Simulation-based Inference %>%

Concept and ELFI - tutorial %>%

ProbAI2022

Henri Pesonen
Oslo Center for Biostatistics and Epidemiology

@henri_pesonen

Concepts

- Bulding blocks of Likelihood-free inference
- Sampling based LFI methods
- Surrogate based methods
- Active learning

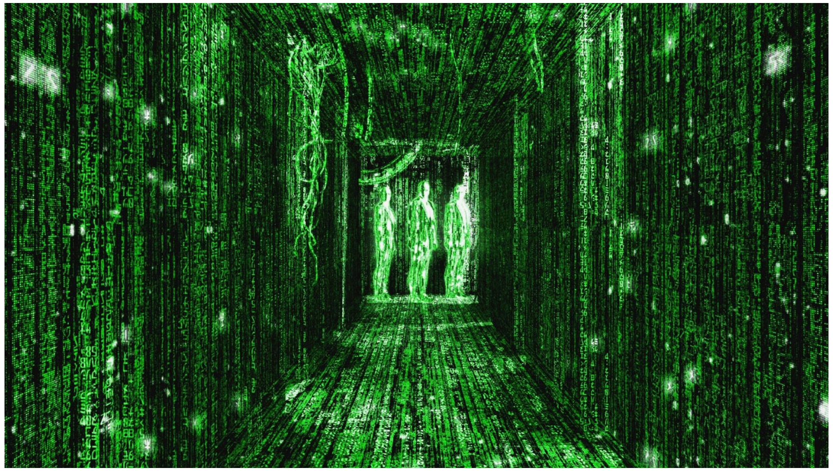
ELFI

Engine for Likelihood-Free Inference

Tutorial available as a notebook in Google Colab

- Basics of model building and exploration
- Choosing and using inference methods

Simulation-based Inference

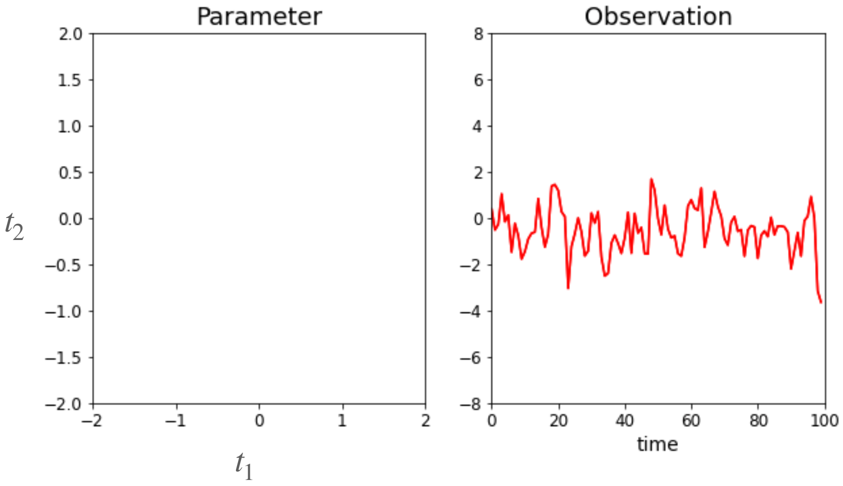


- The phenomena of the world are frequently investigated *in silico* - i.e. using complex computer simulations with great detail
- The simulators are controlled by a set of parameters that we want to infer based on the observations collected of the phenomena
- Complexity of the simulators often prohibits the access to the most important tool of statistical inference - likelihood function

Example

MA(2) model

- Simple time series model
- $x_t = w_t + t_1 w_{t-1} + t_2 w_{i-2}, \quad w_t \sim \text{Normal}(0,1)$



Example

Transmissions of bacterial infections in daycare centers.

- Cross-sectional data from a stochastic SIS-model
- Continuous-time Markov process with transition probabilities:

$$P(I_{is}(t+dt)=1\,|\,I_{is}(t)=0)=\theta_1\cdot E_s(I(t))+\theta_2\cdot P_s,\quad \text{if}\quad I_{i1}(t)+\cdots+I_{iN_s}(t)=0$$

$$P(I_{is}(t+dt)=1\,|\,I_{is}(t)=0)=\theta_3\cdot(\theta_1\cdot E_s(I(t))+\theta_2\cdot P_s),\quad \text{otherwise}$$

$$P(I_{is}(t+dt)=0\,|\,I_{is}(t)=1)=\gamma$$

- $I_{is}(t)$ is the status of carriage of strain s for individual i.
- $E_s(I(t))$ is the probability of sampling the strain s
- θ_1 is the rate of transmission from other children at the DCC
- θ_2 is the rate of transmission from the community outside the DCC
- θ_3 scales the rate of an infected child being infected with another strain
- γ is the relative probability of healing from a strain (scaled to 1)

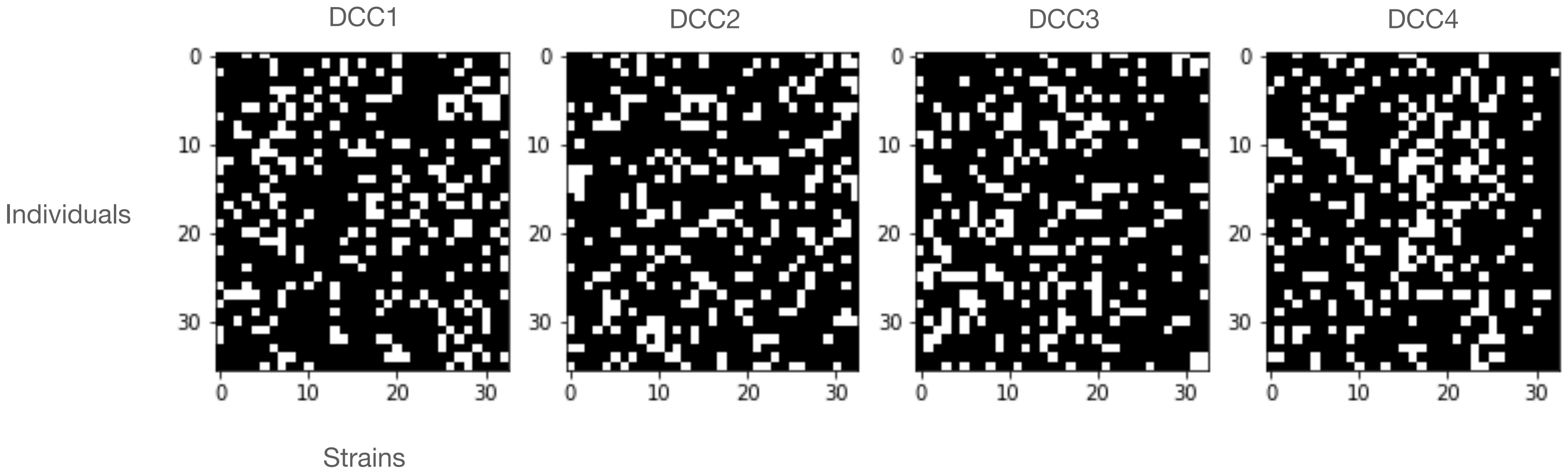
Example

Transmissions of bacterial infections in daycare centers.

$$P(I_{is}(t+dt)=1\,|\,I_{is}(t)=0)=\theta_1\cdot E_s(I(t))+\theta_2\cdot P_s, \quad \text{if} \quad I_{i1}(t)+\cdots+I_{iN_s}(t)=0$$

$$P(I_{is}(t+dt)=1\,|\,I_{is}(t)=0)=\theta_3\cdot(\theta_1\cdot E_s(I(t))+\theta_2\cdot P_s), \quad \text{otherwise}$$

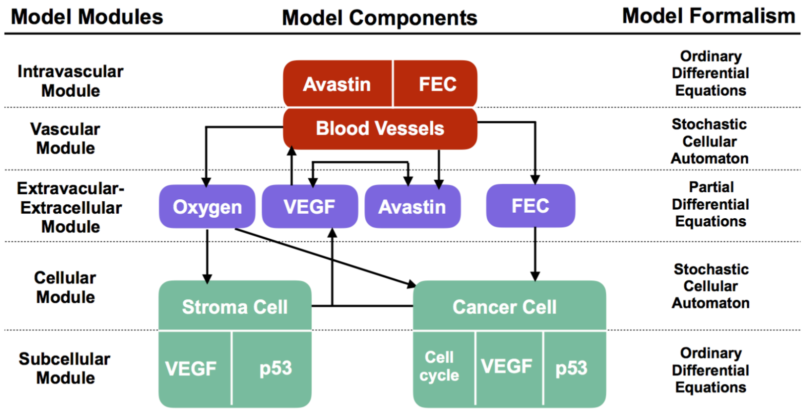
$$P(I_{is}(t+dt)=0\,|\,I_{is}(t)=1)=\gamma$$



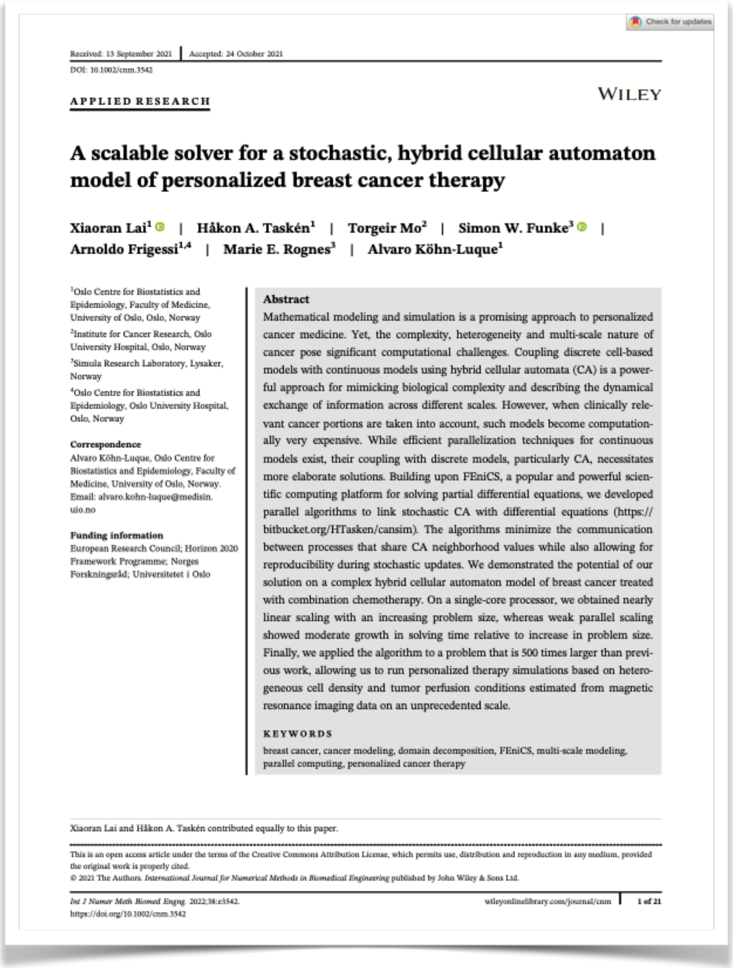
Example

Personalised medicine

- Model for evolution of breast cancer treated with combination chemotherapy
- Describes the evolution of cancer cells, blood vessels, Oxygen,VEGF and Avastin



Model Modules	Model Components	Model Formalism
Intravascular Module	Avastin, FEC	Ordinary Differential Equations
Vascular Module	Blood Vessels	Stochastic Cellular Automaton
Extravacular-Extracellular Module	Oxygen, VEGF, Avastin, FEC	Partial Differential Equations
Cellular Module	Stroma Cell, Cancer Cell	Stochastic Cellular Automaton
Subcellular Module	VEGF, p53; Cell cycle, VEGF, p53	Ordinary Differential Equations



Received: 13 September 2021 | Accepted: 24 October 2021

DOI: 10.1002/cnm.3542

APPLIED RESEARCH

WILEY

A scalable solver for a stochastic, hybrid cellular automaton model of personalized breast cancer therapy

Xiaoran Lai[1] | Håkon A. Taskén[1] | Torgeir Mo[2] | Simon W. Funke[3] | Arnoldo Frigessi[1,4] | Marie E. Rognes[3] | Alvaro Köhn-Luque[1]

[1]Oslo Centre for Biostatistics and Epidemiology, Faculty of Medicine, University of Oslo, Oslo, Norway

[2]Institute for Cancer Research, Oslo University Hospital, Oslo, Norway

[3]Simula Research Laboratory, Lysaker, Norway

[4]Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway

Correspondence
Alvaro Köhn-Luque, Oslo Centre for Biostatistics and Epidemiology, Faculty of Medicine, University of Oslo, Norway.
Email: alvaro.kohn-luque@medisin.uio.no

Funding information
European Research Council; Horizon 2020 Framework Programme; Norges Forskningsråd; Universitetet i Oslo

Abstract

Mathematical modeling and simulation is a promising approach to personalized cancer medicine. Yet, the complexity, heterogeneity and multi-scale nature of cancer pose significant computational challenges. Coupling discrete cell-based models with continuous models using hybrid cellular automata (CA) is a powerful approach for mimicking biological complexity and describing the dynamical exchange of information across different scales. However, when clinically relevant cancer portions are taken into account, such models become computationally very expensive. While efficient parallelization techniques for continuous models exist, their coupling with discrete models, particularly CA, necessitates more elaborate solutions. Building upon FEniCS, a popular and powerful scientific computing platform for solving partial differential equations, we developed parallel algorithms to link stochastic CA with differential equations (https://bitbucket.org/HTasken/cansim). The algorithms minimize the communication between processes that share CA neighborhood values while also allowing for reproducibility during stochastic updates. We demonstrated the potential of our solution on a complex hybrid cellular automaton model of breast cancer treated with combination chemotherapy. On a single-core processor, we obtained nearly linear scaling with an increasing problem size, whereas weak parallel scaling showed moderate growth in solving time relative to increase in problem size. Finally, we applied the algorithm to a problem that is 500 times larger than previous work, allowing us to run personalized therapy simulations based on heterogeneous cell density and tumor perfusion conditions estimated from magnetic resonance imaging data on an unprecedented scale.

KEYWORDS

breast cancer, cancer modeling, domain decomposition, FEniCS, multi-scale modeling, parallel computing, personalized cancer therapy

Xiaoran Lai and Håkon A. Taskén contributed equally to this paper.

Example

Personalised medicine

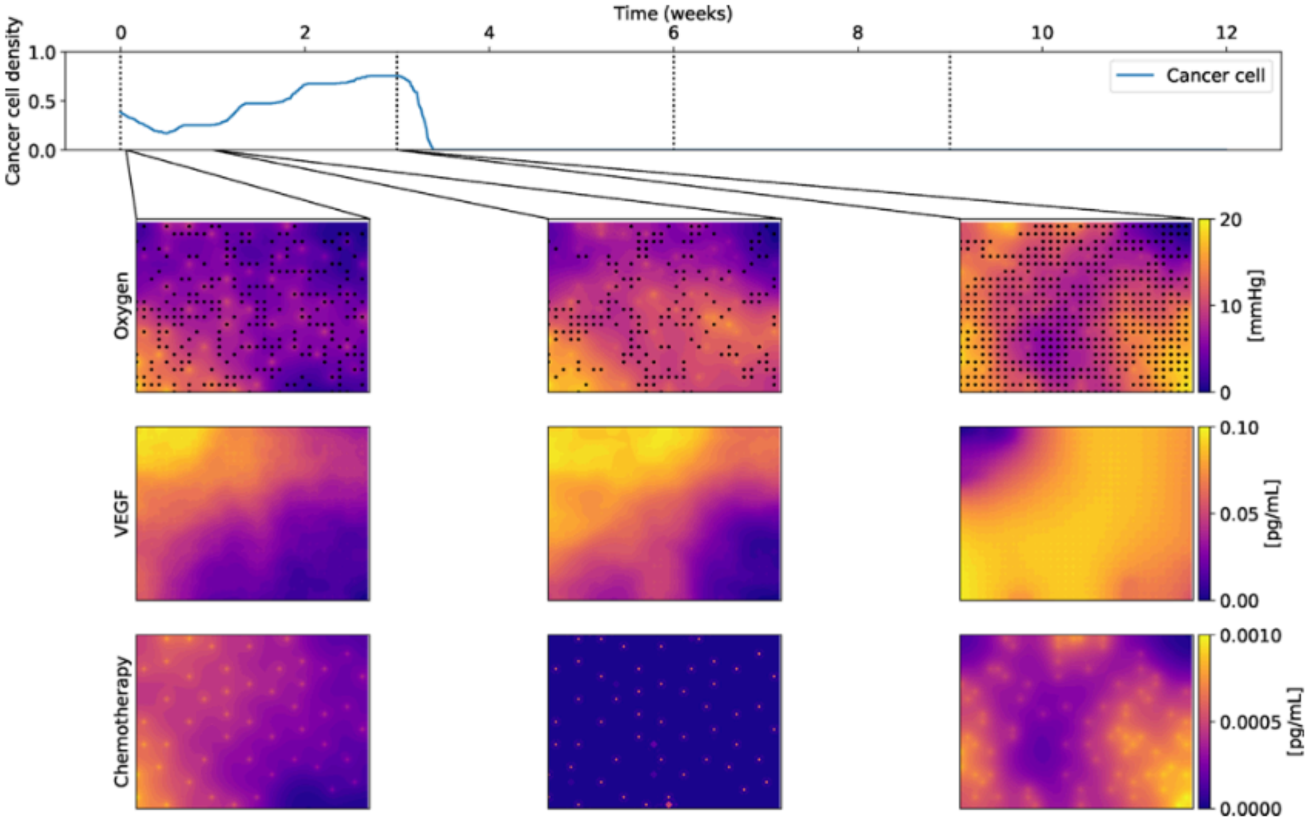
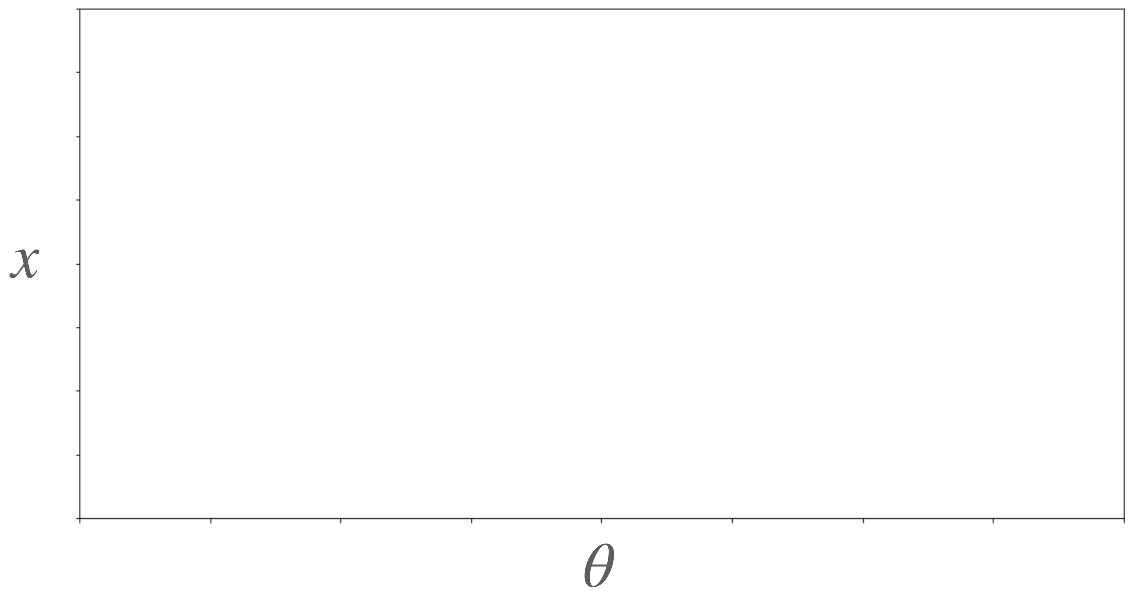


Figure from Lai et al, 2021

Simulators %>% Likelihood-free Inference

Simulator

- A computer program defined as $x \sim p(x \mid \theta)$ that has
  - input parameters θ
  - stochastic output x



Simulator

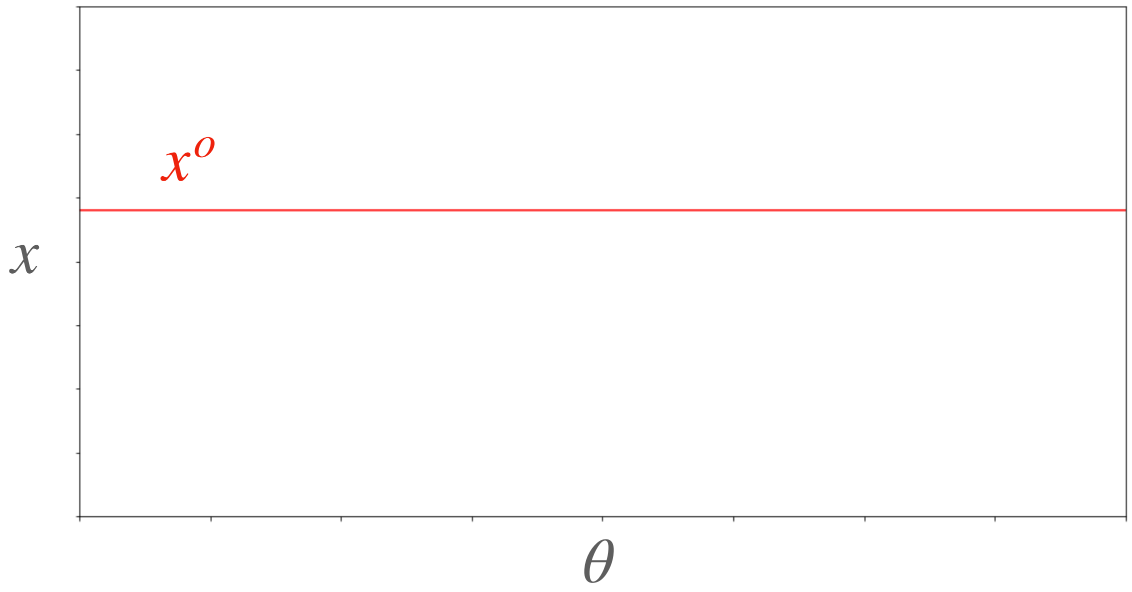
- The data produced by the simulator can be basically in any format
- It can be e.g.
  - Single time series
  - Set of independent data
  - Images
  - Distribution of data points

Inference

- Observe data and infer the values of the parameters that generated them
  - Often based on likelihood $p(x^o \mid \theta)$
  - Bayesian approach $p(\theta \mid x^o) \propto p(x^o \mid \theta)p(\theta)$
  - Maximum likelihood $\arg\max_\theta p(x^o \mid \theta)$
- When data generating process (simulator) is defined as a set of rules to draw $x \sim p(x \mid \theta)$ it is often infeasible to formulate the analytical likelihood $p(x^o \mid \theta)$

Inference without likelihood

- Use the capability to draw simulated data conditioned on the input parameters
- Likely true parameter values are thought to produce data that are *similar* to the observed data



Inference without likelihood

- Use the capability to draw simulated data conditioned on the input parameters
- Likely true parameter values are thought to produce data that are *similar enough* to the observed data
- Approximate the posterior distribution as

$$p(\theta \mid x^o) \propto p(x^o \mid \theta)p(\theta) = \int \mathbb{I}_{\Omega(x^o)}(x)p(x \mid \theta)\mathrm{d}x p(\theta)$$

- Region $\Omega(x^o) = \{x : d(x, x^o) \leq \epsilon\}$ contains data that are similar enough to the realised observation

Distance metric

- Acceptation region is defined by the distance metric
- Reasonable distance metrics depend on the data format, e.g.
  - Euclidean distance for low dimensional numerical outputs
  - L1 distance for data containing outliers
  - Quantiles or Wasserstein distance for distributions

Data dimensionality

- Dimensionality of the data often causes problem when running LFI methods
  - In high dimensions it becomes increasing improbable to generate data close to the observed data
- The (current) standard approach is still to use summary statistics $S(\cdot)$
  - E.g $d(x, x^o) \approx d(S(x), S(x^o)) \quad (\approx \rho(x, x^o))$
- Sufficient statistics usually do not exist
- How to choose them?

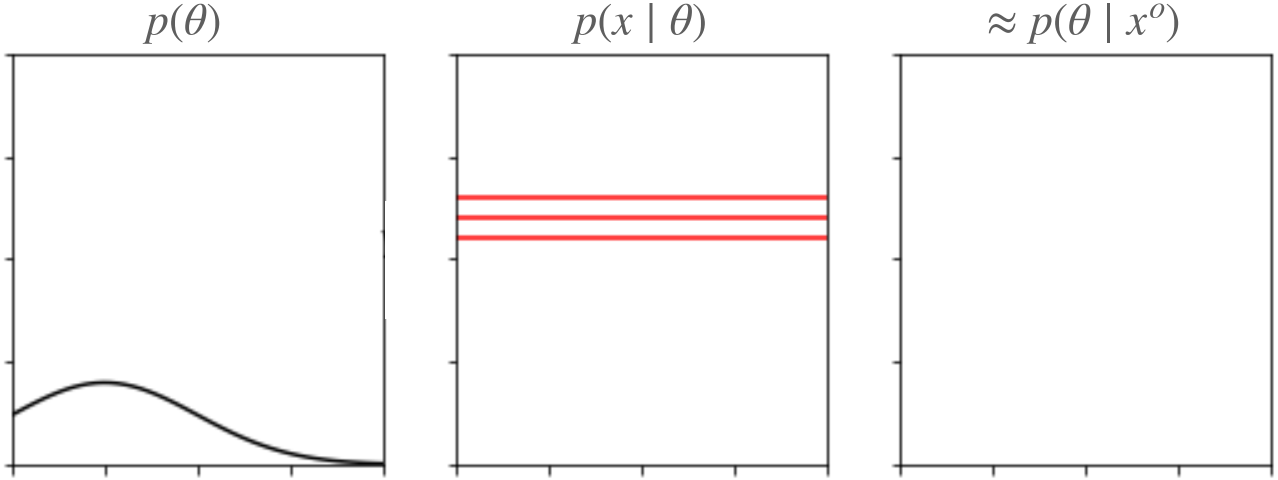
Selecting summary statistics

- An open problem
- Often we use bespoke summary statistics
  - Use **domain expertise** if available
  - **Explore** the simulator prior to inference
  - **Diagnose** the inference results
- Automatic algorithms for **selecting/constructing** the summary statistics

Summary statistics %>%
Distance function %>%
Threshold

Rejection Approximate Bayesian Computation
(ABC)

```
For $i = 1,\ldots,N$
    $\theta^* \sim p(\theta)$
    $x^* \sim p(x \mid \theta^*)$
    while $d(x^*, x^o) > \epsilon$ :
        $\theta^* \sim p(\theta)$
        $x^* \sim p(x \mid \theta^*)$
    set $\theta_i = \theta^*$
```



Rejection ABC

Alternative approach

- Instead of choosing a fixed threshold we can instead sample a very large artificial data set and choose a fraction of samples that are most similar to the observed data
- Threshold can then be calculated after the simulation as the largest distance that was still accepted
- The approach can be used in smaller scale to find out reasonable threshold levels for sampling based ABC methods

Rejection ABC uses samples from the prior

- Unless we have plenty of prior information about the parameters, sampling from prior is hardly effective
- Sequentially formulate importance sampling distributions with more probability mass in interesting regions
  - Sequential Monte Carlo ABC

SMC-ABC

- Round 1 : SMC-ABC is initialised as rejection ABC with a loose threshold ϵ_1
  - $\theta_i^{(1)} \sim p(\theta \mid d(S(x), S(x^o)) \leq \epsilon_1), \quad i = 1, \ldots, N$
  - Set weight $w_i^{(1)} = N^{-1}$
  - Calculate sample variances (for each dimension $j = 1, \ldots, M$)

$$\hat{\mu} = \sum_{i=1}^{N} \frac{\theta_i}{N}, \quad \hat{\sigma}_j^2 = \sum_{i=1}^{N} \frac{1}{N} (\theta_{i_j} - \hat{\mu}_{i_j})^2$$

  - Set proposal density $q(\theta^{(2)} \mid \theta^{(1)}) = \text{Normal}(\theta^{(1)}, 2 \cdot \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_M^2))$

SMC-ABC

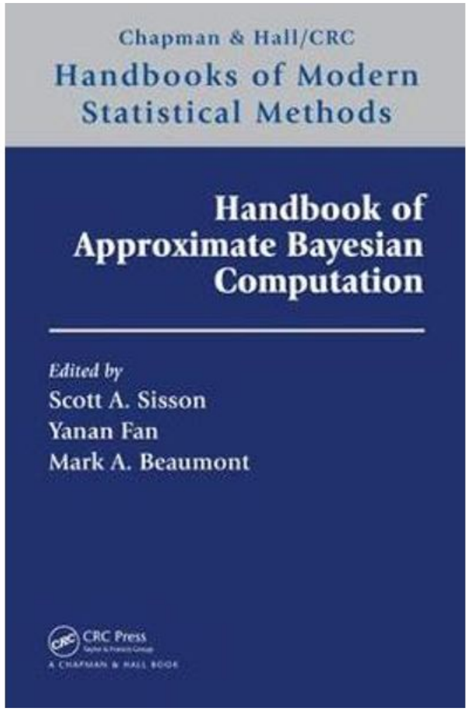
- Rounds $t = 2,\ldots,T$: Set a tighter threshold $\epsilon_t < \epsilon_{t-1}$
  - (1) Select $\theta_i^{(t-1)}$ with probability $\propto w_i^{(t-1)}$
  - (2) Draw $\theta* \sim \text{Normal}(\theta_i^{(t-1)}, \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_M^2))$
  - (3) Simulate $x* \sim p(x \mid \theta*)$
  - Repeat (1)-(3) until $d(S(x*), S(x^o)) < \epsilon_t$ and set $\theta_i^{(t)} = \theta*$
    - Set weights $w_i^{(t)} \propto p(\theta_i^{(t)}) \cdot \left[\sum_{k=1}^{N} w_k^{(t-1)} \sum_{j=1}^{M} \phi \left(\frac{\theta_{i_j}^{(t)} - \theta_{k_j}^{(t)}}{\hat{\sigma}_j^{(t-1)}} \right) \right]^{-1}$
    - Calculate weighted sample variances $\hat{\sigma}_j^2$ (for each dimension $j = 1,\ldots,M$)
    - Set proposal density $q(\theta^{(t+1)} \mid \theta^{(t)}) = \text{Normal}(\theta^{(t)},\ 2 \cdot \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_M^2))$

Summary statistics `%>%`
Distance function `%>%`
Threshold `%>%`
Things that make you go *hmmm?*

Issues to be aware of

- Summary statistics may not catch relevant features of the data
  - can decrease dimension too much - lose information
  - don’t decrease dimension enough - didn’t solve the problem
  - can be correlated - redundant dimensions
  - etc
- How multiple summary statistics/elements of a summary statistic contribute to the distance?
  - E.g. Multiple summary statistics can each have wildly different scales

Reference



Chapman & Hall/CRC
Handbooks of Modern
Statistical Methods

Handbook of
Approximate Bayesian
Computation

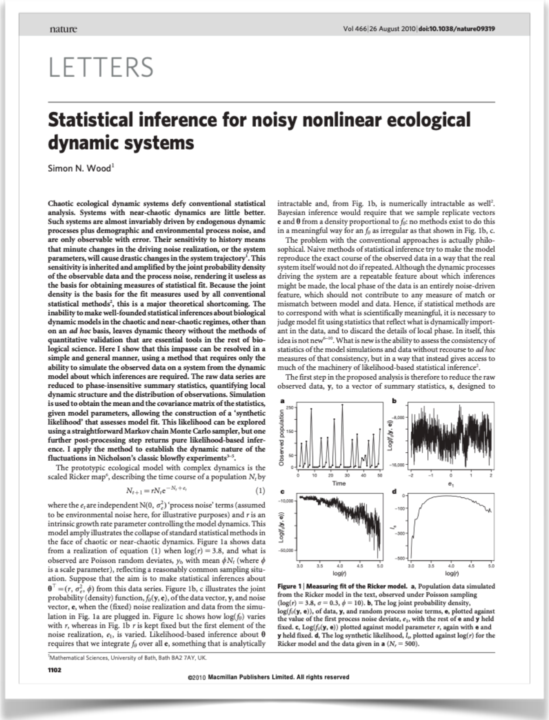
Edited by
Scott A. Sisson
Yanan Fan
Mark A. Beaumont

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

- For those who want to read more

Surrogate models

- Alternative approach to likelihood-free inference is to construct surrogate models for various parts of the system
  - Synthetic likelihood is one of the first surrogate methods
  - At θ approximate $p(x \mid \theta) \approx \text{Normal}(x \mid \hat{\mu}_N(\theta), \hat{\Sigma}_N(\theta))$ with an sample of size N drawn from the model



nature Vol 466|26 August 2010|doi:10.1038/nature09319

LETTERS

Statistical inference for noisy nonlinear ecological dynamic systems

Simon N. Wood[1]

Chaotic ecological dynamic systems defy conventional statistical analysis. Systems with near-chaotic dynamics are little better. Such systems are almost invariably driven by endogenous dynamic processes plus demographic and environmental process noise, and are only observable with error. Their sensitivity to history means that minute changes in the driving noise realization, or the system parameters, will cause drastic changes in the system trajectory[1]. This sensitivity is inherited and amplified by the joint probability density of the observable data and the process noise, rendering it useless as the basis for obtaining measures of statistical fit. Because the joint density is the basis for the fit measures used by all conventional statistical methods[2], this is a major theoretical shortcoming. The inability to make well-founded statistical inferences about biological dynamic models in the chaotic and near-chaotic regimes, other than on an *ad hoc* basis, leaves dynamic theory without the methods of quantitative validation that are essential tools in the rest of biological science. Here I show that this impasse can be resolved in a simple and general manner, using a method that requires only the ability to simulate the observed data on a system from the dynamic model about which inferences are required. The raw data series are reduced to phase-insensitive summary statistics, quantifying local dynamic structure and the distribution of observations. Simulation is used to obtain the mean and the covariance matrix of the statistics, given model parameters, allowing the construction of a 'synthetic likelihood' that assesses model fit. This likelihood can be explored using a straightforward Markov chain Monte Carlo sampler, but one further post-processing step returns pure likelihood-based inference. I apply the method to establish the dynamic nature of the fluctuations in Nicholson's classic blowfly experiments[3–5].

The prototypic ecological model with complex dynamics is the scaled Ricker map[6], describing the time course of a population N_t by

$$N_{t+1} = rN_t e^{-N_t + e_t} \qquad (1)$$

where the e_t are independent $\mathrm{N}(0, \sigma_e^2)$ 'process noise' terms (assumed to be environmental noise here, for illustrative purposes) and r is an intrinsic growth rate parameter controlling the model dynamics. This model amply illustrates the collapse of standard statistical methods in the face of chaotic or near-chaotic dynamics. Figure 1a shows data from a realization of equation (1) when $\log(r) = 3.8$, and what is observed are Poisson random deviates, y_t, with mean ϕN_t (where ϕ is a scale parameter), reflecting a reasonably common sampling situation. Suppose that the aim is to make statistical inferences about $\boldsymbol{\theta}^\top = (r, \sigma_e^2, \phi)$ from this data series. Figure 1b, c illustrates the joint probability (density) function, $f_\theta(\mathbf{y}, \mathbf{e})$, of the data vector, $\mathbf{y}$, and noise vector, $\mathbf{e}$, when the (fixed) noise realization and data from the simulation in Fig. 1a are plugged in. Figure 1c shows how $\log(f_\theta)$ varies with r, whereas in Fig. 1b r is kept fixed but the first element of the noise realization, e_1, is varied. Likelihood-based inference about $\boldsymbol{\theta}$ requires that we integrate f_θ over all $\mathbf{e}$, something that is analytically intractable and, from Fig. 1b, is numerically intractable as well[2]. Bayesian inference would require that we sample replicate vectors $\mathbf{e}$ and $\boldsymbol{\theta}$ from a density proportional to f_θ: no methods exist to do this in a meaningful way for an f_θ as irregular as that shown in Fig. 1b, c.

The problem with the conventional approaches is actually philosophical. Naive methods of statistical inference try to make the model reproduce the exact course of the observed data in a way that the real system itself would not do if repeated. Although the dynamic processes driving the system are a repeatable feature about which inferences might be made, the local phase of the data is an entirely noise-driven feature, which should not contribute to any measure of match or mismatch between model and data. Hence, if statistical methods are to correspond with what is scientifically meaningful, it is necessary to judge model fit using statistics that reflect what is dynamically important in the data, and to discard the details of local phase. In itself, this idea is not new[6–10]. What is new is the ability to assess the consistency of statistics of the model simulations and data without recourse to *ad hoc* measures of that consistency, but in a way that instead gives access to much of the machinery of likelihood-based statistical inference[2].

The first step in the proposed analysis is therefore to reduce the raw observed data, $\mathbf{y}$, to a vector of summary statistics, $\mathbf{s}$, designed to

Figure 1 | Measuring fit of the Ricker model. **a**, Population data simulated from the Ricker model in the text, observed under Poisson sampling ($\log(r) = 3.8$, $\sigma = 0.3$, $\phi = 10$). **b**, The log joint probability density, $\log(f_\theta(\mathbf{y}, \mathbf{e}))$, of data, $\mathbf{y}$, and random process noise terms, $\mathbf{e}$, plotted against the value of the first process noise deviate, e_1, with the rest of $\mathbf{e}$ and $\mathbf{y}$ held fixed. **c**, $\mathrm{Log}(f_\theta(\mathbf{y}, \mathbf{e}))$ plotted against model parameter r, again with $\mathbf{e}$ and $\mathbf{y}$ held fixed. **d**, The log synthetic likelihood, l_s, plotted against $\log(r)$ for the Ricker model and the data given in **a** ($N_r = 500$).

[1]Mathematical Sciences, University of Bath, Bath BA2 7AY, UK.

1102

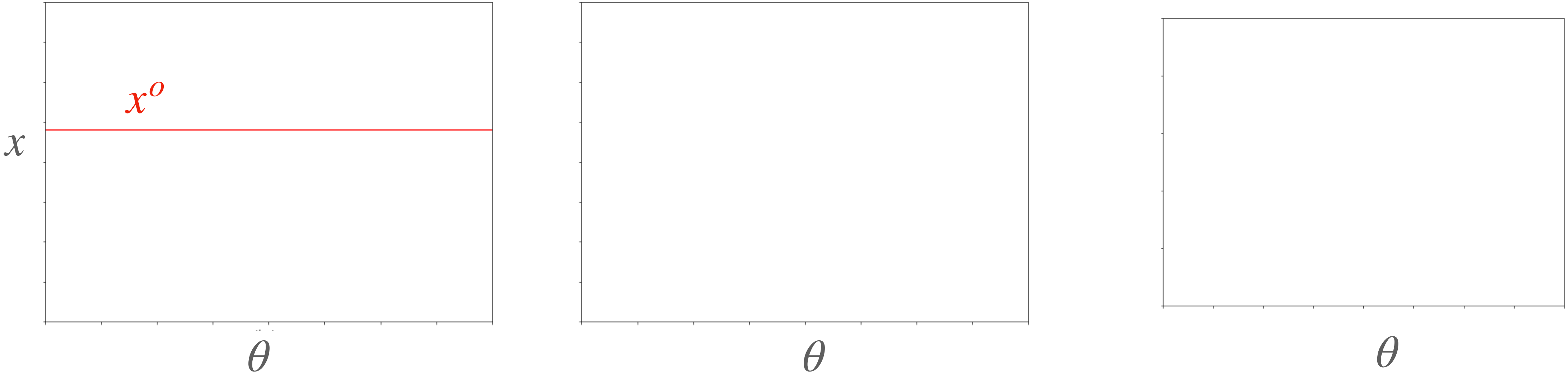
Example

- Construct approximate likelihood at an arbitrary parameter value

Gaussian empirical estimate

Set of samples at θ

$\approx p(\theta \mid x^o)$



Synthetic likelihood is hardly efficient

Surrogate is fitted at each parameter value separately

- BOLFI - Bayesian optimization for likelihood-free inference:
  - Model the discrepancy as a function of the parameter $\Delta(\theta) = d(x(\theta), x^o)$ conditioned on simulated data $\{(\theta_i, \Delta(\theta_i)))\}_{i=1}^t$using a Gaussian process

$$\Delta(\theta) \mid \{(\theta_i, \Delta(\theta_i))\}_{i=1}^t \sim \text{GP}(\mu_{1:t}(\theta), v_{1:t}(\theta) + \sigma_n^2)$$

Synthetic likelihood is hardly efficient

Surrogate is fitted at each parameter value separately

- BOLFI - Bayesian optimization for likelihood-free inference:
    - Likelihood can be approximated from the surrogate (pointwise) by

$$p(x^o \mid \theta) \approx \Phi\left(\frac{\epsilon - \mu_{1:t}(\theta)}{\sqrt{v_{1:t}(\theta) + \sigma_n^2}}\right)$$

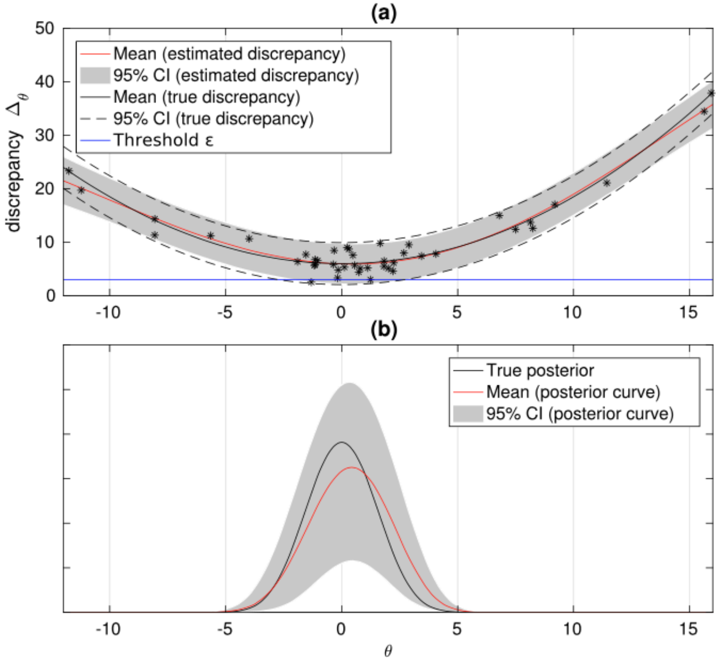
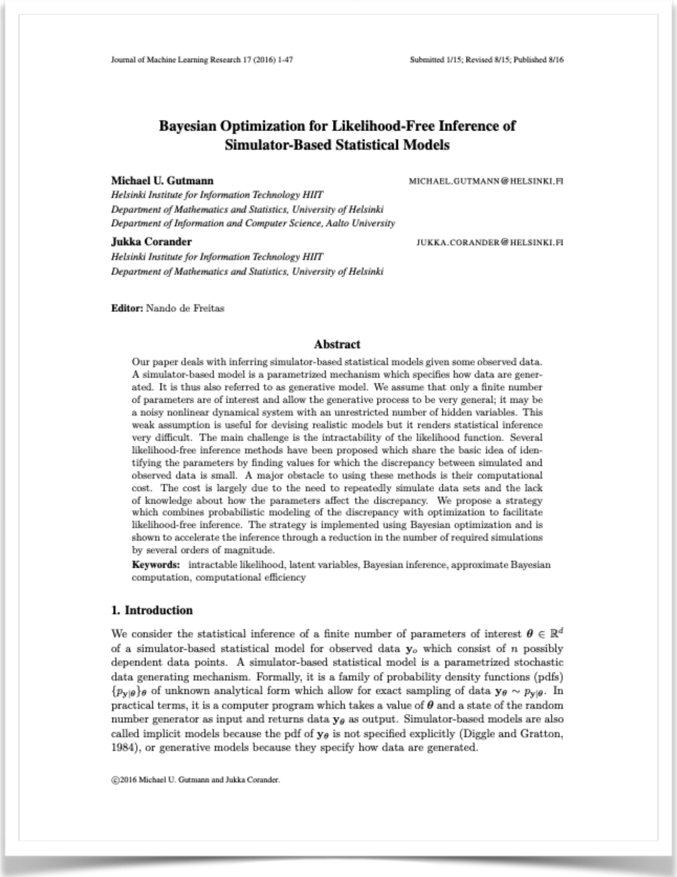


Figure from M. Järvenpää, “Efficient Acquisition Rules for Model-Based Approximate Bayesian Computation”, 2019

GP surrogates can utilise active learning

- Different strategies for selecting parameter values where to query the simulator
  - Reduce the number queries required to produce reasonable approximations to posterior/likelihood
- Usually based on optimization
  - Parameters that produce minimum discrepancies
  - Parameters that decrease most the uncertainty about the posterior

BOLFI



Journal of Machine Learning Research 17 (2016) 1-47 Submitted 1/15; Revised 8/15; Published 8/16

Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models

Michael U. Gutmann MICHAEL.GUTMANN@HELSINKI.FI
Helsinki Institute for Information Technology HIIT
Department of Mathematics and Statistics, University of Helsinki
Department of Information and Computer Science, Aalto University

Jukka Corander JUKKA.CORANDER@HELSINKI.FI
Helsinki Institute for Information Technology HIIT
Department of Mathematics and Statistics, University of Helsinki

Editor: Nando de Freitas

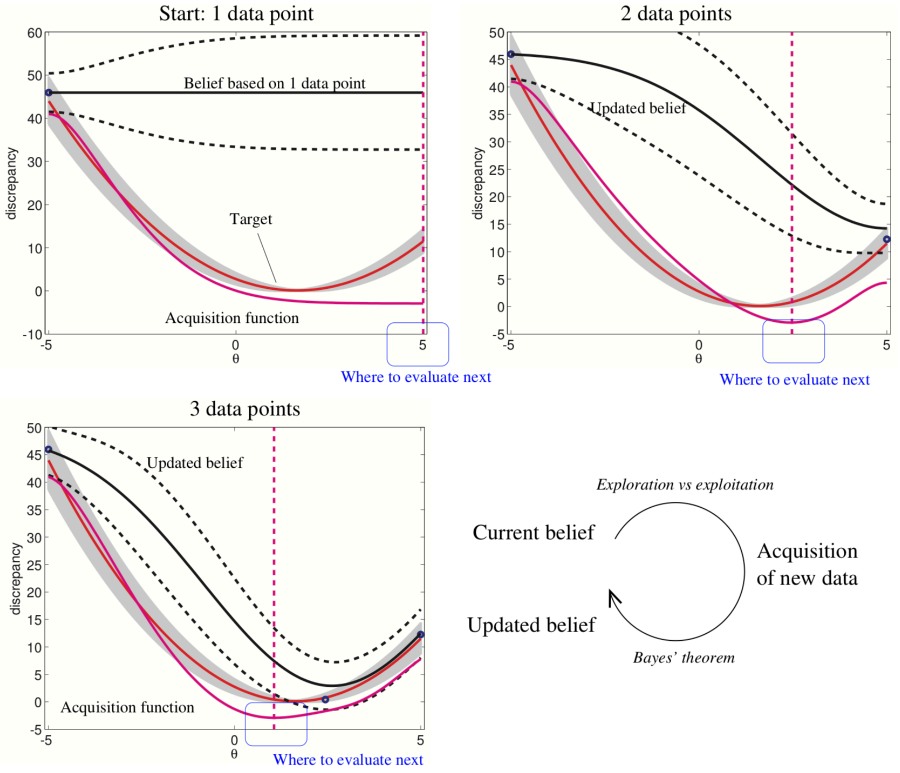
Abstract

Our paper deals with inferring simulator-based statistical models given some observed data. A simulator-based model is a parametrized mechanism which specifies how data are generated. It is thus also referred to as generative model. We assume that only a finite number of parameters are of interest and allow the generative process to be very general; it may be a noisy nonlinear dynamical system with an unrestricted number of hidden variables. This weak assumption is useful for devising realistic models but it renders statistical inference very difficult. The main challenge is the intractability of the likelihood function. Several likelihood-free inference methods have been proposed which share the basic idea of identifying the parameters by finding values for which the discrepancy between simulated and observed data is small. A major obstacle to using these methods is their computational cost. The cost is largely due to the need to repeatedly simulate data sets and the lack of knowledge about how the parameters affect the discrepancy. We propose a strategy which combines probabilistic modeling of the discrepancy with optimization to facilitate likelihood-free inference. The strategy is implemented using Bayesian optimization and is shown to accelerate the inference through a reduction in the number of required simulations by several orders of magnitude.

Keywords: intractable likelihood, latent variables, Bayesian inference, approximate Bayesian computation, computational efficiency

1. Introduction

We consider the statistical inference of a finite number of parameters of interest $\boldsymbol{\theta} \in \mathbb{R}^d$ of a simulator-based statistical model for observed data $\mathbf{y}_o$ which consist of n possibly dependent data points. A simulator-based statistical model is a parametrized stochastic data generating mechanism. Formally, it is a family of probability density functions (pdfs) $\{p_{\mathbf{y}|\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$ of unknown analytical form which allow for exact sampling of data $\mathbf{y}_{\boldsymbol{\theta}} \sim p_{\mathbf{y}|\boldsymbol{\theta}}$. In practical terms, it is a computer program which takes a value of $\boldsymbol{\theta}$ and a state of the random number generator as input and returns data $\mathbf{y}_{\boldsymbol{\theta}}$ as output. Simulator-based models are also called implicit models because the pdf of $\mathbf{y}_{\boldsymbol{\theta}}$ is not specified explicitly (Diggle and Gratton, 1984), or generative models because they specify how data are generated.



Exploration vs exploitation

Current belief

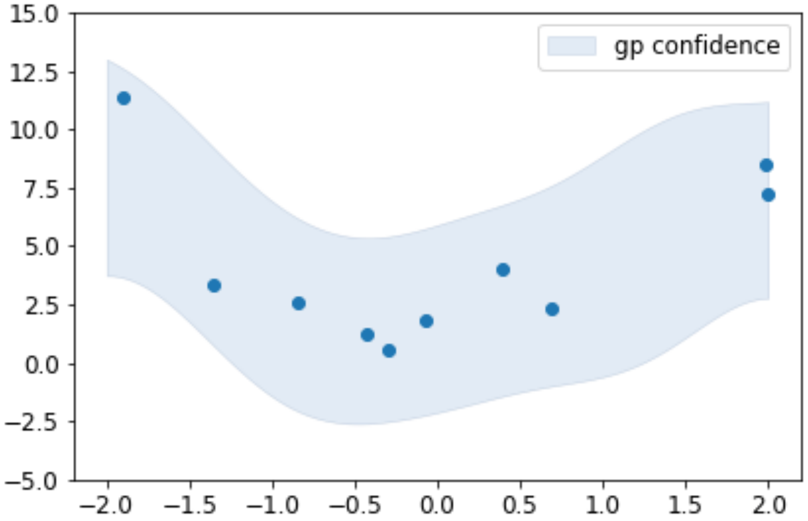
Acquisition of new data

Updated belief

Bayes' theorem

BOLFI

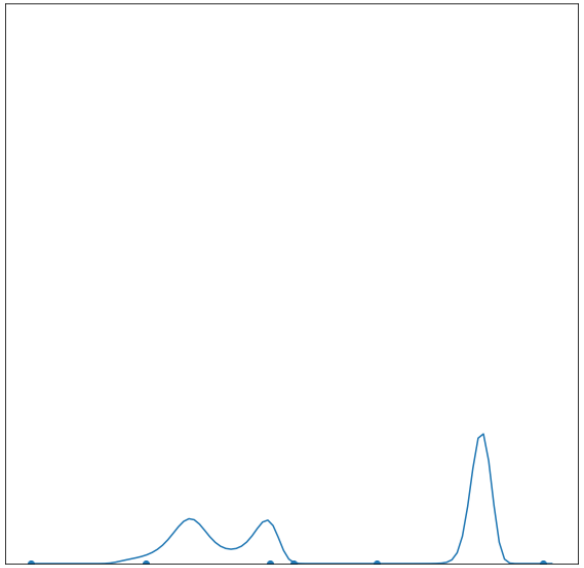
- Find parameter values that minimize discrepancy function
  - Black-box optimization
  - Acquisition strategies balance exploration and exploitation



BOLFI

- Find parameter values that minimize discrepancy function
  - Black-box optimization
  - Acquisition strategies balance exploration and exploitation

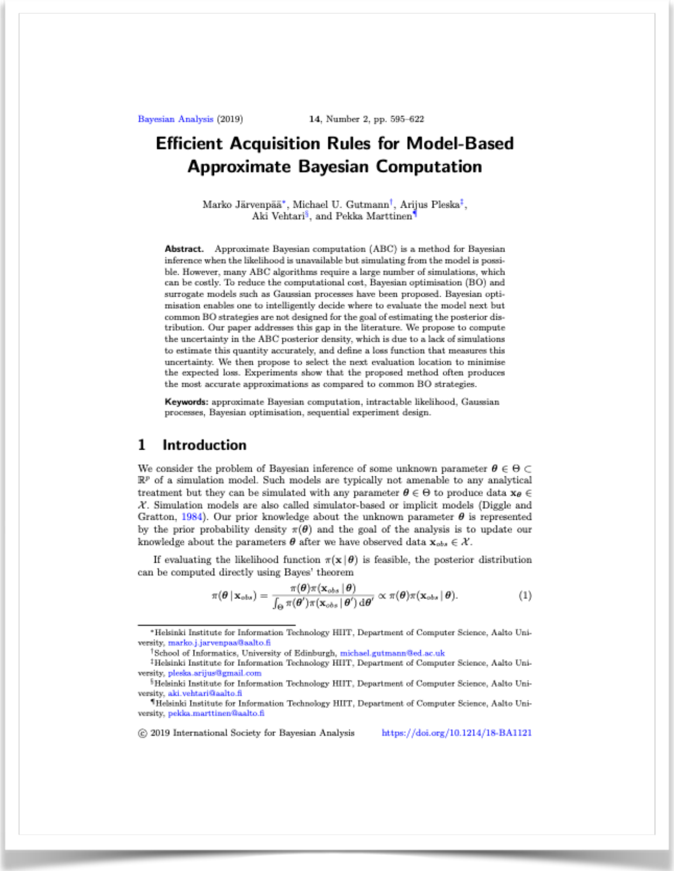
$$p(x^o \mid \theta) \approx \Phi\left(\frac{\epsilon - \mu_{1:t}(\theta)}{\sqrt{v_{1:t}(\theta) + \sigma_n^2}}\right)$$



Minimizing distance is often not optimal

Ultimate goal is to approximate posterior distribution

- How to choose query points that are most informative about the posterior distribution
- Active learning strategies can be based on more reasonable goals



Bayesian Analysis (2019) **14**, Number 2, pp. 595–622

Efficient Acquisition Rules for Model-Based Approximate Bayesian Computation

Marko Järvenpää[*], Michael U. Gutmann[†], Arijus Pleska[‡], Aki Vehtari[§], and Pekka Marttinen[¶]

Abstract. Approximate Bayesian computation (ABC) is a method for Bayesian inference when the likelihood is unavailable but simulating from the model is possible. However, many ABC algorithms require a large number of simulations, which can be costly. To reduce the computational cost, Bayesian optimisation (BO) and surrogate models such as Gaussian processes have been proposed. Bayesian optimisation enables one to intelligently decide where to evaluate the model next but common BO strategies are not designed for the goal of estimating the posterior distribution. Our paper addresses this gap in the literature. We propose to compute the uncertainty in the ABC posterior density, which is due to a lack of simulations to estimate this quantity accurately, and define a loss function that measures this uncertainty. We then propose to select the next evaluation location to minimise the expected loss. Experiments show that the proposed method often produces the most accurate approximations as compared to common BO strategies.

Keywords: approximate Bayesian computation, intractable likelihood, Gaussian processes, Bayesian optimisation, sequential experiment design.

1 Introduction

We consider the problem of Bayesian inference of some unknown parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ of a simulation model. Such models are typically not amenable to any analytical treatment but they can be simulated with any parameter $\boldsymbol{\theta} \in \Theta$ to produce data $\mathbf{x}_{\boldsymbol{\theta}} \in \mathcal{X}$. Simulation models are also called simulator-based or implicit models (Diggle and Gratton, 1984). Our prior knowledge about the unknown parameter $\boldsymbol{\theta}$ is represented by the prior probability density $\pi(\boldsymbol{\theta})$ and the goal of the analysis is to update our knowledge about the parameters $\boldsymbol{\theta}$ after we have observed data $\mathbf{x}_{\mathrm{obs}} \in \mathcal{X}$.

If evaluating the likelihood function $\pi(\mathbf{x} \,|\, \boldsymbol{\theta})$ is feasible, the posterior distribution can be computed directly using Bayes' theorem

$$\pi(\boldsymbol{\theta} \,|\, \mathbf{x}_{\mathrm{obs}}) = \frac{\pi(\boldsymbol{\theta})\pi(\mathbf{x}_{obs} \,|\, \boldsymbol{\theta})}{\int_{\Theta} \pi(\boldsymbol{\theta}')\pi(\mathbf{x}_{obs} \,|\, \boldsymbol{\theta}')\,\mathrm{d}\boldsymbol{\theta}'} \propto \pi(\boldsymbol{\theta})\pi(\mathbf{x}_{\mathrm{obs}} \,|\, \boldsymbol{\theta}). \qquad (1)$$

[*] Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, marko.j.jarvenpaa@aalto.fi

[†] School of Informatics, University of Edinburgh, michael.gutmann@ed.ac.uk

[‡] Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, pleska.arijus@gmail.com

[§] Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, aki.vehtari@aalto.fi

[¶] Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, pekka.marttinen@aalto.fi

https://doi.org/10.1214/18-BA1121

Acquisition functions

- Lower Confidence Bound Selection Criterion
- Maximum Variance
- Randomized Maximum Variance
- Expected Integrated Variance

LCBSC

Lower Confidence Bound Selection Criteria for minimizing the distance

- Selecting the query point for round t is a two part process. First optimise LCBSC

$$\theta * = \arg\min_{\theta} \mu_{1:t}(\theta) - \sqrt{\eta_t^2 v_{1:t}(\theta)} \quad , \quad \eta_t^2 = 2 \cdot \log\left(\frac{t^{2 \cdot d+2}\pi^2}{3 \cdot \xi_\eta}\right)$$

- Then sample the next query point from truncated Gaussian

$$\theta_{t+1} \sim \mathsf{TN}(\theta *, \Sigma_{\mathsf{acq}}, \Omega)$$

- where Σ_{acq} and Ω are a tunable parameter balancing exploration/explotation and the optimization region, respectively

MaxVar

The maximum variance acquisition method

- The next evaluation point is acquired where the variance of the unnormalised approximate posterior is maximised

$$\theta_{t+1} = \arg\max_{\theta} \mathsf{Var}(p(\theta) \cdot p_a(\theta))$$

$$p_a(\theta) = \Phi\left(\frac{\epsilon - \mu_{1:t}(\theta)}{\sqrt{v_{1:t}(\theta) + \sigma_n^2}}\right)$$

- ϵ is the ABC threshold, $\mu_{1:t}$ and $v_{1:t}$ are determined by the GP surrogate, σ_n^2 is the noise.

RandMaxVar

The randomized maximum variance acquisition method

- The next evaluation point is drawn randomly from the density corresponding to the variance of the posterior

$$\theta_{t+1} \sim q(\theta), \text{ where } q(\theta) \propto \text{Var}(p(\theta) \cdot p_a(\theta))$$

$$p_a(\theta) = \Phi\left(\frac{\epsilon - \mu_{1:t}(\theta)}{\sqrt{v_{1:t}(\theta) + \sigma_n^2}}\right)$$

- ϵ is the ABC threshold, $\mu_{1:t}$ and $v_{1:t}$ are determined by the GP surrogate, σ_n^2 is the noise.

ExpIntVar

The Expected Integrated Variance

- Loss function measures the overall uncertainty in the unnormalised ABC posterior over the parameter space.
- The value of the loss function depends on the next simulation so the next evaluation location $\theta*$ is chosen to minimise the expected loss

$$\theta_{t+1} = \arg\min_{\theta*} L_{1:t}(\theta*)$$

- The expected loss $L(\cdot)$ approximated as:

$$L_{1:t}(\theta*) \approx 2 \cdot \sum_{i=1}^{s} \omega^i \cdot p^2(\theta^i) \cdot w_{1:t+1}(\theta^i, \theta*)$$

- ω^i is an importance weight, $p^2(\theta^i)$ is the prior squared, and $w_{1:t+1}(\theta^i, \theta*)$ is the expected variance of the unnormalised ABC posterior at θ^i after running the simulation model with parameter $\theta*$

Sampling from surrogate

- To represent the posterior distribution we require a set of samples drawn from it
- ABC methods produce an approximate sample from the posterior
- Surrogate methods provide an approximate posterior curve
- We use MCMC methods to draw a posterior sample

After inference

After inference

How reliable are the results

- Interpret inference results
    - Different error sources
        - Algorithm performance
        - Model performance
        - Simulator performance

- Likelihood-free inference methods are based on several levels of approximations
    - All add to the total error
        - Which parts contribute most?
            - Run algorithm longer/sample more?
            - Change models?