrow$ Model their distribution as a Gaussian
   - $\Rightarrow$ Compute likelihood function with $T(\boldsymbol{x}^o)$ as observed data
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   - $\Rightarrow$ Use obtained "synthetic" likelihood function as part of a Monte Carlo method

Difficulties:

- Choice of $T()$
- Gaussianity assumption may not hold
- Typically high computational cost

Program

Background

Previous work

Our approach

Overview of some of my work

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ Use classification (Gutmann et al, 2014, 2017)

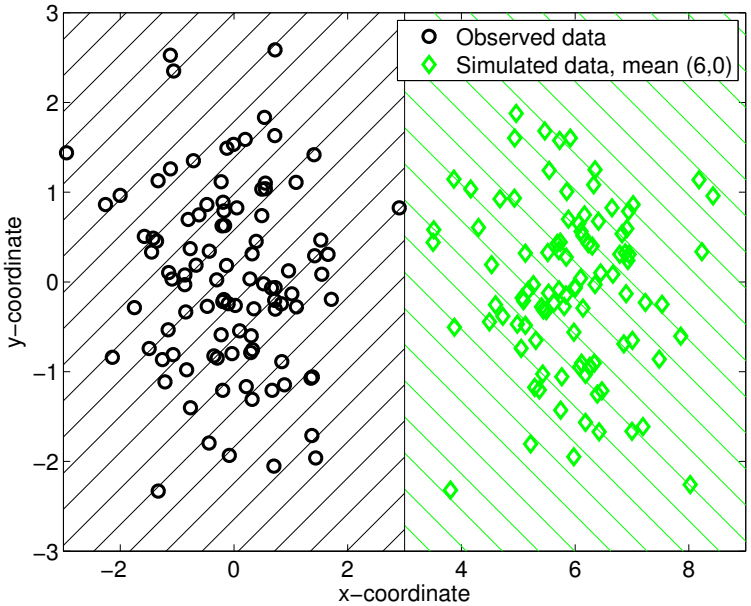
1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ Use density ratio estimation (Dutta et al, 2016, arXiv:1611.10242)

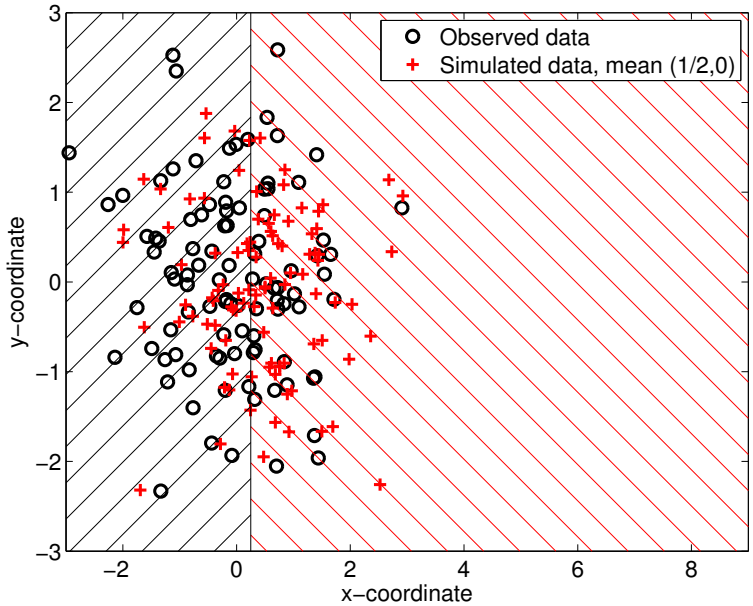
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   $\Rightarrow$ Use Bayesian optimisation (Gutmann and Corander, 2013-2016)

Overview of some of my work

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ Use classification (Gutmann et al, 2014, 2017)

- Basic idea: Classification accuracy (discriminability) serves as distance measure
- Value of 1: far; Value of 1/2: close





Overview of some of my work

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?

   $\Rightarrow$ Use density ratio estimation (Dutta et al, 2016, arXiv:1611.10242)

- Basic idea: frame posterior estimation as ratio estimation problem

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})} = r(\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{\theta}) \qquad (2)$$

- Estimate $\hat{r}(\boldsymbol{x}, \boldsymbol{\theta})$ yields estimate of the likelihood function and posterior

$$\hat{L}(\boldsymbol{\theta}) \propto \hat{r}(\boldsymbol{x}^o, \boldsymbol{\theta}), \qquad \hat{p}(\boldsymbol{\theta}|\boldsymbol{x}^o) = \hat{r}(\boldsymbol{x}^o, \boldsymbol{\theta})p(\boldsymbol{\theta}). \qquad (3)$$

Overview of some of my work

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ Use classification (Gutmann et al, 2014, 2017)

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ Use density ratio estimation (Dutta et al, 2016, arXiv:1611.10242)

2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   $\Rightarrow$ Use Bayesian optimisation (Gutmann and Corander, 2013-2016)

Overview of some of my work

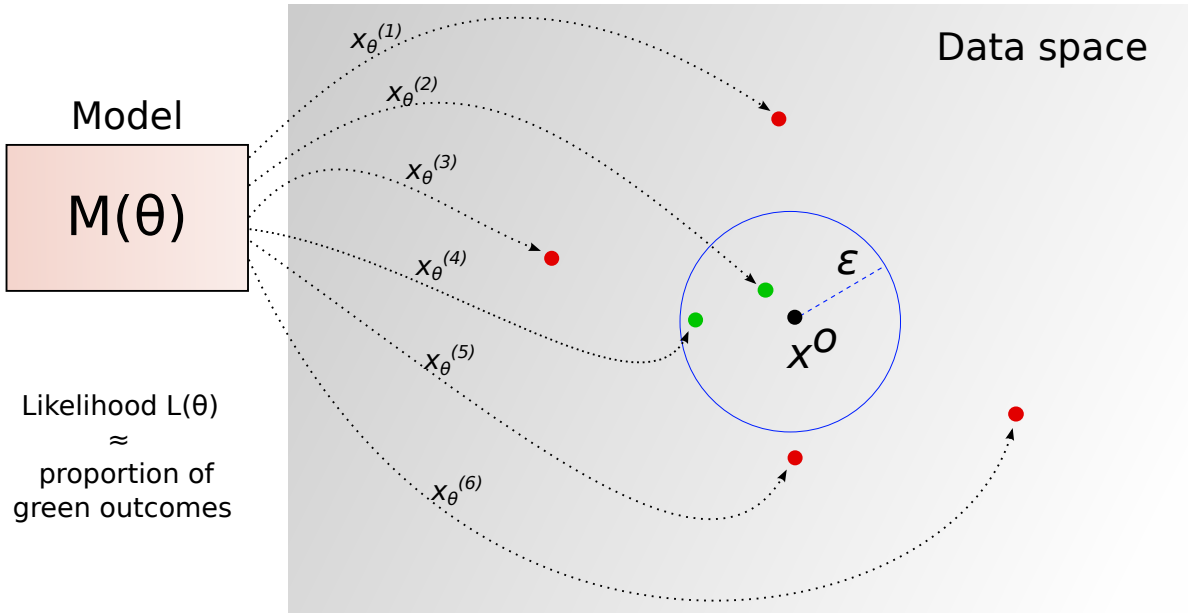
1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ Use classification (Gutmann et al, 2014, 2017)

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ Use density ratio estimation (Dutta et al, 2016, arXiv:1611.10242)

2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   $\Rightarrow$ Use Bayesian optimisation (Gutmann and Corander, 2013-2016)

Why is the ABC algorithm so expensive?

1. It rejects most samples when ϵ is small
2. It does not make assumptions about the shape of $L(\boldsymbol{\theta})$
3. It does not use all information available
4. It aims at equal accuracy for all parameters



$$L(\boldsymbol{\theta}) \approx \mathbb{P}^N(d(\boldsymbol{x_\theta}, \boldsymbol{x^o}) \leq \epsilon) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left(d(\boldsymbol{x_\theta^{(i)}}, \boldsymbol{x^o}) \leq \epsilon\right)$$

Proposed solution

(Gutmann and Corander, JMLR, 2016)

1. It rejects most samples when ϵ is small
   $\Rightarrow$ Don't reject samples – learn from them
2. It does not make assumptions about the shape of $L(\theta)$
   $\Rightarrow$ Model the distances, assume average distance is smooth
3. It does not use all information available
   $\Rightarrow$ Use Bayes' theorem to update the model
4. It aims at equal accuracy for all parameters
   $\Rightarrow$ Prioritize parameter regions with small distances

equivalent strategy applies to inference with synthetic likelihood

Modelling (points 1 & 2)

- Data are tuples $(\boldsymbol{\theta}_i, d_i)$, where $d_i = d(\boldsymbol{x}_{\boldsymbol{\theta}}^{(i)}, \boldsymbol{x}^o)$
- Model the conditional distribution of d given $\boldsymbol{\theta}$
- Estimated model yields approximation $\hat{L}(\boldsymbol{\theta})$ for any choice of ϵ

$$\hat{L}(\boldsymbol{\theta}) \propto \widehat{\mathbb{P}}\left(d \leq \epsilon \mid \boldsymbol{\theta}\right)$$

  $\widehat{\mathbb{P}}$ is probability under the estimated model.
- Here: Use (log) Gaussian process with squared exponential covariance function as model
- Approach not restricted to this model or Gaussian processes
  (comparison of different GP models: Järvenpää et al, 2016, arXiv:1610.06462)

Data acquisition (points 3 & 4)

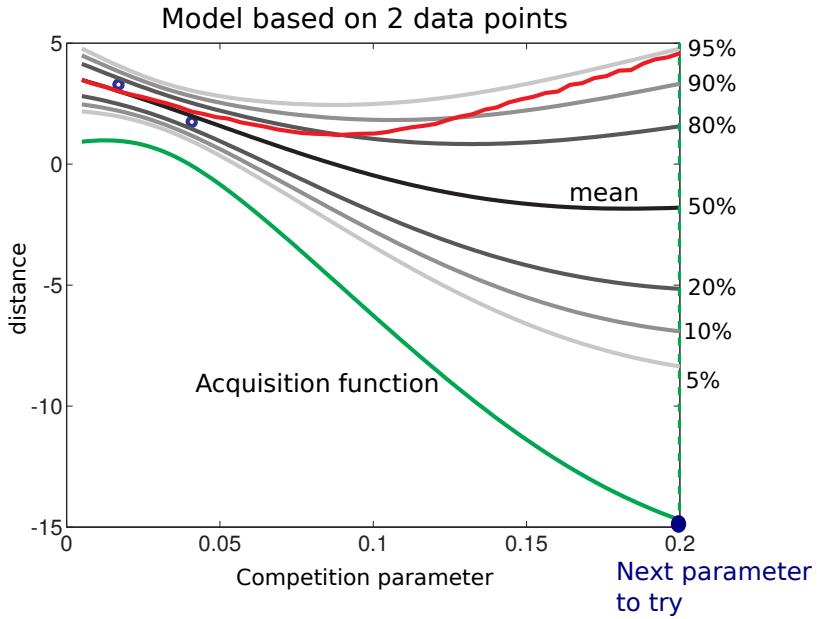
- Samples of θ could be obtained by sampling from the prior or some adaptively constructed proposal distribution
- Give priority to regions in the parameter space where distance d tends to be small.
- Use Bayesian optimization to find such regions
- Here: Use lower confidence bound acquisition function (e.g. Cox and John, 1992; Srinivas et al, 2012)

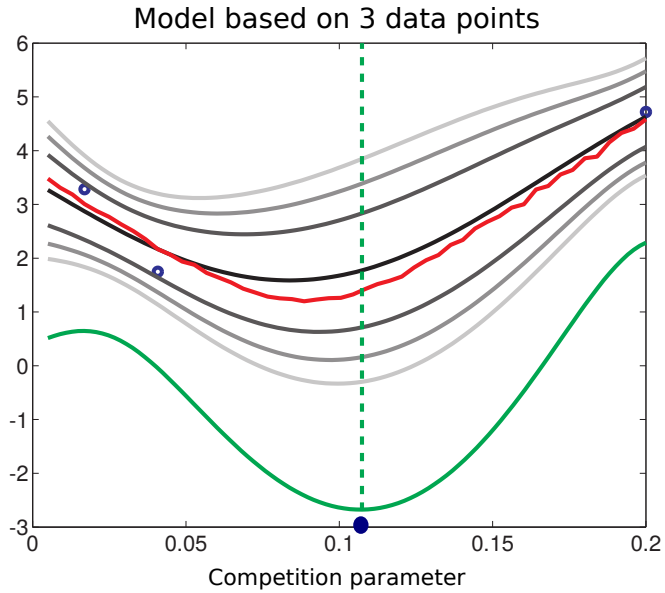
$$\mathcal{A}_t(\theta) = \underbrace{\mu_t(\theta)}_{\text{post mean}} - \sqrt{\underbrace{\eta_t^2}_{\text{weight}} \underbrace{v_t(\theta)}_{\text{post var}}} \tag{4}$$

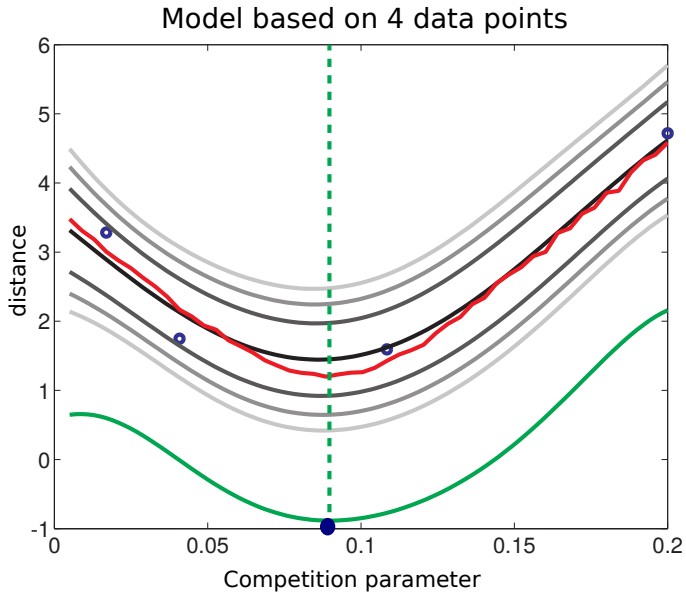
  t: number of samples acquired so far

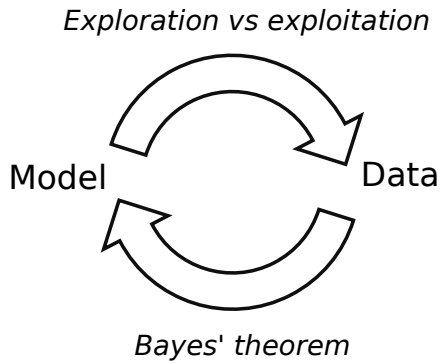
- Approach not restricted to this acquisition function.
  (new acquisition function: Järvenpää et al, 2017, arXiv:1704.00520)

Bayesian optimization for likelihood-free inference



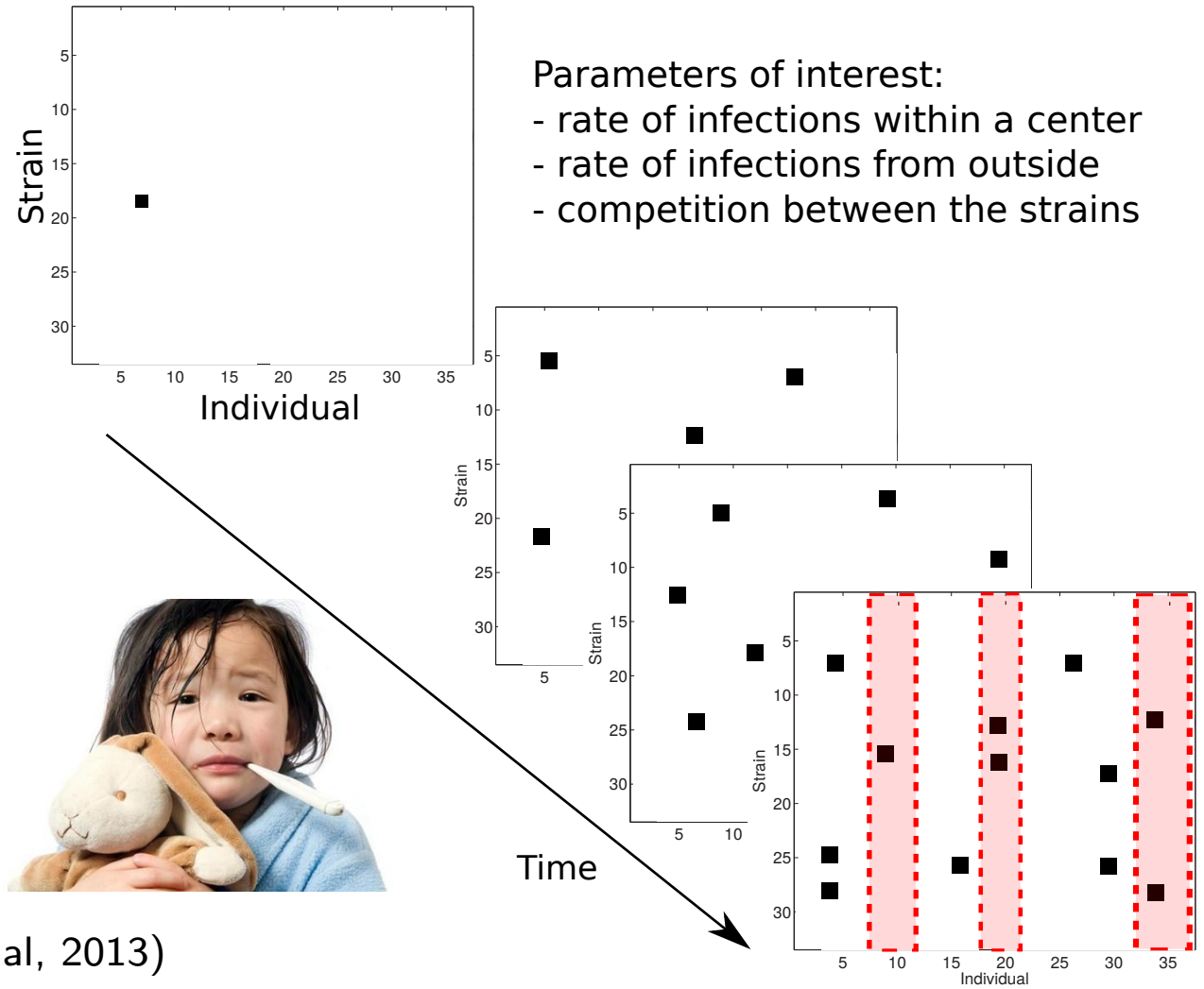






Example: Bacterial infections in child care centers

- Likelihood intractable for cross-sectional data
- But generating data from the model is possible



(Numminen et al, 2013)

Example: Bacterial infections in child care centers

- Comparison of the proposed approach with a standard population Monte Carlo ABC approach.
- Roughly equal results using 1000 times fewer simulations.

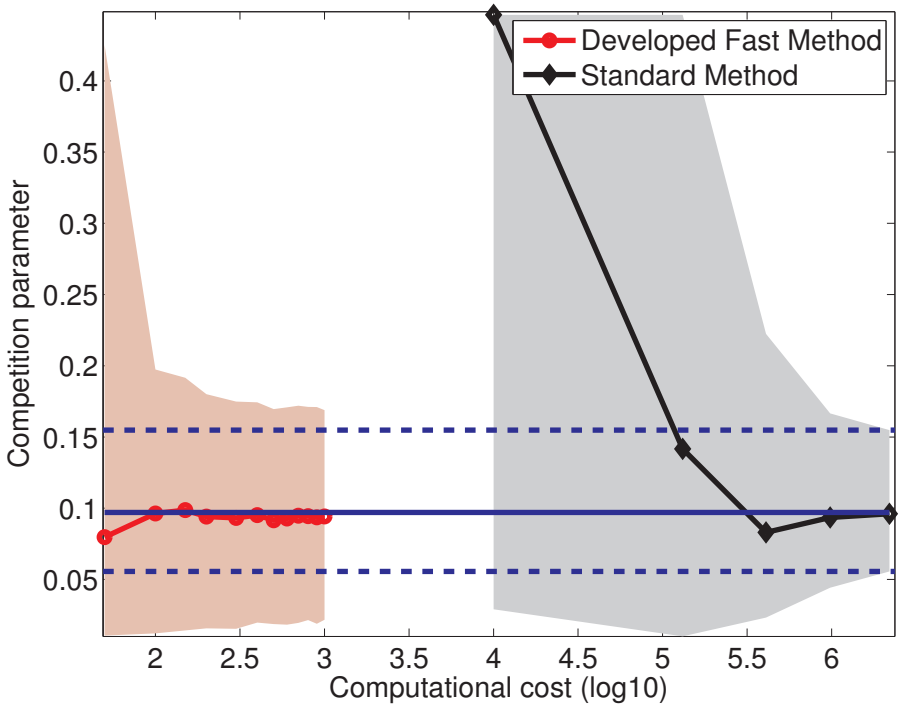
4.5 days with 200 cores

↓

90 minutes with seven cores

Posterior means: solid lines,

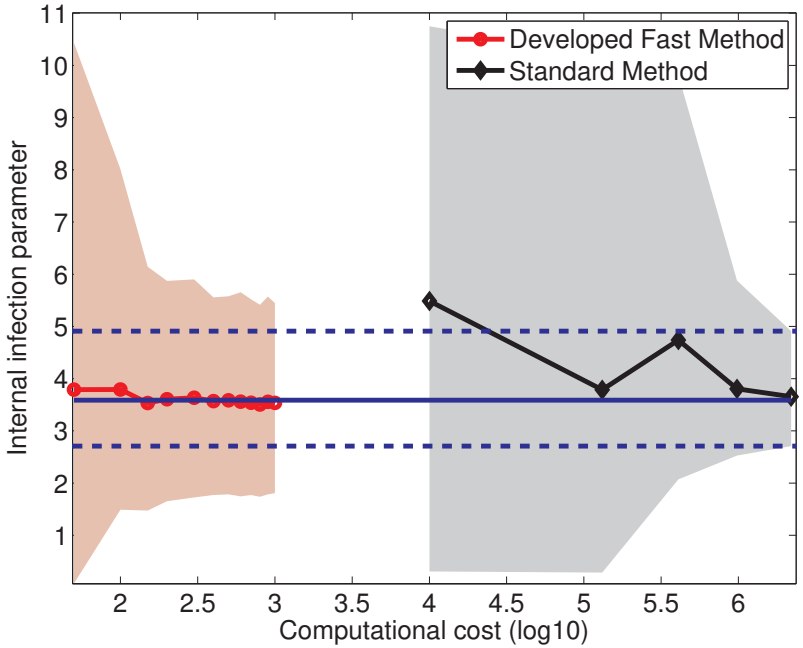
credibility intervals: shaded areas or dashed lines.

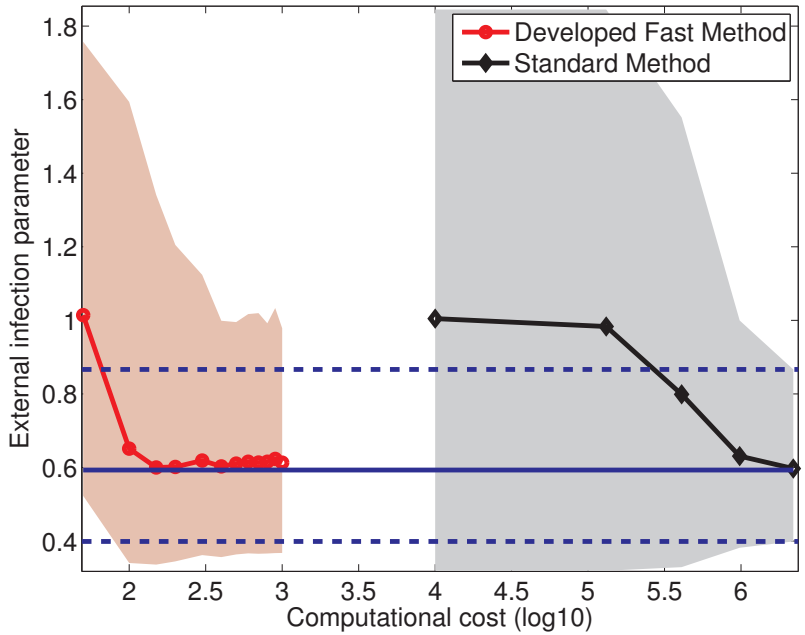


(Gutmann and Corander, JMLR, 2016)

Example: Bacterial infections in child care centers

- Comparison of the proposed approach with a standard population Monte Carlo ABC approach.
- Roughly equal results using 1000 times fewer simulations.





Posterior means are shown as solid lines, credibility intervals as shaded areas or dashed lines.

Benefits

- The proposed method makes the inference more efficient.
  - allowed us to perform far more comprehensive data analysis than with standard approach (Numminen et al, 2016)
- Enables inference for models which were out of reach till now
  - model of evolution where simulating a single data set took us 12-24 hours (Marttinen et al, 2015)
- Enables easier assessment of parameter identifiability for complex models
  - model about transmission dynamics of tuberculosis (Lintusaari et al, 2016)

Open questions

- Model: How to best model the distance between simulated and observed data?
- Acquisition function: Can we find strategies which are optimal for parameter inference?
- Efficient high-dimensional inference: Can we use the approach to infer the joint distribution of 1000 variables?

  see Gutmann and Corander, JMLR, 2016 for a discussion

  for first answers: `http://homepages.inf.ed.ac.uk/mgutmann`

Summary

- Topic: Inference for models where the likelihood is intractable but sampling is possible
- Inference principle: Find parameter values for which the distance between simulated and observed data is small
- Problem considered: Computational cost
- Proposed approach: Combine statistical modeling of the distance with decision making under uncertainty (Bayesian optimization)
- Outcome: Approach increases the efficiency of the inference by several orders of magnitude