Likelihood-free statistical inference

Michael Gutmann

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

Bayesian Statistics Group
University of Helsinki

22nd January 2015

Presentation based on:

- M.U. Gutmann, R. Dutta, S. Kaski, and J. Corander
  Likelihood-free inference via classification
  `http://arxiv.org/abs/1407.4981`
- M.U. Gutmann and J. Corander
  Bayesian optimization for likelihood-free inference of simulator-based statistical models
  `http://arxiv.org/abs/1501.03291`

Big picture of statistical inference

- Given: A statistical model which describes data $\mathbf{y} = (y_1, \dots, y_n)$, with model parameters $\boldsymbol{\theta}$
- Given: Observed data $\mathbf{y}^o$
- Possibly given: A (prior) probability density function (pdf) for $\boldsymbol{\theta}$, $p_{\boldsymbol{\theta}}$
- Wanted: Some probabilistic statement about $\boldsymbol{\theta}$
  - which value has generated $\mathbf{y}^o$ most likely?
  - what is the mean value of $\boldsymbol{\theta}$ given $\mathbf{y}^o$?
  - given $\mathbf{y}^o$, which interval contains θ_1 with probability 0.95 ?
  - ...

Three types of statistical models

1. Statistical model as family of pdfs, e.g.

$$p_{\mathbf{y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right), \quad \boldsymbol{\theta} = (\mu, \sigma)$$

2. Unnormalized models
(scale of $p_{\mathbf{y}|\boldsymbol{\theta}}$, that is, the partition function, is not known)

$$p^0_{\mathbf{y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right)$$

3. Simulator-based models
(shape and scale of $p_{\mathbf{y}|\boldsymbol{\theta}}$ are not known but sampling is possible)

$$\mathbf{y} \sim p_{\mathbf{y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta}), \qquad y_i = \mu + \sigma n_i \quad n_i \sim \mathcal{N}(0, 1)$$

Likelihood-based statistical inference

- Likelihood function: pdf of the observed data $\mathbf{y}^o$ as a function of the model parameters

$$L(\boldsymbol{\theta}) \propto p_{\mathbf{y}|\boldsymbol{\theta}}(\mathbf{y}^o|\boldsymbol{\theta})$$

- Plays a central role in statistical inference
  - Maximum likelihood estimation:

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

  - Bayesian inference:

$$p_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta}|\mathbf{y}^o) \propto L(\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

- Allows to make probabilistic statements about $\boldsymbol{\theta}$.
- Generally not computable for unnormalized and simulator-based models.

Likelihood-free inference (LFI)

- LFI: Procedure to obtain probabilistic statements about $\boldsymbol{\theta}$ if likelihood is not available, as for unnormalized or simulator-based statistical models.
- Existing methods for unnormalized models:
  - pseudo-likelihood (Besag, JRSSB, 1974)
  - contrastive divergence (Hinton, NeCo, 2002)
  - score matching (Hyvärinen, JMLR, 2005)
  - noise-contrastive estimation (Gutmann and Hyvärinen, JMLR, 2012)
- Here: simulator-based models

Why simulator-based models?

- Allows to implement hypotheses of how the data were generated without having to make excessive compromises in the modeling.
- Neat interface with physical or biological models of data.
- Can handle (infinitely many) unobserved variables.

Principle behind the inference algorithms

- There are several flavors of likelihood-free inference for simulator-based models, e.g.
  - Approximate Bayesian computation (ABC) (for recent review: Marin, Statistics and Computing, 2012)
  - Synthetic likelihood (Wood, Nature, 2010)
- Basic idea: Identify values of $\boldsymbol{\theta}$ for which simulated data resemble the observed data (discrepancy $\Delta_{\boldsymbol{\theta}}$ between simulated and observed data is small).

Example

- Inference of the mean θ of a Gaussian of variance one.
- Discrepancy:

$$\Delta_\theta = (\hat{\mu}^o - \hat{\mu}_\theta)^2,$$

$$\hat{\mu}^o = \frac{1}{n}\sum_{i=1}^{n} y_i^o,$$

$$\hat{\mu}_\theta = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$y_i \sim \mathcal{N}(\theta, 1)$$

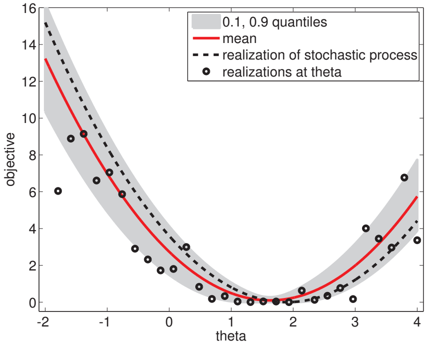


Figure 1 : Distribution of Δ_θ, the squared distance between the sample averages

Example

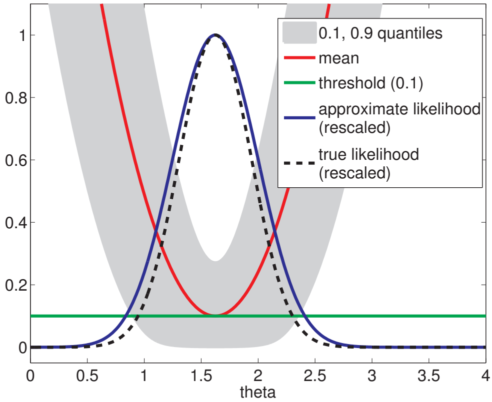


Figure 2 : Probability that Δ_θ is below some threshold h approximates the likelihood.

Example

- In this simple example, the probability can be computed in closed form, with $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) \mathrm{d}u$

$$\Pr(\Delta_\theta \leq h) = \Phi\left(\sqrt{n}(\hat{\mu}^o - \theta) + \sqrt{nh}\right) - \Phi\left(\sqrt{n}(\hat{\mu}^o - \theta) - \sqrt{nh}\right)$$

- For nh small: $\Pr(\Delta_\theta \leq h) \propto \sqrt{h} L(\theta)$
- For realistic models, sample average is used

$$\Pr(\Delta_\theta \leq h) \approx \frac{1}{N} \sum_{i=1}^{N} 1_{[0,h]}(\Delta_\theta^{(i)}) \underset{\sim}{\propto} L(\theta)$$

- Good news: For small enough h and large enough N, good approximation of likelihood.
- Bad news: Procedure is computationally costly

Two major difficulties in likelihood-free inference

1. How to measure the discrepancy between simulated and observed data
2. How to handle the computational burden of the inference

Two major difficulties in likelihood-free inference

1. How to measure the discrepancy between simulated and observed data
   → Use classification
   M.U. Gutmann, R. Dutta, S. Kaski, and J. Corander
   Likelihood-free inference via classification
   `http://arxiv.org/abs/1407.4981`
2. How to handle the computational burden of the inference
   → Use Bayesian optimization
   M.U. Gutmann and J. Corander
   Bayesian optimization for likelihood-free inference of simulator-based statistical models
   `http://arxiv.org/abs/1501.03291`

Discrepancy measurement via classification

- Correctly classifying data into two categories is usually easier if the two data sets were generated with very different values of $\boldsymbol{\theta}$ (left) than with similar values (right).

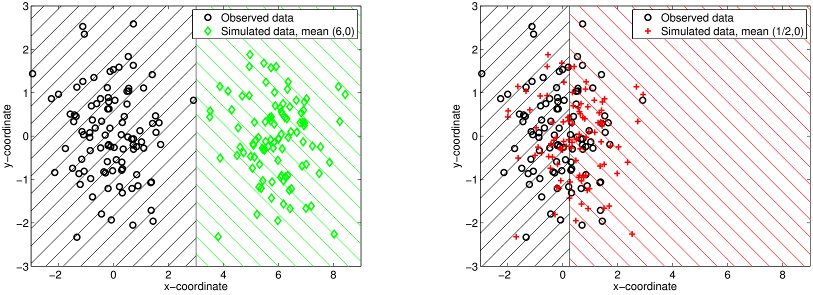


Figure 3 : Discriminability (classifiability) as discrepancy measure $\Delta_{\boldsymbol{\theta}}$.

Discrepancy measurement via classification

- We proposed to use the discriminability of the observed and simulated data as discrepancy measure.
- Complete arsenal of classification methods becomes available to likelihood-free inference.

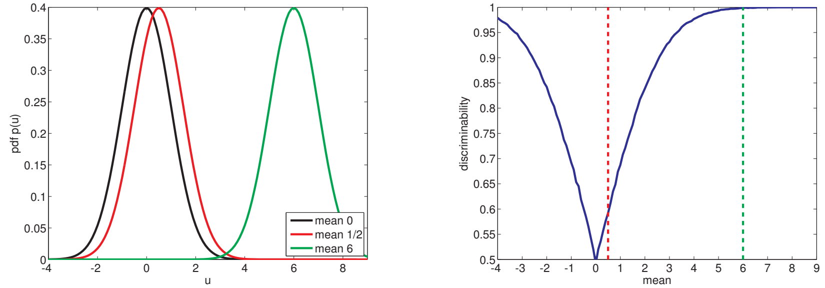


Figure 4 : Discriminability of 1/2 indicates similarity.

Application to epidemiology of infectious diseases

Data: Colonization states of sampled attendees of 29 day care centers (DCCs).

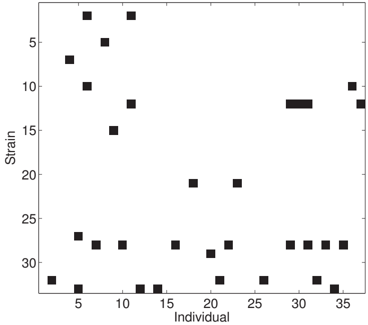


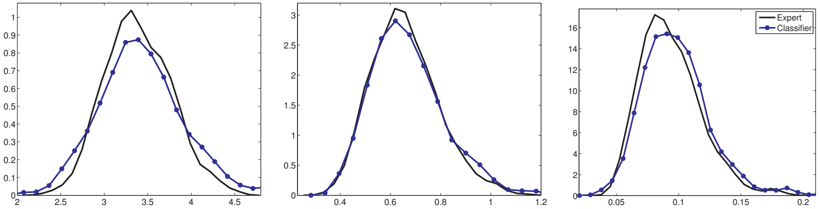
Figure 5 : Example data from a DCC. Each square indicates an attendee colonized with a strain of the bacterium *Streptococcus pneumoniae*.

Application to epidemiology of infectious diseases

- Simulator-based model: latent continuous-time Markov chain for the transmission dynamics in a DCC and an observation model (Numminen et, Biometrics, 2013).
- The model has three parameters:
  - β: rate of infections within a DCC
  - Λ: rate of infections outside a DCC
  - θ: possibility to be infected with multiple strains
- Likelihood is intractable because there are infinitely many unobserved variables (data at a single time point are available only).

Application to epidemiology of infectious diseases

- Our classification-based discrepancy measure does not use domain/expert knowledge.
- Performs as well as a discrepancy measure based on domain knowledge (Numminen et, Biometrics, 2013).

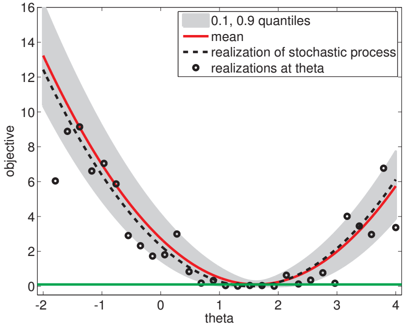


(a) Posterior pdf for β (b) Posterior pdf for Λ (c) Posterior pdf for θ

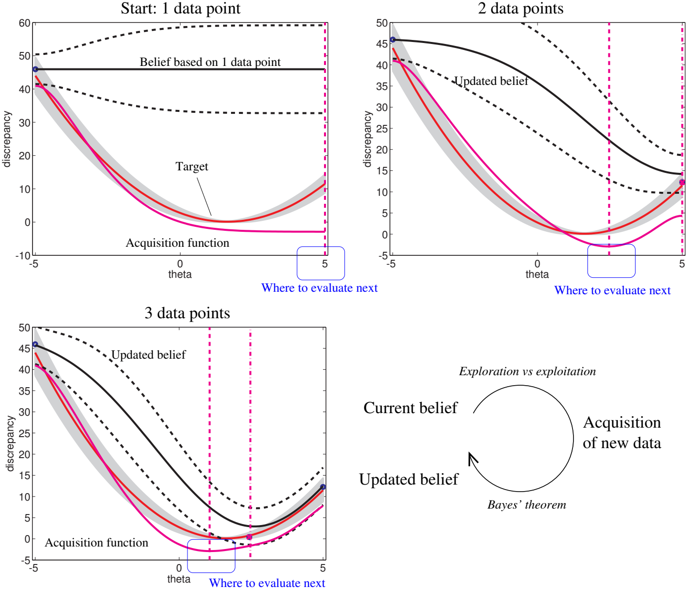
Figure 6 : The results are kernel density estimates of 1000 samples.

Increasing efficiency via Bayesian optimization

- Increase computational efficiency by evaluating $\Delta_{\boldsymbol{\theta}}$ where it tends to be small.
- This is possible by combining probabilistic modeling of $\Delta_{\boldsymbol{\theta}}$ with optimization.



Increasing efficiency via Bayesian optimization



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)

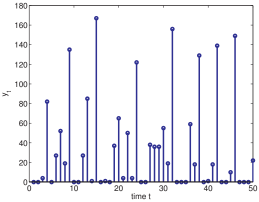


Figure 7 : Example data, $\boldsymbol{\theta}^o = (3.8, 0.3, 10)$.

Application to parameter inference in chaotic systems

- Discrepancy $\Delta_{\boldsymbol{\theta}}$ given by synthetic likelihood (Wood, Nature, 2010)
- Speed up: $\approx$ 600 times fewer evaluations of $\Delta_{\boldsymbol{\theta}}$
- Slight shift in posterior mean towards the data generating parameter $\boldsymbol{\theta}^o$ (marked by green circles)

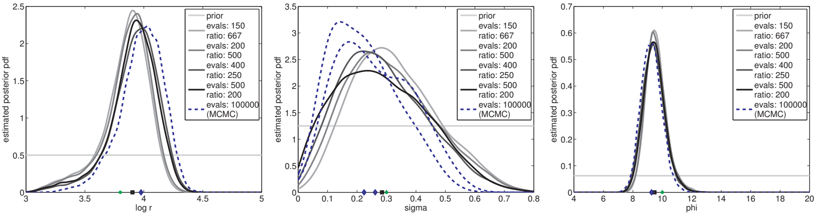


Figure 8 : Comparison with results using MCMC (Wood, Nature, 2010)

Summary

- The topic was likelihood-free inference for simulator-based models.
- Two difficulties:
  1. measurement of discrepancy (similarity)
  2. computational efficiency
- We proposed to measure the discrepancy via classification
  - Reduces the difficult problem of choosing an appropriate discrepancy measure to a standard problem
- We proposed to increase the computational efficiency via Bayesian optimization
  - Yields speed-ups of the order of one day versus one year of computation time.