Bayesian Optimization for Likelihood-Free Inference

Michael Gutmann

https://sites.google.com/site/michaelgutmann

University of Edinburgh

14th September 2016

Reference

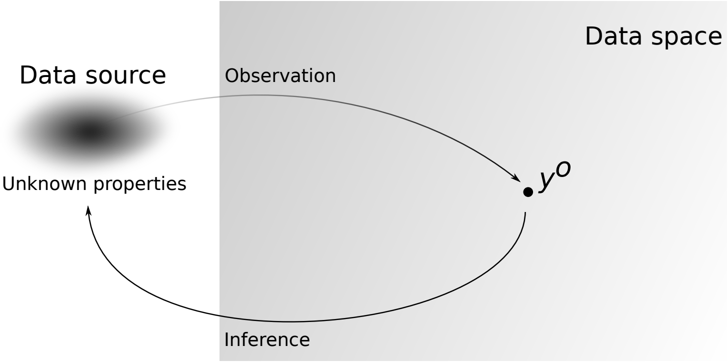
For further information:

M.U. Gutmann and J. Corander
Bayesian optimization for likelihood-free inference of simulator-based statistical models
Journal of Machine Learning Research, 17(125): 1–47, 2016

J. Lintusaari, M.U. Gutmann, R. Dutta, S. Kaski, and J. Corander
Fundamentals and Recent Developments in Approximate Bayesian Computation
Systematic Biology, in press, 2016

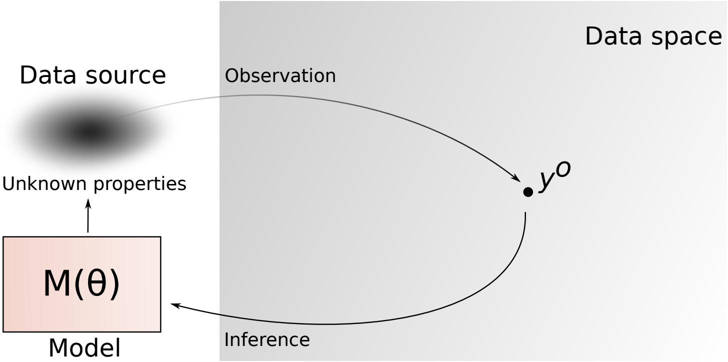
Overall goal

- Inference: Given data y^o, learn about properties of its source
- Enables decision making, predictions, ...



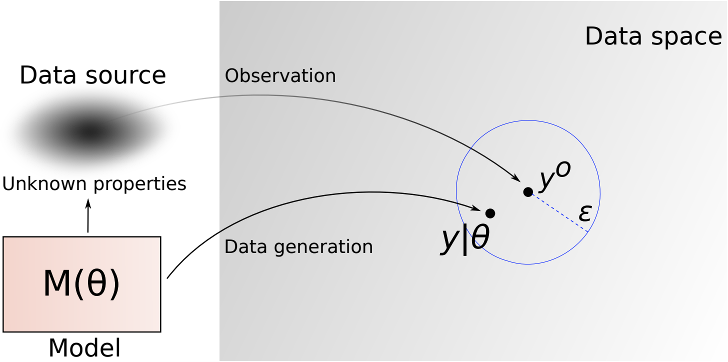
Approach

- Set up a model with potential properties θ (hypotheses)
- See which θ are in line with the observed data y^o



The likelihood function $L(\theta)$

- Measures agreement between θ and the observed data y^o
- Probability to generate data like y^o if hypothesis θ holds



Performing statistical inference

- If $L(\theta)$ is known, inference is straightforward
- Maximum likelihood estimation

$$\hat{\theta} = \mathrm{argmax}_\theta \, L(\theta)$$

- Bayesian inference

$$p(\theta|y) \propto p(\theta) \times L(\theta)$$
$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Allows us to learn from data by updating probabilities

Likelihood-free inference

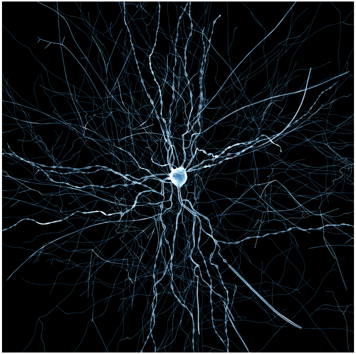
Statistical inference for models where

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

Importance of likelihood-free inference

One reason: Such generative / simulator-based models occur widely

- Astrophysics:
  Simulating the formation of galaxies, stars, or planets
- Evolutionary biology:
  Simulating the evolution of life
- Neuroscience:
  Simulating neural circuits
- Computer vision:
  Simulating natural scenes
- Health science:
  Simulating the spread of an infectious disease
- ...



Simulated neural activity in rat somatosensory cortex
(Figure from `https://bbp.epfl.ch/nmc-portal`)

Flavors of likelihood-free inference

- There are several flavors of likelihood-free inference. In Bayesian setting e.g.
  - Approximate Bayesian computation (ABC)
  - Synthetic likelihood (Wood, 2010)
- General idea: Identify the values of the parameters of interest θ for which simulated data resemble the observed data
- Simulated data resemble the observed data if some distance measure $d \geq 0$ is small.

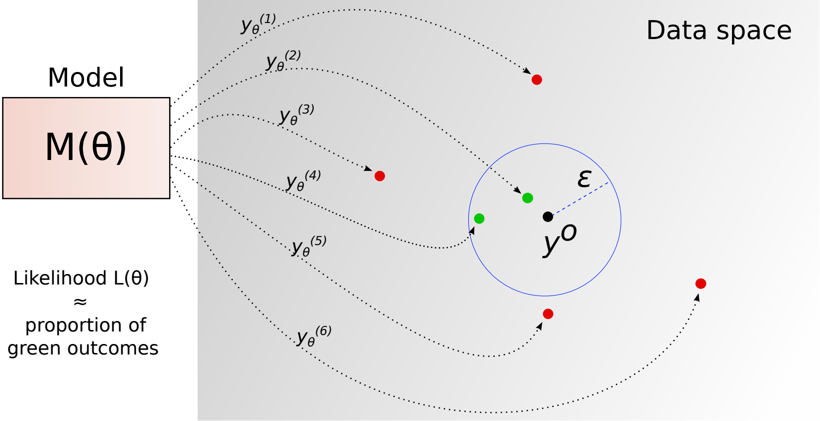
Here: Focus on ABC, see JMLR paper for synthetic likelihood

Meta ABC algorithm

- Let y^o be the observed data.
- Iterate many times:
  1. Sample θ from a proposal distribution $q(\theta)$
  2. Sample $y|\theta$ according to the model
  3. Compute distance $d(y, y^o)$ between simulated and observed data
  4. Retain θ if $d(y, y^o) \leq \epsilon$
- Different choices for $q(\theta)$ give different algorithms
- Produces samples from the (approximate) posterior when ϵ is small

Implicit likelihood approximation

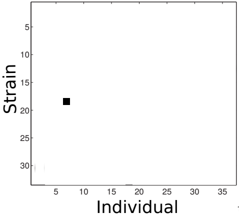
Likelihood: Probability to generate data like y^o if hypothesis θ holds



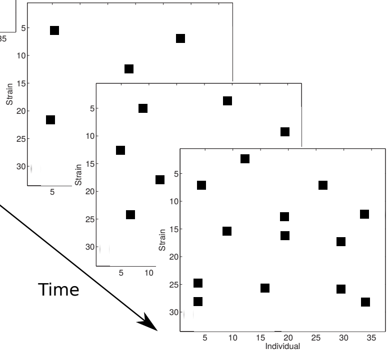
$$L(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left(d(y_\theta^{(i)}, y^o) \leq \epsilon\right)$$

Example: Bacterial infections in child care centers

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



Parameters of interest:
- rate of infections within a center
- rate of infections from outside
- competition between the strains



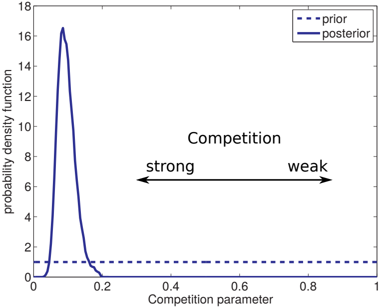
(Numminen et al, 2013)

Example: Bacterial infections in child care centers

- Data: *Streptococcus pneumoniae* colonization for 29 centers
- Inference with Population Monte Carlo ABC
- Reveals strong competition between different bacterial strains

Expensive:

- 4.5 days on a cluster with 200 cores
- More than one million simulated data sets

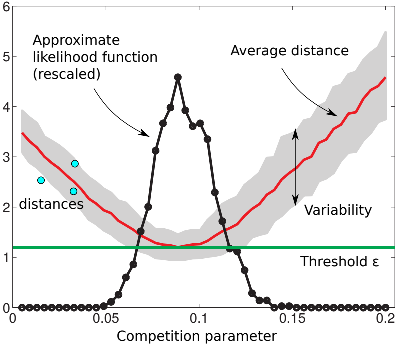


Why is the ABC algorithm so expensive?

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

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

Approximate lik function for competition parameter. $N = 300$.



Proposed solution

(Gutmann and Corander, 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

Modeling (points 1 & 2)

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

$$\hat{L}(\theta) \propto \widehat{\Pr}\left(d \le \epsilon \mid \theta\right)$$

  $\widehat{\Pr}$ is probability under the estimated model.
- Here: Use (log) Gaussian process as model (with squared exponential covariance function)
- Approach not restricted to Gaussian processes.

Data acquisition (points 3 & 4)

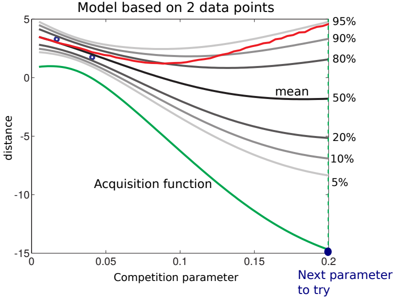
- Samples of θ could be obtained by sampling from the prior or some adaptively constructed proposal distribution
- Give priority to regions in the parameter space where distance d tends to be small.
- Use Bayesian optimization to find such regions
- Here: Use lower confidence bound acquisition function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

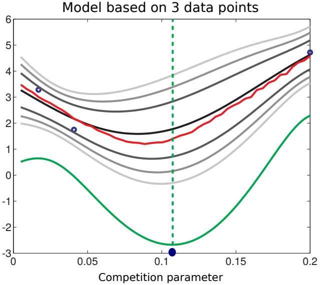
$$\mathcal{A}_t(\theta) = \underbrace{\mu_t(\theta)}_{\text{post mean}} - \sqrt{\underbrace{\eta_t^2}_{\text{weight}} \underbrace{v_t(\theta)}_{\text{post var}}} \tag{1}$$

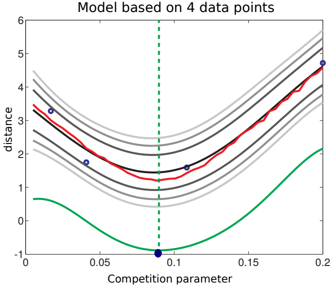
t: number of samples acquired so far

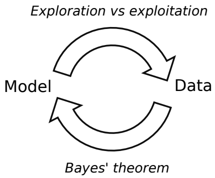
- Approach not restricted to this acquisition function.

Bayesian optimization for likelihood-free inference









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.

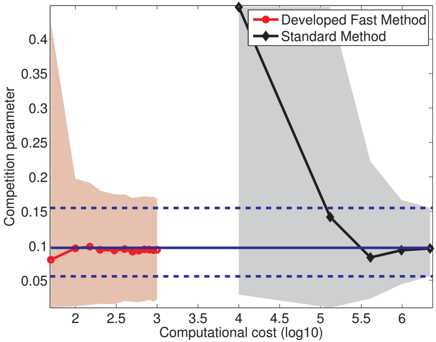
4.5 days with 200 cores

↓

90 minutes with seven cores

Posterior means: solid lines,

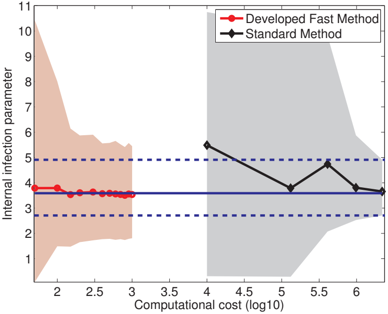
credibility intervals: shaded areas or dashed lines.

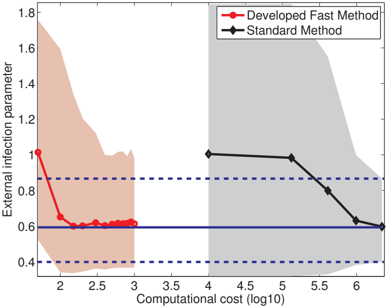


(Gutmann and Corander, 2016)

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.

Further 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 JMLR paper for a discussion

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

References

- M.U. Gutmann and J. Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models, *Journal of Machine Learning Research*, 17(125): 1–47, 2016
- J. Lintusaari, M.U. Gutmann, R. Dutta, S. Kaski, and J. Corander. Fundamentals and Recent Developments in Approximate Bayesian Computation, *Systematic Biology*, in press, 2016
- E. Numminen, M.U. Gutmann, M. Shubin, et al. The impact of host metapopulation structure on the population genetics of colonizing bacteria *Journal of Theoretical Biology*, 396: 53–62, 2016
- J. Lintusaari, M.U. Gutmann, S. Kaski, and J. Corander. On the identifiability of transmission dynamic models for infectious disease *Genetics*, 202(3): 911–918, 2016
- P. Marttinen, N.J. Croucher, M.U. Gutmann, J. Corander, and W.P. Hanage. Recombination produces coherent bacterial species clusters in both core and accessory genomes, *Microbial Genomics*, 1(5), 2015
- Numminen et al. Estimating the Transmission Dynamics of Streptococcus pneumoniae from Strain Prevalence Data. *Biometrics* 9, 2013.
- N. Srinivas, A. Krause, S.M. Kakade, and M. Seeger. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.
- S.N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems, *Nature*, 466: 1102–1104, 2010
- D. Cox and S. John. A statistical method for global optimization, *Proc. IEEE Conference on Systems, Man and Cybernetics*, 2: 1241–1246, 1992