Simulation Based Inference
The Frequentist Perspective
AISSAI

Harrison B. Prosper

Department of Physics, Florida State University

28 November, 2023

Outline

1. Introduction

2. Amortized Likelihood Free Frequentist Inference

3. Examples

4. Summary

Outline

1. Introduction

2. Amortized Likelihood Free Frequentist Inference

3. Examples

4. Summary

As described in the talk by Gilles Louppe, simulation-based inference is a way to make inferences about the parameters θ of a statistical model $p(X|\theta)$ without using the model explicitly. The model is implicitly defined via a forward simulator, $D \sim \mathbb{F}(\theta)$, that can simulate data sets D given parameters θ.

Recently, a method called likelihood-free frequentist inference (LF2I) was introduced by Prof. Ann Lee[1] and her group at Carnegie Mellon University that features correct conditional coverage.

I'll begin with a brief reminder of what that means and follow with a description of a modified version of this method, called amortized likelihood-free frequentist inference (ALFFI), which is illustrated with a couple of simple examples.

[1] 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.

Consider a large ensemble of experiments (say all those that have been performed since the discovery of the electron).

For each experiment, we assert that the true value of some parameter θ lies in some subset $R(D)$ of the parameter space associated with the experiment. Each such statement is either True or False.

In frequentist inference, it is required that the fraction of true statements, that is, the coverage probability, over an ensemble of statements of the form $\theta \in R(D)$ never fall below the claimed confidence level (CL) $1 - \alpha$, where α is typically a small number.

Random sets $\{R(D)\}$ with this property, of which a confidence interval is a special case, are called confidence sets.

The LF2I approach of Lee et al. is a method for constructing confidence sets, $R(D)$, which

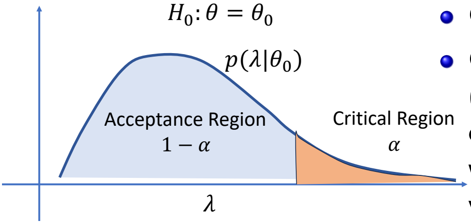
1. does not presume the validity of Wilks' theorem and its variants[2] and, therefore, works for finite data samples and
2. does not require knowledge of the statistical model, and, therefore, the likelihood function.

The method

1. exploits the fact that confidence sets for all the parameters taken together can always be constructed;
2. exploits the close relationship between classical hypothesis tests and confidence sets, and
3. leverages the availability of high-fidelity simulators and machine learning.

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

Consider the hypothesis $H_0 : \theta = \theta_0$. It is typically tested as follows.



- Choose a small probability α;
- 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 the p-value $= \mathbb{P}(\lambda > \lambda_{\text{obs}}|\theta_0) = 1 - \mathbb{C}(\lambda_{\text{obs}}|\theta_0)$, where $\lambda_{\text{obs}} = \lambda(D, \theta_0)$ is the observed value of the test statistic, and the cumulative distribution function is given by

$$\mathbb{C}(\lambda_{\text{obs}}|\theta_0) = \int_{Y \leq \lambda_{\text{obs}}} dY \int dX \, \delta(Y - \lambda(X, \theta_0)) \, p(X|\theta_0). \qquad (1)$$

- If the p-value $< \alpha$ then the test statistic has landed in the so-called critical region in which case the parameter value θ_0 is rejected.

If $H_0 : \theta = \theta_0$ is true, then by construction the probability to reject θ_0 is α[3]. Therefore, the probability not to reject θ_0 is $1 - \alpha$.

In other words, we keep θ_0 whenever the p-value $\geq \alpha$ or, equivalently, whenever $\mathbb{C}(\lambda_{\text{obs}}|\theta) \leq 1 - \alpha$.

For a given data set D, the confidence set $R(D)$ is constructed by collecting together all values of θ that are kept.

Therefore, the task is to approximate either the p-value or the cumulative distribution function, which is the basis of the LF2I method and its recent variant ALFFI.

[3] Rejecting a true hypothesis is called a Type 1 error.

Outline

1. Introduction
2. Amortized Likelihood Free Frequentist Inference
3. Examples
4. Summary

Algorithm 1 Amortized likelihood-free frequentist inference (ALFFI)

```
1. Initial training 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. Produce a second data sample, $\mathbb{Y} = \{(\theta_k, X_k)\}$, to serve as instances
of "observed" data by randomly shuffling the $X_k$ relative to the $\theta_k$
while $k \in [1, \cdots K]$ do
    6. Compute test statistic $\lambda_k \leftarrow \lambda(X_k, \theta_k)$
    7. Compute test statistic $\lambda'_k \leftarrow \lambda(Y_k, \theta_k)$
    8. Compute indicator $Z_k \leftarrow \mathbb{I}(\lambda_k \leq \lambda'_k)$
    9. Update training sample $\mathbb{T} \leftarrow \mathbb{T} \cup \{(\theta_k, \lambda'_k, Z_k)\}$
end while
10. Train an ML model, $f(\theta, \lambda_{\text{obs}}; \omega)$, to approximate $\mathbb{C}(\lambda_{\text{obs}}|\theta)$.
```

Since $Z = \mathbb{I}(\lambda \leq \lambda_{\text{obs}})$ then, for a given θ, the probability that $\lambda \leq \lambda_{\text{obs}}$ is the same as the probability that $Z = 1$, which, in turn, is the same as the conditional expectation value $\mathbb{E}[Z|\theta]$, which is approximated with a neural network.

The neural network models of interest are trained by minimizing an empirical risk function (aka cost function, average loss function), given by

$$\mathbb{R}(\omega) = \frac{1}{K}\sum_{i=1}^{K} L(f_i, t_i), \quad f_i \equiv f(x_i;\omega), \tag{2}$$

where t_i are known targets associated with known inputs x_i and $L(f, t)$ is a loss function.

In the limit of an infinite training sample the empirical risk function becomes the risk functional $\mathbb{R}[f]$,

$$\mathbb{R}[f] = \int \left[\int L(f, t)\, p(t|x)\, dt\right] p(x)\, dx. \tag{3}$$

If

- the training data sample is large enough, and
- the ML model has sufficient capacity, and
- a good approximation to the minimum of the risk functional can be found,

then, provided that $p(x) > 0 \,\forall\, x$, minimizing the risk functional $\mathbb{R}[f]$ yields the important result,

$$\int \frac{\partial L}{\partial f} p(t|x)\, dt = 0, \tag{4}$$

which is a generalization of a result from the early 1990s.

LF2I uses the quadratic loss $L(f, t) = (f - t)^2$ with the targets set to $t = Z$. According to the Eq. (4), this implies that the best-fit ML model parameters ω^* yield a trained ML model that satisfies,

$$f(\theta; \omega^*) \approx p(Z = 1|\theta) \equiv \mathbb{P}(\lambda \leq \lambda_{\text{obs}}|\theta). \tag{5}$$

Outline

1. Introduction
2. Amortized Likelihood Free Frequentist Inference
3. Examples
4. Summary

ON/OFF Experiment

In an ON/OFF experiment, 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.

When data D are entered into the model, we arrive at the likelihood function

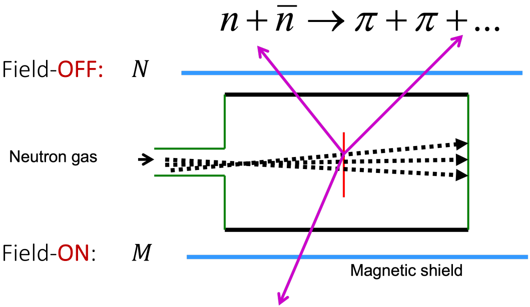
$$p(D|\theta) = \text{Poisson}(N, \mu + \nu)\text{Poisson}(M, \nu).$$

Usually, we don't care about ν, the mean background, but we'll pretend that we do!

The search for neutron-antineutron oscillations at the Institut Laue Langevin (ILL) in Grenoble, France (1980 - 1985) is a pedagogically perfect example of an ON/OFF experiment in particle physics.

H18 cold neutron beam: neutron flux 1.5×10^9n/s, neutron temperature $\sim$ 1.5K (neutron speed $\sim$ 160m/s).

The CERN-Rutherford-ILL-Sussex-Padova Collaboration[4] conducted the experiment sketched below.



Results:

$$N = 3 \quad \text{field-OFF events,}$$
$$M = 7 \quad \text{field-ON events.}$$

[4] G. Fidecaro et al., "Experimental search for neutron-antineutron transitions with free neutrons", Phys. Lett. B 156, 122 (1985).

We use the following test statistic

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

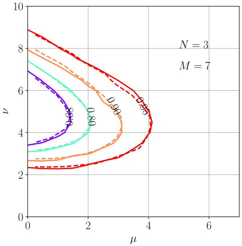
where $\hat{\mu}$ and $\hat{\nu}$ are the best-fit values of the parameters. Since $\mu \geq 0$, we take the estimate of the mean signal to be

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

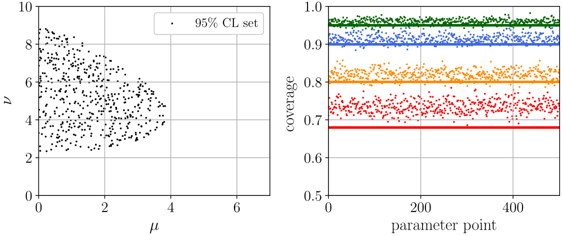
which explicitly violates one of the regularity conditions for the validity of Wilks' theorem, namely, that estimates must lie within the interior of the parameter space. For the estimate of the mean background, we take

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

A simple feed-forward neural network is trained that yields the following confidence sets and coverage probabilities.



Confidence sets



Coverage

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 of the Grenoble experiment.

In our second example the statistical model is intractable, which of course is where simulation-based inference is most needed.

The susceptible-infected-recovered (SIR) model is the prototypical model of an epidemic. In this model, individuals in the susceptible class, S, can migrate to the infected class, I, and from there to the recovered (or removed) class, R. We apply this model to a widely used data set from a flu outbreak more than a century ago at an English Boarding School.

The mean counts in the three classes are governed by the equations

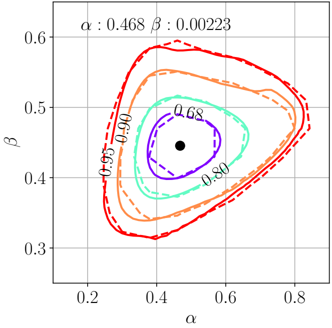
$$\begin{aligned}\frac{dS}{dt} &= -\beta S I, \\ \frac{dI}{dt} &= -\alpha I + \beta S I, \\ \frac{dR}{dt} &= \alpha I, \end{aligned} \tag{9}$$

where $\theta = \alpha, \beta$ are the model parameters.

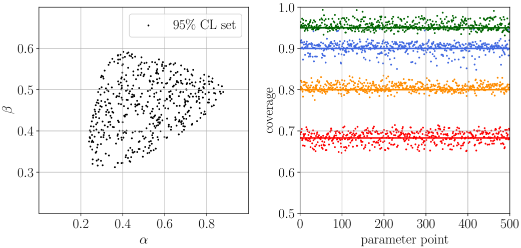
The test statistic is chosen to be

$$\lambda(D, \theta) = \sqrt{\sum_{n=1}^{N} \frac{(x_n - I_n(\theta))^2}{I_n(\theta)}}, \tag{10}$$

where x_n are the observed number of infected school children on a given day. The likelihood function is intractable because the counts are correlated across time and the fluctuations are super-Poissonian.



Again, a relatively simple neural network is trained to approximate $\mathbb{P}(\lambda \leq \lambda_{\text{obs}} | \theta)$ and is used to compute the solid contours in the figure to the left. The dashed lines are obtained, as before, with the histogram approximation. We see good agreement between the two approximations.



There is some under-coverage, but overall the results are reasonable.

The examples chosen for illustration are very simple.

It remains to be seen how well the method scales to large problems and whether a way can be found to map the confidence sets to confidence intervals for individual parameters in a way that gives correct coverage.

Outline

1. Introduction
2. Amortized Likelihood Free Frequentist Inference
3. Examples
4. Summary

- If a high-fidelity simulator is available, the LF2I approach can be used to create confidence sets with good coverage and, in principle, exact coverage.
- A simple modification (ALFFI) makes it possible to use the same network to construct confidence sets and check their coverage explicitly.
- Two simple examples illustrate the potential of simulation-based frequentist inference, but work is needed to find algorithm to map from confidence sets to confidence intervals.
- The LF2I approach contains methods to compute confidence sets for subsets of the parameters, but, alas, without frequentist guarantees for small samples.

This work was performed in collaboration with Ali Kadhim (FSU) and Prof. Olivia Prosper (U. of Tennessee).

I wish to thank Prof. Ann Lee for fruitful discussions at Aspen last year.

- This work was performed in part at the Aspen Center for Physics, which is supported by US NSF grant PHY-1607611.
- This work was supported in part by US DoE grant DE-SC0010102 (AK, HP).
- This work was supported in part by US NSF grant DMS-2045843 (OP)

The likelihood-free frequentist inference (LF2I) approach comprises several components, including an algorithm for approximating the p-value or the cumulative distribution function, which is shown below.

Algorithm 2 LF2I approximation of $\mathbb{C}(\lambda_{\text{obs}}|\theta)$ given a simulator $\mathbb{F}(\theta)$

```
1. Initialize training sample $\mathbb{T} \leftarrow \varnothing$
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. Compute test statistic $\lambda_k \leftarrow \lambda(X_k, \theta_k)$
    5. Compute test statistic $\lambda_{\text{obs},k} \leftarrow \lambda(D, \theta_k)$
    6. Compute indicator $Z_k \leftarrow \mathbb{I}(\lambda_k \leq \lambda_{\text{obs},k})$
    7. Update training sample $\mathbb{T} \leftarrow \mathbb{T} \cup \{(\theta_k, Z_k)\}$
end while
8. Use $\mathbb{T}$ to train a machine learning (ML) model, $f(\theta;\omega)$, to approximate $\mathbb{C}(\lambda_{\text{obs}}|\theta)$, where $\theta$ are the inputs to $f(\theta;\omega)$, and $\omega$ are the ML model parameters.
```

As noted, the power of simulation-based inference is that knowledge of the statistic model is not needed. Moreover, LF2I and ALFFI work for samples of all sizes. However, it is useful to have simple benchmark models, with known likelihoods, to validate and illustrate the method.

We first apply ALFFI to a cosmological model that is fitted to the Union 2.1 compilation of data for 580 Type 1a supernova[5].

For the test statistic, we use the function

$$\lambda = \sum_{i=1}^{N} \left(\frac{x_i - \mu(z_i, \theta)}{\sigma_i} \right)^2, \tag{11}$$

where $x_i \pm \sigma_i$ are the measured distance moduli, $\mu(z, \theta)$ the predicted distance modulus function, and z_i the measured supernovae red shifts, which are accurately known.

[5] https://www.supernova.lbl.gov/

Our toy cosmological model is defined by the rather odd equation of state

$$\mathcal{P} = -\frac{1}{3} n a^n \Omega(a), \tag{12}$$

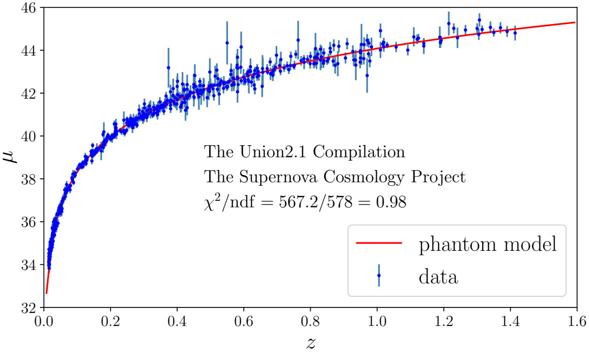
where n is a free parameter, and $a(t)$, $\Omega(a)$, and $\mathcal{P}$ are the dimensionless universal scale factor, the dimensionless energy density, and the dimensionless pressure, respectively, and t is the time since the Big Bang.

This equation of state yields the energy density

$$\Omega(a) = \exp(a^n - 1) \, / \, a^3. \tag{13}$$

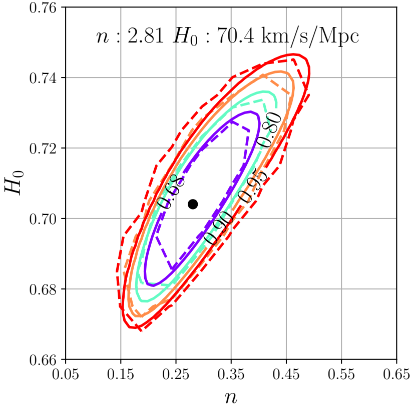
About the only virtue of this model is that it has only two parameters, the other being the Hubble constant $\mathcal{H}_0$ (not to be confused with an hypothesis), and the model can be exactly integrated.

When the cosmological model is fitted to the Type 1a data by minimizing λ (using, for example, iminuit), the following excellent fit is found.



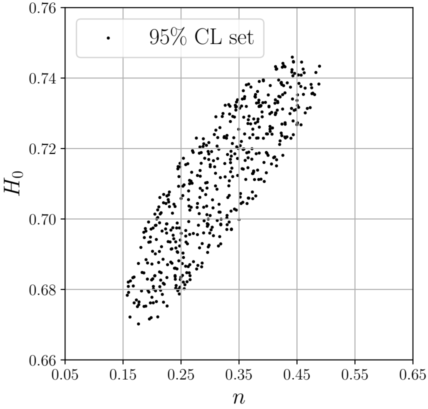
By the way, the model predicts that the universe will self-destruct in a Big Rip at about 1.4 times its current age!

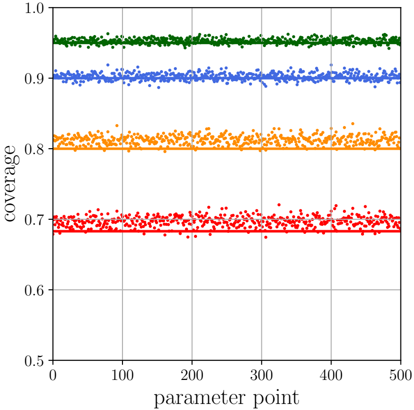
We approximate $\mathbb{P}(\lambda \leq \lambda_{\text{obs}})$ using a 5-layer fully-connected feed-forward neural network, with 20 nodes per layer, a single output, and `ReLU` non-linearities. The confidence sets are shown in the figure below.



The solid contours are computed with ALFFI, while the dashed contours are computed by approximating $\mathbb{E}[Z|\theta]$ using the ratio $\mathbb{H}_Z/\mathbb{H}_1$ of two 2D histograms, one ($\mathbb{H}_Z$) in which entries are weighted by the indicators Z and the other ($\mathbb{H}_1$) uses unit weights.

In ALFFI, unlike LF2I, the "observed" test statistic is an input to the neural network model. Therefore, we can directly check the coverage by simulating ensembles of data sets at many randomly selected points within the parameter space and explicitly counting how often the confidence sets at each point contain that point.


95% CL set
H_0
n


coverage
parameter point