Confidence Sets and Hypothesis Testing in a Likelihood-Free Inference Setting

Nic Dalmasso[1], Rafael Izbicki[2], Ann B. Lee[1]

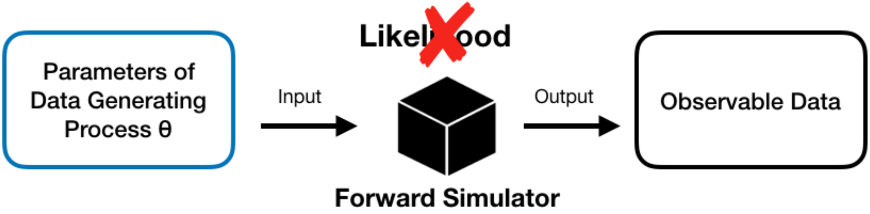
[1] Department of Statistics & Data Science, Carnegie Mellon University
[2] Department of Statistics, Federal University of Sao Carlõs

International Conference on Machine Learning (ICML)
July 12-18 2020

Motivation: Likelihood in Studying Complex Phenomena

However, for some complex phenomena in the science and engineering, an explicit likelihood function might not be available.

Likelihood-Free Inference



1. True likelihood cannot be evaluated
2. Samples can be generated for fixed settings of θ, so the likelihood is implicitly defined

Inference on parameters θ in this setting is known as likelihood-free inference (LFI).

Likelihood-Free Inference Literature

- Approximate Bayesian computation[1]
- More recent developments:
  - Direct posterior estimation (bypassing the likelihood)[2]
  - Likelihood estimation[3]
  - Likelihood ratio estimation[4]

Hypothesis testing and confidence sets can be considered cornerstones of classical statistics, but have not received much attention in LFI.

[1] Beaumont et al. 2002, Marin et al. 2012, Sisson et al. 2018

[2] Marin et al., 2016; Izbicki et al., 2019; Greenberg et al., 2019

[3] Thomas et al., 2016; Price et al., 2018; Ong et al., 2018; Lueckmann et al., 2019; Papamakarios et al., 2019

[4] Izbicki et al., 2014; Cranmer et al., 2015; Frate et al., 2016

A Frequentist Approach to LFI

Our goal is to develop:

1. valid hypothesis testing procedures
2. confidence intervals with the correct coverage

Main Challenges:

- Dealing with high-dimensional and different types of simulated data
- Computational efficiency
- Assessing validity and coverage

Hypothesis Testing and Confidence Sets

Key ingredients:

- data $\mathcal{D} = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$
- a test statistic, such as likelihood ratio statistic $\Lambda(\mathcal{D}; \theta_0)$
- an α-level critical value $C_{\theta_0,\alpha}$

Reject the null hypothesis H_0 if $\Lambda(\mathcal{D}; \theta_0) < C_{\theta_0,\alpha}$

Theorem (Neyman inversion, 1937)

Building a $1-\alpha$ confidence set for θ is equivalent to testing

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

for θ_0 across the parameter space.

Approximate Computation via Odds Ratio Estimation

Key Realization:

1. Likelihood ratio statistic $\log \Lambda(\mathcal{D}; \Theta_0)$,
2. Critical value of the test $C_{\theta_0,\alpha}$,
3. Coverage of the confidence sets

Are conditional distribution functions which often vary smoothly as a function of the (unknown) parameters of interest θ.

Rather than relying solely on samples at fixed parameter settings (standard Monte Carlo solutions), we can interpolate across the parameter space with ML models.

Likelihood Ratio Statistic (I)

1. Forward simulator F_θ
   - Identifiable model, i.e. $F_{\theta_1} \neq F_{\theta_2}$ for $\theta_1 \neq \theta_2 \in \Theta$
2. Proposal distribution for the parameters $r(\theta)$ over Θ
3. Reference distribution G over the data space $\mathcal{X}$
   - Does not depend on θ
   - G needs to be a dominating measure of F_θ for every θ
     - ⋆ OK if $G = F_\theta$ for one specific $\theta \in \Theta$

Train a probabilistic classifier m to discriminate samples from G $(Y = 0)$ between samples from F_θ $(Y = 1)$ given θ.

$$m : (\theta, \mathbf{x}) \longrightarrow \mathbb{P}(Y = 1|\mathbf{x}, \theta) \implies \mathbb{O}(\theta_0; \mathbf{x}) = \frac{\mathbb{P}(Y = 1|\mathbf{x}, \theta)}{\mathbb{P}(Y = 0|\mathbf{x}, \theta)} = \frac{F_\theta(\mathbf{x})}{G(\mathbf{x})}$$

Likelihood Ratio Statistic (II)

$$\log \mathbb{OR}(\mathbf{x};\theta_0,\theta_1) = \log \frac{\mathbb{O}(\theta_0;\mathbf{x})}{\mathbb{O}(\theta_1;\mathbf{x})} \text{ (log-odds ratio)}$$

Suppose we want to test:

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \notin \Theta_0$$

We define the test statistics:

$$\tau(\mathcal{D};\Theta_0) := \sup_{\theta_0 \in \Theta_0} \inf_{\theta_1 \in \Theta} \sum_{i=1}^{n} \log \left(\widehat{\mathbb{OR}}(\mathbf{X}_i^{\text{obs}};\theta_0,\theta_1) \right)$$

Theorem (Fisher's Consistency)

If $\widehat{\mathbb{P}}(Y=1|\theta,\mathbf{x}) = \mathbb{P}(Y=1|\theta,\mathbf{x}) \, \forall \theta,\mathbf{x} \implies \tau(\mathcal{D};\Theta_0) = \log \Lambda(\mathcal{D};\Theta_0)$

Likelihood Ratio Statistic (III)

Suppose we want to test:

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \notin \Theta_0$$

We define the test statistics:

$$\tau(\mathcal{D}; \Theta_0) := \sup_{\theta_0 \in \Theta_0} \inf_{\theta_1 \in \Theta} \sum_{i=1}^{n} \log \left(\widehat{\mathbb{OR}}(\mathbf{X}_i^{\text{obs}}; \theta_0, \theta_1) \right)$$

By fitting a classifier m we can:

- estimate $\widehat{\mathbb{OR}}(\mathbf{x}; \theta_0, \theta_1)$ for all $\mathbf{x}, \theta_0, \theta_1$,
- leverage ML probabilistic classifier to deal with high-dimensional $\mathbf{x}$,
- use loss-function as relative comparison of which classifier performs best among a set of classifiers.

Determine Critical Values $C_{\theta_0,\alpha}$

We reject the null hypothesis when $\tau(\mathcal{D};\Theta_0) \leq C_{\theta_0,\alpha}$, where $C_{\theta_0,\alpha}$ is chosen so that the test has a size α.

$$C_{\theta_0,\alpha} = \arg\sup_{C \in \mathbb{R}} \left\{ C : \sup_{\theta_0 \in \Theta_0} \mathbb{P}\left(\tau(\mathcal{D};\Theta_0) < C_{\theta_0} \mid \theta_0\right) \leq \alpha \right\},$$

Problem: Need to estimate $\mathbb{P}\left(\tau(\mathcal{D};\Theta_0) < C_{\theta_0} \mid \theta_0\right)$ over any $\theta \in \Theta$.

Solution: $\mathbb{P}(\tau(\mathcal{D};\Theta_0) < C_{\theta_0}|\theta_0)$ is a (conditional) CDF, so we can estimate its α quantile via quantile regression.

Assessing Confidence Set Coverage

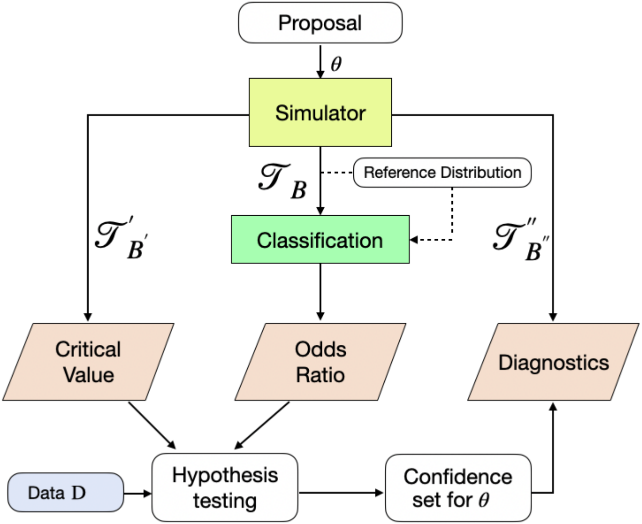
$$\text{Set Coverage: } \mathbb{E}[\mathbb{I}(\theta_0 \in R(\mathcal{D}))] = \mathbb{P}\left(\theta_0 \in R(\mathcal{D})\right) \geq 1 - \alpha$$

- **Marginal Coverage** ✗
  Build R for different $\theta_0^1, ..., \theta_0^n$ and check overall coverage
- **Estimate Via Regression** ✓
  Run `ACORE` for different $\theta_0^1, ..., \theta_0^n$ and estimate coverage:

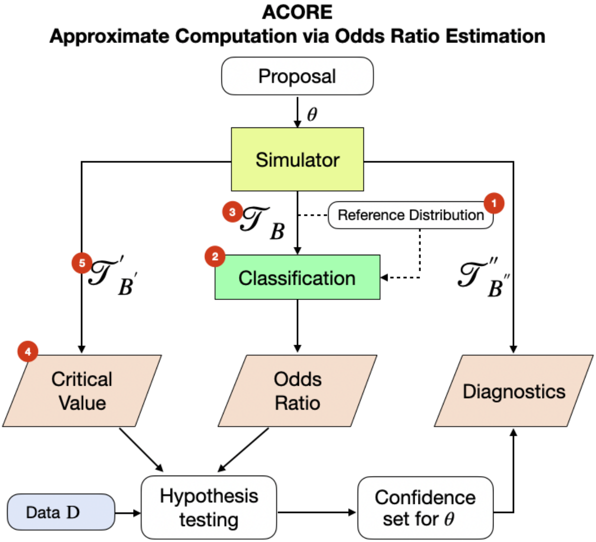
$$\{\theta_0^i, R(\mathcal{D}^i)\}_{i=1}^n \longrightarrow \text{learn } \mathbb{E}[\mathbb{I}(\theta_0 \in R(\mathcal{D}))]$$

  We can check that $1 - \alpha$ is within prediction interval for each θ_0

ACORE
Approximate Computation via Odds Ratio Estimation



ACORE Relies on 5 Key Components



A Practical Strategy

To apply ACORE, we need to choose five key components:

- a reference distribution G
- a probabilistic classifier
- a training sample size B for learning odds ratios
- a quantile regression algorithm
- a training sample size $B^{'}$ for estimating critical values

Empirical Strategy:

1. Use prior knowledge or marginal distribution of a separate simulated sample to build G;
2. Use the cross entropy loss to select the classifier and B;
3. Use the goodness-of-fit procedure to select the quantile regression method and $B^{'}$.

Also included in our work

1. Theoretical results
2. Toy examples to showcase `ACORE` in situations where the true likelihood is known
3. Signal detection example inspired by the particle physics literature
4. Comparison with existing methods
5. Open source `Python` implementation[5]
    - based on `numpy`, `sklearn` and `PyTorch`

[5] Github: Mr8ND/ACORE-LFI

THANKS FOR WATCHING!