Simulation-Base Inference

LPC Workshop

Harrison B. Prosper

Department of Physics, Florida State University

April 25, 2024

Outline

1. Introduction
   - Overview
   - Recent developments
2. Example
   - ON/OFF Problem
3. Outstanding Issues

Outline

1. Introduction
   - Overview
   - Recent developments
2. Example
   - ON/OFF Problem
3. Outstanding Issues

Suppose X are potential observations and Θ the parameter space of a theoretical model, e.g., an EFT.

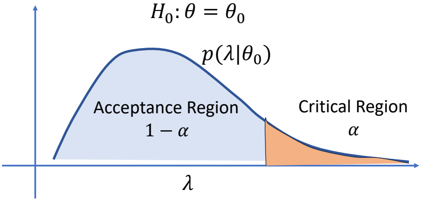
Given data D, the associated statistical model $p(X|\theta)$ and, therefore, the likelihood function $L(\theta) = p(D|\theta)$, in principle one can construct *random* sets $R(D) \in \Theta$ that satisfy

$$\mathbb{P}(\theta \in R(D)|\theta) \geq 1 - \alpha, \quad \forall\theta \in \Theta, \tag{1}$$

where $\mathbb{P}(\theta \in R(D)|\theta)$ is the coverage probability and $\tau = 1 - \alpha$ is the desired confidence level.

Sets $\{R(D)\}$ that satisfy the above conditional coverage criterion are called confidence sets. (A confidence interval is a 1-dimensional confidence set.)

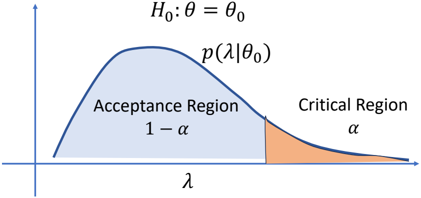
One way to construct a confidence set is via repeated hypothesis tests: $H_0 : \theta = \theta_0$. One proceeds as follows.



- Construct a function of (potential) observations X called a test statistic, $\lambda(X, \theta)$, with the property that large values of λ cast doubt on the validity of the hypothesis H_0.

- Compute $\mathbb{C}(\lambda_{\text{obs}}|\theta_0) = \mathbb{P}(\lambda \leq \lambda_{\text{obs}}|\theta_0)$ the cumulative distribution function, where $\lambda_{\text{obs}} = \lambda(D, \theta_0)$ is the observed value of the test statistic.

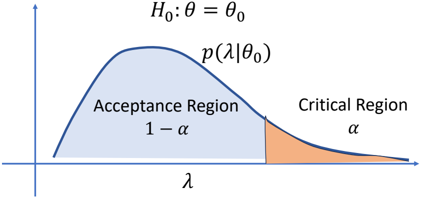
One way to construct a confidence set is via repeated hypothesis tests: $H_0 : \theta = \theta_0$. One proceeds as follows.



- Construct a function of (potential) observations X called a test statistic, $\lambda(X, \theta)$, with the property that large values of λ cast doubt on the validity of the hypothesis H_0.

- Compute $\mathbb{C}(\lambda_{\text{obs}}|\theta_0) = \mathbb{P}(\lambda \leq \lambda_{\text{obs}}|\theta_0)$ the cumulative distribution function, where $\lambda_{\text{obs}} = \lambda(D, \theta_0)$ is the observed value of the test statistic.
- Choose the confidence level $\tau = 1 - \alpha$.

One way to construct a confidence set is via repeated hypothesis tests: $H_0 : \theta = \theta_0$. One proceeds as follows.



- Construct a function of (potential) observations X called a test statistic, $\lambda(X, \theta)$, with the property that large values of λ cast doubt on the validity of the hypothesis H_0.
- Compute $\mathbb{C}(\lambda_{\text{obs}}|\theta_0) = \mathbb{P}(\lambda \leq \lambda_{\text{obs}}|\theta_0)$ the cumulative distribution function, where $\lambda_{\text{obs}} = \lambda(D, \theta_0)$ is the observed value of the test statistic.
- Choose the confidence level $\tau = 1 - \alpha$.
- If $\mathbb{C}(\lambda_{\text{obs}}|\theta_0) > \tau$ then λ_{obs} has landed in the critical region in which case reject θ_0, otherwise add it to the confidence set $R(D)$.

Simulation-based inference (SBI) (also known by the misnomer likelihood-free inference) in the frequentist approach[1] and in the context of EFT fits[2] that leverages parameterized machine learning[3] is a widely-applicable approach to inference[4] that does not require knowledge of the statistical model $p(X|\theta)$.

Here X denotes potential observations and θ the parameters of the theoretical model.

[1] *Approximating Likelihood Ratios with Calibrated Discriminative Classifiers*, Kyle Cranmer, Juan Pavez, and Gilles Louppe.

[2] *Constraining Effective Field Theories with Machine Learning*, Johann Brehmer, Kyle Cranmer, Gilles Louppe, Juan Pavez; *A Guide to Constraining Effective Field Theories with Machine Learning*, Johann Brehmer, Kyle Cranmer, Gilles Louppe, Juan Pavez.

[3] *Parameterized Machine Learning for High-Energy Physics*, P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, arXiv preprint arXiv:1601.07913.

[4] *MadMiner: Machine learning-based inference for particle physics*, J. Brehmer, F. Kling, I. Espejo, K. Cranmer, Comput.Softw.Big Sci. 4 (2020) 1, 3.

Recently, Lee et al. introduced a method they call likelihood-free frequentist inference (LF2I)[5]. As in the SBI methods cited, LF2I,

1. does not presume the validity of Wilks' theorem and its variants[6];

[5] *Likelihood-Free Frequentist Inference: Confidence Sets with Correct Conditional Coverage*, Niccolò Dalmasso, Luca Masserano, David Zhao, Rafael Izbicki, Ann B. Lee, arXiv:2107.03920v6 [stat.ML] 6 Apr 2023.

[6] G. Cowan, K. Cranmer, E. Gross, O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur.Phys.J.C71:1554, 2011.

Recently, Lee et al. introduced a method they call likelihood-free frequentist inference (LF2I)[5]. As in the SBI methods cited, LF2I,

1. does not presume the validity of Wilks' theorem and its variants[6];
2. does not require knowledge of the statistical model;

[5] *Likelihood-Free Frequentist Inference: Confidence Sets with Correct Conditional Coverage*, Niccolò Dalmasso, Luca Masserano, David Zhao, Rafael Izbicki, Ann B. Lee, arXiv:2107.03920v6 [stat.ML] 6 Apr 2023.

[6] G. Cowan, K. Cranmer, E. Gross, O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur.Phys.J.C71:1554, 2011.

Recently, Lee et al. introduced a method they call likelihood-free frequentist inference (LF2I)[5]. As in the SBI methods cited, LF2I,

1. does not presume the validity of Wilks' theorem and its variants[6];
2. does not require knowledge of the statistical model;
3. exploits the fact that confidence sets for all parameters taken together can always be constructed;

[5] *Likelihood-Free Frequentist Inference: Confidence Sets with Correct Conditional Coverage*, Niccolò Dalmasso, Luca Masserano, David Zhao, Rafael Izbicki, Ann B. Lee, arXiv:2107.03920v6 [stat.ML] 6 Apr 2023.

[6] G. Cowan, K. Cranmer, E. Gross, O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur.Phys.J.C71:1554, 2011.

Recently, Lee et al. introduced a method they call likelihood-free frequentist inference (LF2I)[5]. As in the SBI methods cited, LF2I,

1. does not presume the validity of Wilks' theorem and its variants[6];
2. does not require knowledge of the statistical model;
3. exploits the fact that confidence sets for all parameters taken together can always be constructed;
4. exploits the relationship, we've just sketched, between classical hypothesis tests and confidence sets, and
5. leverages high-fidelity simulators and machine learning.

[5] *Likelihood-Free Frequentist Inference: Confidence Sets with Correct Conditional Coverage*, Niccolò Dalmasso, Luca Masserano, David Zhao, Rafael Izbicki, Ann B. Lee, arXiv:2107.03920v6 [stat.ML] 6 Apr 2023.

[6] G. Cowan, K. Cranmer, E. Gross, O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur.Phys.J.C71:1554, 2011.

ALFFI Recently LF2I was extended[7] to make it possible to construct confidence sets and check their coverage using $\mathbb{C}(\lambda_{\text{obs}}|\theta)$.

Given the discrete random variable $Z = \mathbb{I}[\lambda(X,\theta) \leq \lambda(X',\theta)]$, where X is conditional on θ while X' is not, and a simulator $X \sim \mathbb{F}[\theta]$, the method approximates

$$C(\lambda_{\text{obs}}|\theta) = \mathbb{E}[Z|\lambda_{\text{obs}},\theta], \tag{2}$$

by minimizing the average quadratic loss,

$$\mathcal{R}(\omega) = \frac{1}{N}\sum_{i=1}^{N}[Z_i - f(\lambda_i',\theta_i;\omega)]^2, \tag{3}$$

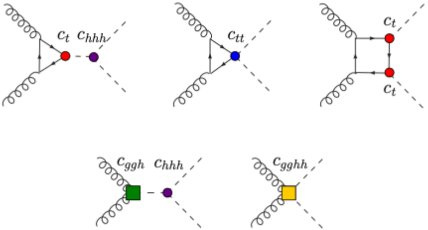
where $f(\lambda',\theta;\omega)$ is a deep neural network (DNN). Details next slide.

[7] Amortized simulation-based frequentist inference for tractable and intractable likelihoods, Ali Al Kadhim, HBP, and Olivia F Prosper, Mach. Learn.: Sci. Technol. 5 (2024) 015020.

Algorithm 1 Amortized Likelihood-Free Inference (ALFFI)

```
1. Initialize samples: $\mathbb{X} \leftarrow \varnothing$, $\mathbb{T} \leftarrow \varnothing$
while $k \in [1, \cdots K]$ do
    2. Sample $\theta_k \sim \pi(\theta)$
    3. Sample $X_k \equiv X_{1,k}, \cdots, X_{n,k} \sim \mathbb{F}(\theta_k)$
    4. Update training sample $\mathbb{X} \leftarrow \mathbb{X} \cup \{(\theta_k, X_k)\}$
end while
5. Define $\mathbb{Y} = \{(\theta_k, X'_k)\}$ by randomly shuffling $X_k$ relative to $\theta_k$
while $k \in [1, \cdots K]$ do
    6. Compute test statistic $\lambda_k \leftarrow \lambda(X_k, \theta_k)$
    7. Compute test statistic $\lambda'_k \leftarrow \lambda(X'_k, \theta_k)$
    8. Compute indicator $Z_k \leftarrow \mathbb{I}(\lambda_k \leq \lambda'_k)$
    9. Update training sample $\mathbb{T} \leftarrow \mathbb{T} \cup \{(Z_k, \lambda'_k, \theta_k)\}$
end while
10. Train a DNN, $f(\lambda', \theta; \omega)$, to approximate $\mathbb{C}(\lambda' | \theta)$.
```

Un-binned test statistic for EFTs. Consider $pp \to hh$ where the lowest order diagrams are shown below. In the HEFT[8], the process is characterized by 5 Wilson coefficients, $\theta = c_{hhh}, c_t, c_{tt}, c_{ggh}, c_{gghh}$.



The Standard Model $(\theta = 1, 1, 0, 0, 0)$ (density $s(X)$) is nested within the HEFT (density $e(X|\theta)$). Therefore, for a *single* event a natural test statistic to consider is

$$\lambda(X, \theta) = \ln \frac{s(X)}{e(X|\theta)},$$

which is being explored by FSU graduate student Bobby Goff in the context of $t\bar{t}\gamma$. For N events, the statistic is $\lambda_N = \frac{1}{N} \sum_{i=1}^{N} \lambda(X_i, \theta)$.

[8] Effective Field Theory descriptions of Higgs boson pair production, arXiv:2304.01968v1

Un-binned test statistic for EFTs Given a balanced data set of SM and EFT events where the data comprise pairs (X, θ), the test statistic λ can be approximated directly by minimizing the average exponential loss or indirectly by minimizing the average cross entropy.
Minimizing the latter yields an approximation to the discriminant

$$\begin{aligned} \mathcal{D}(X, \theta) &= \frac{e(X, \theta)}{e(X, \theta) + s(X, \theta)}, \\ &= \frac{e(X|\theta)\pi_\theta}{e(X|\theta)\pi_\theta + s(X)\pi_\theta}, \end{aligned} \tag{4}$$

where $e(*)$ and $s(*)$ are the EFT and SM densities, respectively. (The SM density factorizes because X and θ are statistically independent.)
A rearrangement yields the desired result

$$\lambda(X, \theta) = \ln \frac{1 - D}{D}. \tag{5}$$

Outline

1. Introduction
   - Overview
   - Recent developments
2. Example
   - ON/OFF Problem
3. Outstanding Issues

In an ON/OFF experiment[9], the data comprise two independent counts $D = N, M$ obtained under the signal plus background condition (ON) or the background-only condition (OFF). In the simplest case, the statistical model is

$$p(X, Y|\theta) = \text{Poisson}(X, \mu + \nu)\text{Poisson}(Y, \nu),$$

where X and Y are random counts.

Obviously, this model does not require simulation-based inference! But it does serve as an example of a problem that violates two of the regularity conditions for the validity of Wicks' theorem.

[9] T. P. Li and Y. Q. Ma, *Analysis method for results in gamma-ray astronomy*, Astrophys. J. **272**, 313 (1983); J. T. Linnemann, *Measures of Significance in HEP and Astrophysics*, PHYSTAT2003, SLAC, Stanford CA, September 8-11, 2003; R. D. Cousins, J. T. Linnemann, J. Tucker, *Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process*, NIM A **595** 480-501 (2008).

We use the following test statistic

$$\lambda(D, \theta) = -2 \log \left[\frac{p(D|\mu, \nu)}{p(D|\hat{\mu}, \hat{\nu})} \right], \tag{6}$$

where $\hat{\mu}$ and $\hat{\nu}$ are the best-fit values of the parameters.

We use the following test statistic

$$\lambda(D, \theta) = -2 \log \left[\frac{p(D|\mu, \nu)}{p(D|\hat{\mu}, \hat{\nu})} \right], \tag{6}$$

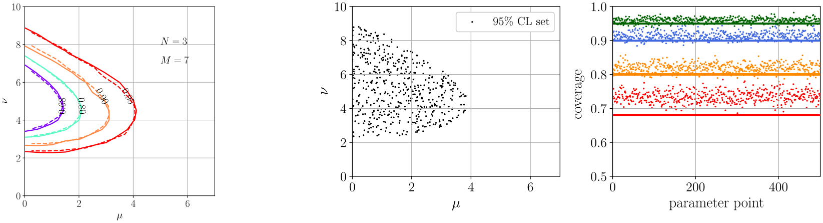
where $\hat{\mu}$ and $\hat{\nu}$ are the best-fit values of the parameters. Since $\mu \geq 0$, we take the estimate of the mean signal to be

$$\hat{\mu} = \begin{cases} N - M & \text{if} \quad N > M \\ 0 & \quad\quad \text{otherwise,} \end{cases}, \tag{7}$$

which explicitly violates the regularity condition that estimates must lie in the interior of the parameter space. We also choose low counts $N = 3$ and $M = 7$, which are a tad short of the asymptotic regime. For the estimate of the mean background, we take

$$\hat{\nu} = \begin{cases} M & \text{if} \quad \hat{\mu} = N - M \\ (M + N)/2 & \quad \text{otherwise.} \end{cases} \tag{8}$$

Training a simple few-layer neural network yields the following confidence sets and coverage probabilities.



The coverage probabilities shown in the rightmost plot at the parameter points displayed in the middle plot are indeed bounded by the confidence levels $1 - \alpha$ even for the sparse data.

Outline

1. Introduction
   - Overview
   - Recent developments
2. Example
   - ON/OFF Problem
3. Outstanding Issues

Simulation-based inference (SBI) is particularly useful when the statistical model is intractable or extremely complicated. In contrast to the benchmark approach, where large numbers of events are simulated at a few parameter points, SBI requires simulating a few events at each of a large number of parameter points. Several outstanding issues remain, a few of which are listed below.

- Accuracy of confidence sets in many dimensions (can we use conformal inference to improve accuracy?)

Simulation-based inference (SBI) is particularly useful when the statistical model is intractable or extremely complicated. In contrast to the benchmark approach, where large numbers of events are simulated at a few parameter points, SBI requires simulating a few events at each of a large number of parameter points. Several outstanding issues remain, a few of which are listed below.

- Accuracy of confidence sets in many dimensions (can we use conformal inference to improve accuracy?)
- Given $\mathbb{C}(\lambda|\theta)$, we can compute the pdf of λ using $f(\lambda|\theta) = \partial\mathbb{C}/\partial\lambda$ via automatic differentiation. How good is this approach compared with the likelihood ratio trick?

Simulation-based inference (SBI) is particularly useful when the statistical model is intractable or extremely complicated. In contrast to the benchmark approach, where large numbers of events are simulated at a few parameter points, SBI requires simulating a few events at each of a large number of parameter points. Several outstanding issues remain, a few of which are listed below.

- Accuracy of confidence sets in many dimensions (can we use conformal inference to improve accuracy?)
- Given $\mathbb{C}(\lambda|\theta)$, we can compute the pdf of λ using $f(\lambda|\theta) = \partial\mathbb{C}/\partial\lambda$ via automatic differentiation. How good is this approach compared with the likelihood ratio trick?
- Correct conditional coverage is achieved for *all* parameters taken together. Can one map a confidence set to a confidence interval with a desired confidence level?