Fast Likelihood-Free Inference via Bayesian Optimization

Michael Gutmann

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

University of Helsinki   Aalto University

Helsinki Institute for Information Technology

Joint work with Jukka Corander

17 May 2016

References

For all the details:
M.U. Gutmann and J. Corander
Bayesian optimization for likelihood-free inference of simulator-based statistical models
Journal of Machine Learning Research, in press.
http://arxiv.org/abs/1501.03291

Early results:
Bayesian Optimization for Likelihood-Free Estimation
Poster at ABC in Rome, 2013.

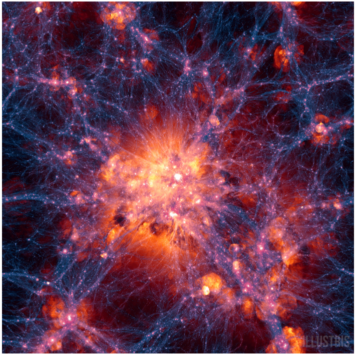
Likelihood-free inference

Statistical inference for models where

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

Why does it matter?

- Such models occur widely:
  - Astrophysics:
    Simulating the formation of galaxies, stars, or planets
  - Evolutionary biology:
    Simulating the evolution of life
  - Health science:
    Simulating the spread of an infectious disease
  - ...
- Enables inference for models with complex data generating mechanisms (e.g. scientific models)



Dark matter density simulated by the Illustris collaboration
(Figure from http://www.illustris-project.org)

Likelihood-free inference is an umbrella term

- There are several flavors of likelihood-free inference. In Bayesian setting e.g.
  - Approximate Bayesian computation (ABC)
    (for review, see e.g. Marin et al, Statistics and Computing, 2012)
  - Synthetic likelihood (Wood, Nature, 2010)
- General idea: Identify the values of the parameters of interest $\boldsymbol{\theta}$ for which simulated data resemble the observed data
- Simulated data resemble the observed data if some discrepancy measure $\Delta \geq 0$ is small.

Here: Focus on ABC, see reference paper for more

Meta ABC algorithm

- Let $\mathbf{y}^o$ be the observed data.
- Iterate many many times:
  1. Sample $\boldsymbol{\theta}$ from a proposal distribution $q(\boldsymbol{\theta})$
  2. Sample $\mathbf{y}|\boldsymbol{\theta}$ according to the model
  3. Compute discrepancy Δ between $\mathbf{y}^o$ and $\mathbf{y}$
  4. Retain $\boldsymbol{\theta}$ if $\Delta \leq \epsilon$

Meta ABC algorithm

- Let $\mathbf{y}^o$ be the observed data.
- Iterate many many times:
  1. Sample $\boldsymbol{\theta}$ from a proposal distribution $q(\boldsymbol{\theta})$
  2. Sample $\mathbf{y}|\boldsymbol{\theta}$ according to the model
  3. Compute discrepancy Δ between $\mathbf{y}^o$ and $\mathbf{y}$
  4. Retain $\boldsymbol{\theta}$ if $\Delta \leq \epsilon$
- Different choices for $q(\boldsymbol{\theta})$ give different algorithms
  - rejection ABC (Tavaré et al, 1997; Pritchard et al, 1999)
  - MCMC ABC (Marjoram et al, 2003)
  - Population Monte Carlo ABC (Sisson et al, 2007)
- ϵ: trade-off between statistical and computational performance
- Produces samples from an approximate posterior

Two major difficulties

1. How to measure the discrepancy
2. How to handle the computational cost

Two major difficulties

1. How to measure the discrepancy

   → Use classification

   M.U. Gutmann, R. Dutta, S. Kaski, and J. Corander
   Statistical Inference of Intractable Generative Models via Classification
   `http://arxiv.org/abs/1407.4981`

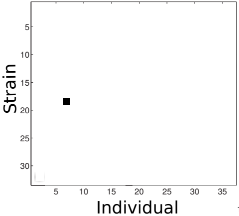
2. How to handle the computational cost

   → Use Bayesian optimization

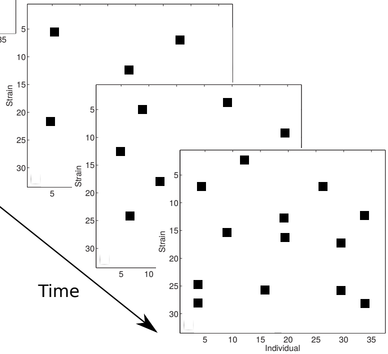
   M.U. Gutmann and J. Corander
   Bayesian optimization for likelihood-free inference of simulator-based statistical models
   Journal of Machine Learning Research, in press.
   `http://arxiv.org/abs/1501.03291`

Example: Bacterial infections in child care centers

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



Parameters of interest:
- β: rate of infections within a DCC
- Λ: rate of infections from outside
- θ: competition between the strains



(Numminen et al, 2013)

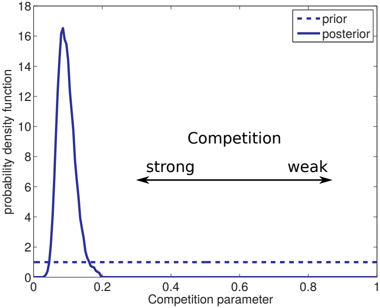
Example: Bacterial infections in child care centers

(Numminen et al, 2013)

- 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 ABC so expensive?

- Let $\mathbf{y}^o$ be the observed data.
- Building block of several ABC algorithms:
  1. Sample $\boldsymbol{\theta}$ from a proposal distribution $q(\boldsymbol{\theta})$
  2. Sample $\mathbf{y}|\boldsymbol{\theta}$ according to the model
  3. Compute discrepancy Δ between $\mathbf{y}^o$ and $\mathbf{y}$
  4. Retain $\boldsymbol{\theta}$ if $\Delta \leq \epsilon$
- Previous work: focus on choice of proposal distribution
- Key bottleneck: presence of the rejection step

  small $\epsilon \Rightarrow$ small acceptance probability $\Pr(\Delta \leq \epsilon \mid \boldsymbol{\theta})$

How to make the rejection step disappear ?

- Conditional acceptance probability corresponds to a likelihood approximation,
$$\tilde{L}(\boldsymbol{\theta}) \propto \Pr\left(\Delta \leq \epsilon \mid \boldsymbol{\theta}\right)$$
- The conditional distribution of Δ determines $\tilde{L}(\boldsymbol{\theta})$.
- If we knew the distribution of Δ we could compute $\tilde{L}(\boldsymbol{\theta})$.
- Suggests an approach based on statistical modeling of Δ.

Proposed approach

1. Model and estimate the distribution of Δ
   - Estimated model yields computable approximation $\hat{L}(\boldsymbol{\theta})$

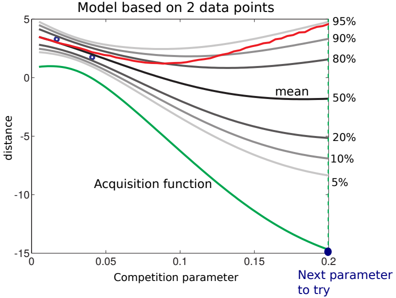
$$\hat{L}(\boldsymbol{\theta}) \propto \widehat{\Pr}\left(\Delta \leq \epsilon \mid \boldsymbol{\theta}\right)$$

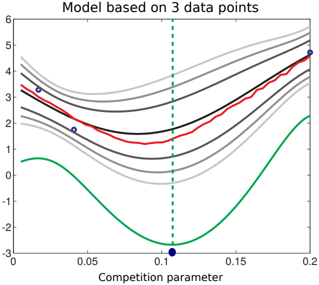
     $\widehat{\Pr}$ is probability under the estimated model.
   - Data for estimation by sampling $\boldsymbol{\theta}$ from the prior or from some other proposal distribution
2. Give priority to regions in the parameter space where discrepancy Δ tends to be small.
   - Prioritize modal regions of the likelihood/posterior
   - Use Bayesian optimization to find the regions where Δ tends to be small.

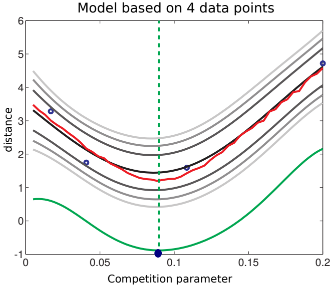
Bayesian optimization

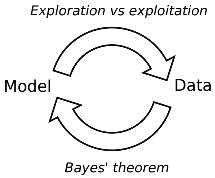
- Set of methods to minimize black-box functions
- Basic idea:
  - A probabilistic model of Δ guides the selection of points $\boldsymbol{\theta}$ where Δ is next evaluated.
  - Observed values of Δ are used to update the model by Bayes' theorem.
- When deciding where to evaluate Δ, balance
  - points where Δ is believed to be small ("exploitation")
  - points where we are uncertain about Δ ("exploration")

Bayesian optimization









Vanilla implementation

- Assume (log) discrepancy follows a Gaussian process model.
- Assume a squared exponential covariance function $\mathrm{cov}(\Delta_{\boldsymbol{\theta}}, \Delta_{\boldsymbol{\theta}'}) = k(\boldsymbol{\theta}, \boldsymbol{\theta}')$,

$$k(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sigma_f^2 \exp\left(\sum_j \frac{1}{\lambda_j^2} (\theta_j - \theta_j')^2 \right). \qquad (1)$$

- Use lower confidence bound acquisition function (e.g. Cox and John, 1992; Srinivas et al, 2012)

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

- Possibly use stochastic acquisition rule: sample from Gaussian centered at $\mathrm{argmin}_{\boldsymbol{\theta}} \mathcal{A}_t(\boldsymbol{\theta})$ while respecting boundaries.

Recipe for fast likelihood-free inference

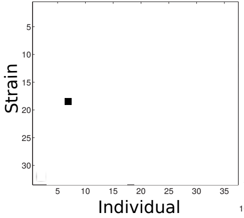
1. Estimate a model of the discrepancy using Bayesian optimization
2. Choose threshold ϵ to obtain the likelihood approximation

$$\hat{L}(\boldsymbol{\theta}) \propto \widehat{\Pr}\left(\Delta \leq \epsilon \mid \boldsymbol{\theta}\right)$$

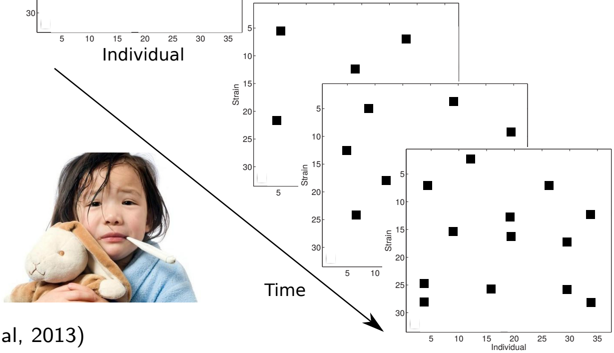
3. MLE or posterior inference with any standard method, using $\hat{L}$ in place of true likelihood function.

Example: Bacterial infections in child care centers

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



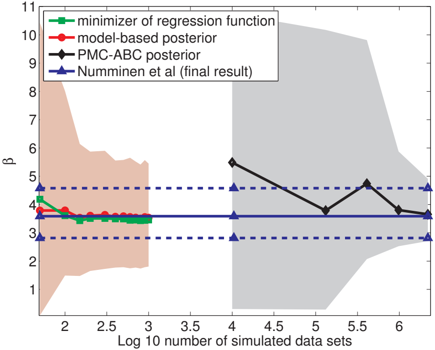
Parameters of interest:
- β: rate of infections within a DCC
- Λ: rate of infections from outside
- θ: competition between the strains



(Numminen et al, 2013)

Inference results

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



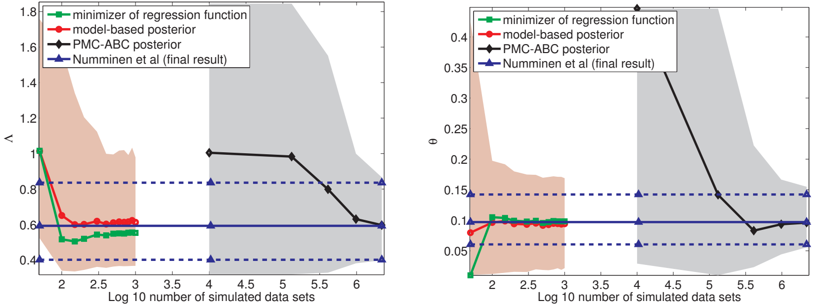
- The minimizer of the regression function under the model does not involve choosing a threshold ϵ.

Posterior means: solid lines with markers,

credibility intervals: shaded areas or dashed lines.

Inference results

- Comparison of the model-based approach with a population Monte Carlo (PMC) ABC approach.



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

Further benefits

- 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)
- Allowed us to perform far more comprehensive data analysis than with standard approach (Numminen et al, 2016)
- Estimated $\hat{L}(\boldsymbol{\theta})$ can be used to assess parameter identifiability for complex models
  - model about transmission dynamics of tuberculosis (Lintusaari et al, 2016)
- For point estimation, minimize $\hat{\mathrm{E}}(\Delta|\boldsymbol{\theta})$
  - no thresholds required

Some open questions

- Modeling of the discrepancy:
  Vanilla GP-model worked surprisingly well but there are likely more suitable models.
- Exploration/exploitation trade-off:
  Can we find strategies which are optimal for parameter inference?

Summary

- Problem considered: Computational cost of likelihood-free inference
- Proposed approach: Combine optimization with modeling of the discrepancy between simulated and observed data
- Outcome: Approach increases the efficiency of the inference by several orders of magnitude
- Talk was on approximate Bayesian computation with uniform kernels. For other kernels and synthetic likelihood see

M.U. Gutmann and J. Corander
Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models, *Journal of Machine Learning Research*, in press.
`http://arxiv.org/abs/1501.03291`

Appendix

Ricker model

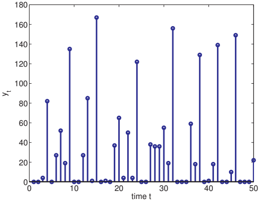
Details of the bacterial transmission model

Application to parameter inference in chaotic systems

- Data: Time series with counts y_t (animal population size)
- Simulator-based model: Stochastic version of the Ricker map followed by an observation model

$$\begin{aligned} \log N_t &= \log(r) + \log N_{t-1} - N_{t-1} + \sigma e_t, \quad e_t \sim \mathcal{N}(0,1) \\ y_t | N_t, \varphi &\sim \mathrm{Poisson}(\varphi N_t) \end{aligned}$$

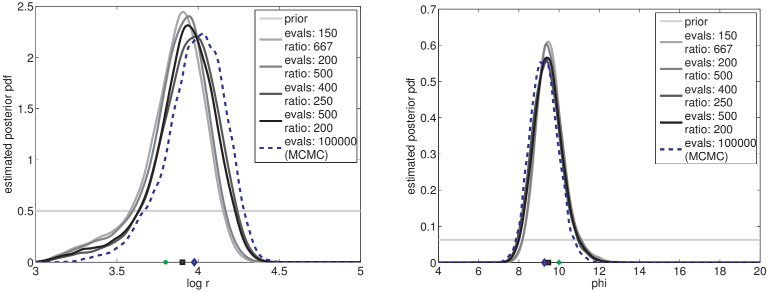
- Parameters $\boldsymbol{\theta}$:
  - $\log r$ (growth rate)
  - σ (noise var),
  - φ (scale parameter)



Example data, $\boldsymbol{\theta}^o = (3.8, 0.3, 10)$.

Application to parameter inference in chaotic systems

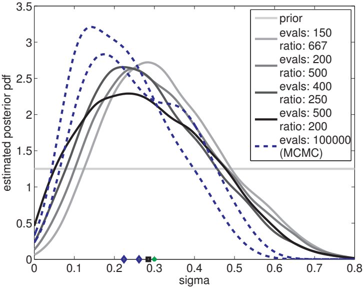
- Speed up: ≈ 600 times fewer evaluations of the distance function.
- Slight shift in posterior mean towards the data generating parameter $\boldsymbol{\theta}^o$ (green circle)



Comparison with results using MCMC (Wood, Nature, 2010)

Application to parameter inference in chaotic systems

- Speed up: $\approx$ 600 times fewer evaluations of the distance function.
- Slight shift in posterior mean towards the data generating parameter $\boldsymbol{\theta}^o$ (green circle)



Comparison with results using MCMC (Wood, Nature, 2010)

Bacterial transmission model (Numminen et al, 2013)

- Latent continuous time Markov chain for the transmissions inside a center

$$\Pr(I_{is}^{t+h} = 0 | I_{is}^{t} = 1) = h + o(h) \quad (3)$$

$$\Pr(I_{is}^{t+h} = 1 | I_{is'}^{t} = 0 \ \forall s') = R_s(t)h + o(h) \quad (4)$$

$$\Pr(I_{is}^{t+h} = 1 | I_{is}^{t} = 0, \ \exists s' : I_{is'}^{t} = 1) = \theta R_s(t)h + o(h) \quad (5)$$

$$R_s(t) = \beta E_s(t) + \Lambda P_s \quad (6)$$

- P_s : infections from outside the group (static)
- $E_s(t) = \sum_i \frac{1}{N-1} I_{is}^{t} \frac{1}{n_i(t)}$: infections from within the group
  $n_i(t) = \sum_{s'} I_{is'}^{t}$: number of strains that individual i carries
- Observation model: Cross-sectional sampling at random time.

Distance measure used

- Summary statistics for each center:
  - the diversity of the strains present
  - the number of different strains present
  - the proportion of infected individuals
  - the proportion of individuals with more than one strain.
- Distance $\equiv$ Distance between the empirical cumulative distribution functions (cdfs) of the four summary statistics.