Likelihood Free Inference

PHYSTAT Seminar

Harrison B. Prosper

Department of Physics, Florida State University

27 September, 2023

Outline

Outline

Let D denote data generated from a (typically unknown) probability distribution G, where G is presumed to be the manifestation of physical processes that we wish to understand.

To that end, we build statistical models, $p(X|\theta)$, of G, where θ denote the model parameters. The most common use case is that only a subset of the parameters are of interest. However, in this talk I'll assume we are interested in all of the parameters.

I'll start with a pedagogical introduction to likelihood-free frequentist inference (LF2I), which was introduced recently by Prof. Ann Lee[1] and her group at Carnegie Mellon University using as an example an ON/OFF experiment performed at the Institut Laue Langevin in Grenoble in the early 1980s. The method will then be illustrated with a few 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.

So what is an ON/OFF experiment?

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

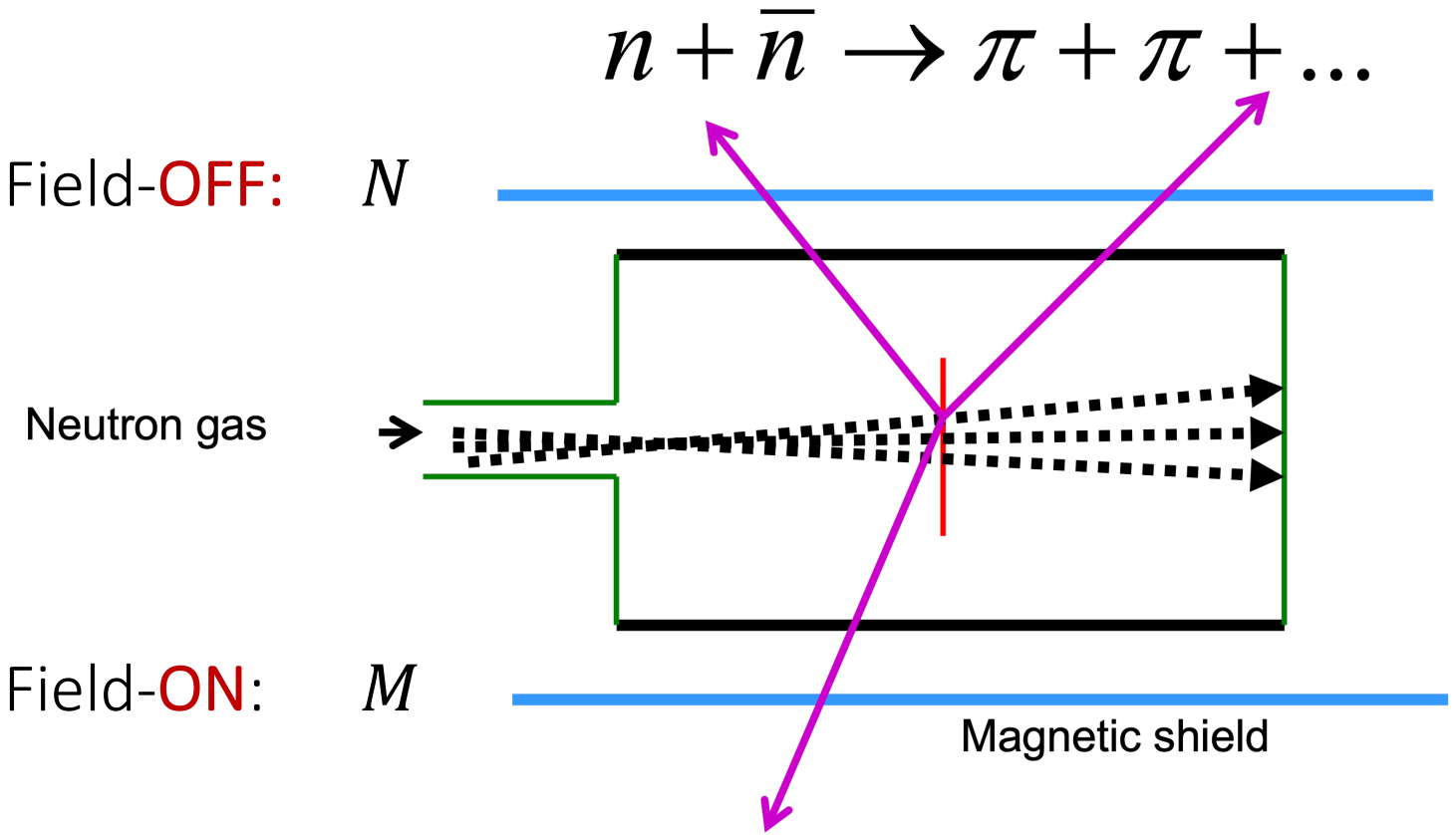
$$p(D|\theta) = \text{Poisson}(N, \mu + \nu)\text{Poisson}(M, \nu).$$

Usually, we don't care about ν, the mean background, but for today 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 $1.5 \times 10^9 \text{n/s}$, neutron temperature $\sim 1.5\text{K}$ (neutron speed $\sim 160\text{m/s}$).

The CERN-Rutherford-ILL-Sussex-Padova Collaboration[2] conducted the experiment sketched below.



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

In this experiment the following results were obtained

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

For the purposes of this talk, we'll assume that the goal is to create a subset $R(D)$ within the $\theta = \mu, \nu$ parameter space of the likelihood function, which, in some sense, is most likely to contain the true value of θ.

By "some sense" and "most likely" we mean the following,

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

that is, the probability that θ lies within the subset $R(D)$ is at least $1 - \alpha$ for every plausible value of θ in the parameter space Θ.

Today we restrict ourselves to the frequentist meaning of the probability $\mathbb{P}$.

Consider all the experiments[3] that have been performed since the discovery of the electron in 1897.

Let's imagine their re-analysis using the same method to construct a set $R(D)$ for each experiment, and let's choose $\alpha = 0.05$. In general, the meaning of D and θ differs from one experiment to the next.

For each experiment, we assert that the true (unknown) value of $\theta \in R(D)$. Each such statement is either True or False.

Our task is to devise a method such that the fraction of true statements, that is, the coverage probability, over the ensemble of statements is greater than or equal to the confidence level (CL) $1 - \alpha = 0.95$.

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

[3] In principle, we need to use an infinitely large ensemble of experiments.

The LF2I approach provides a method for constructing confidence sets, $R(D)$, which

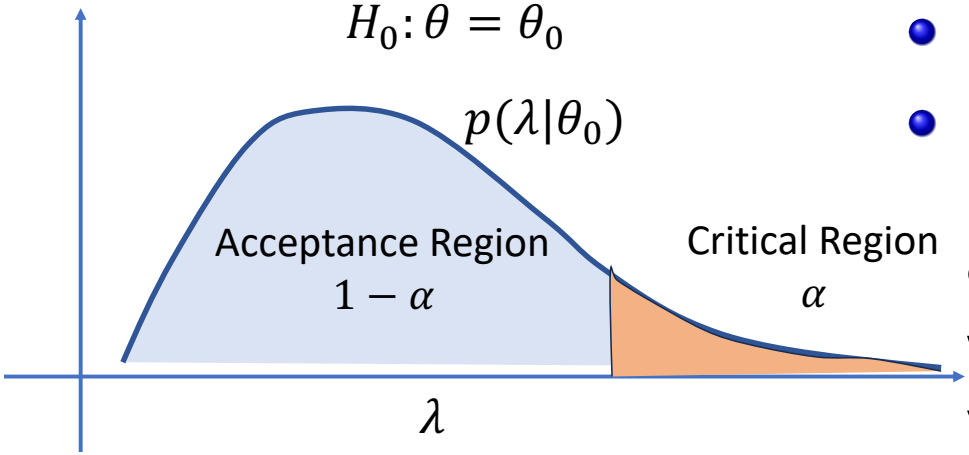
1. does not presume the validity of Wilks' theorem and its variants[4] and, therefore, works for finite data samples and
2. does not require knowledge of the statistical model, and, therefore, the likelihood function.

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 in many scientific fields, and the power of machine learning.

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

For each hypothesis $H_0 : \theta = \theta_0$, we proceed 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

$$\mathbb{C}(\lambda_{\text{obs}}|\theta_0) = \int_{Y \leq \lambda_{\text{obs}}} dY \int dX\, \delta(Y - \lambda(X, \theta_0))\, p(X|\theta_0). \quad (2)$$

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

By construction the probability to reject θ_0 is α[5] if H_0 is true. 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 to which we now turn.

[5] If the hypothesis H_0 is true then the probability of falsely rejecting it is α. Rejecting a true hypothesis is called a Type 1 error.

Outline

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 1 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 θ are the inputs to $f(\theta;\omega)$, and ω are the ML model parameters.

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 $\mathbb{E}[Z|\theta]$.

The ML models of interest are trained through empirical risk minimization, that is, by minimizing an empirical risk function (aka cost function), given by

$$\mathbb{R}(\omega) = \frac{1}{K}\sum_{i=1}^{K} L(f_i, t_i), \quad f_i \equiv f(x_i; \omega), \tag{3}$$

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 $\mathbb{R}[f]$,

$$\mathbb{R}[f] = \int \left[\int L(f, t)\, p(t|x)\, dt\right] p(x)\, dx. \tag{4}$$

If

- the training data sample is large enough, and
- the ML model has sufficient capacity (that is, it is capable of modeling all desired functions), and
- the minimizer can find a good approximation to the minimum of the risk functional,

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

$$\int \frac{\partial L}{\partial f} p(t|x)\, dt = 0. \tag{5}$$

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

$$f(\theta;\omega^*) \approx p(Z=1|\theta) \equiv \mathbb{P}(\lambda \leq \lambda_{\text{obs}}|\theta). \tag{6}$$

The LF2I algorithm generates confidence sets by simply testing whether a given point θ satisfies the condition $f(\theta;\omega^*) \leq 1-\alpha$.

As noted earlier, LF2I comprises several components. One of components is an algorithm to model the coverage probability as a function of θ. In principle, this is a marvelous way to check the quality of the approximation $f(\theta;\omega^*) \approx \mathbb{P}(\lambda \leq \lambda_{\text{obs}}|\theta)$.

But, unfortunately, to use the coverage probability function as a diagnostic for the quality of $f(\theta;\omega^*)$ we need to be sure that the coverage probability function itself is well modeled!

It is these difficulties and the desire to avoid the need for building multiple ML models that inspired the modification we recently proposed, following a fruitful discussion I had with Ann Lee at a workshop in Aspen last year she and I co-organized with Konstantin Matchev. The modified LF2I algorithm is called amortized likelihood-free frequentist inference (ALFFI).

Algorithm 2 Amortized likelihood-free 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 θ_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 $\mathbb{C}(\lambda_{\text{obs}}|\theta)$.

Outline

The power of LF2I and ALFFI is that knowledge of the likelihood is not needed and the method works for samples of all sizes. However, it is useful to have simple benchmark models 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[6].

For the test statistic, we use the function

$$\lambda = \sum_{i=1}^{N} \left(\frac{x_i - \mu(z_i, \theta)}{\sigma_i} \right)^2, \tag{7}$$

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.

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

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

$$\mathcal{P} = -\frac{1}{3} n a^n \Omega(a), \tag{8}$$

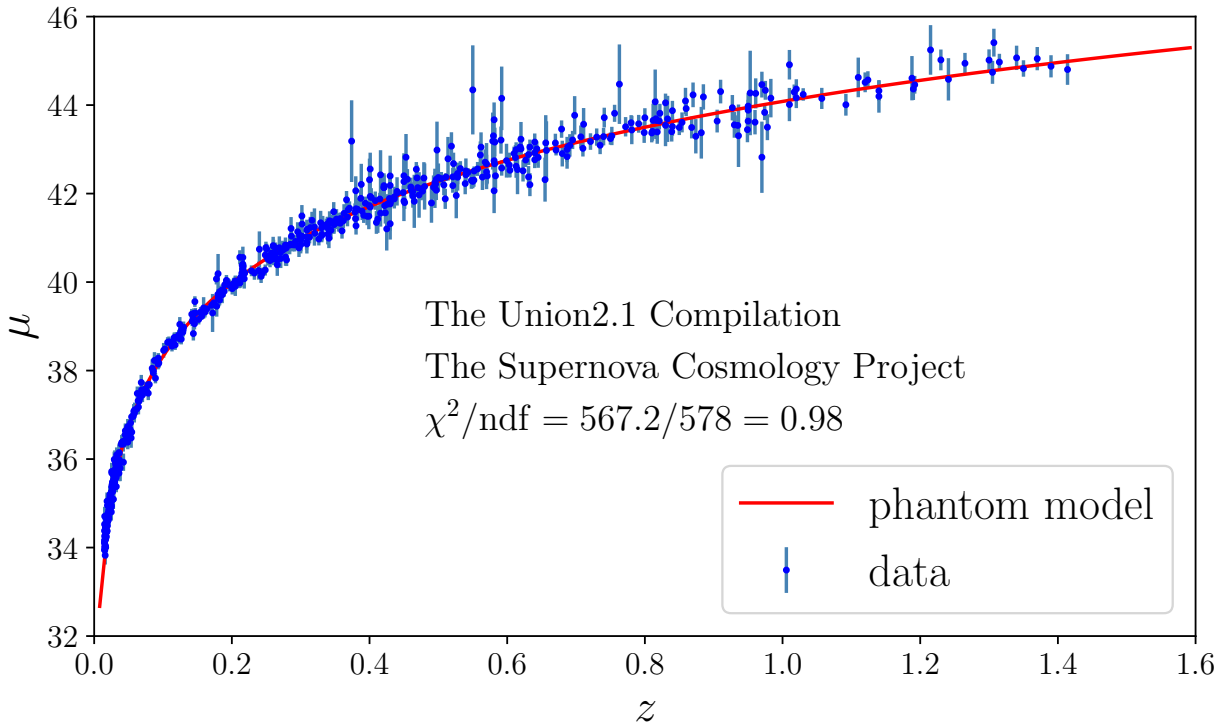
where n is a free parameter, and $a(t)$, $\Omega(a)$, and $\mathcal{P}$ are the dimensionless universal scale factor, the dimensionless energy density, and the dimensionless pressure, respectively, and t is the time since the Big Bang.

This equation of state yields the energy density

$$\Omega(a) = \exp(a^n - 1) \,/\, a^3. \tag{9}$$

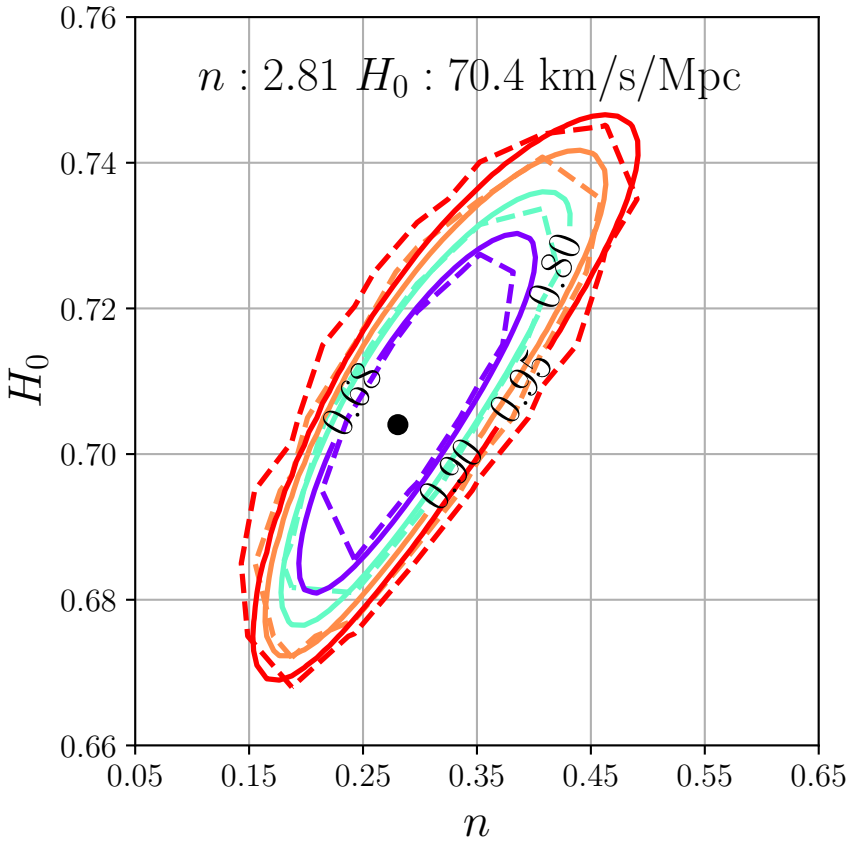
About the only virtue of this model is that it has only two parameters, the other being the Hubble constant $\mathcal{H}_0$ (not to be confused with an hypothesis), and the model can be exactly integrated.

When the cosmological model is fitted to the Type 1a data by simply minimizing the $\lambda\lambda$ (using, for example, iminuit), we find the following excellent fit.



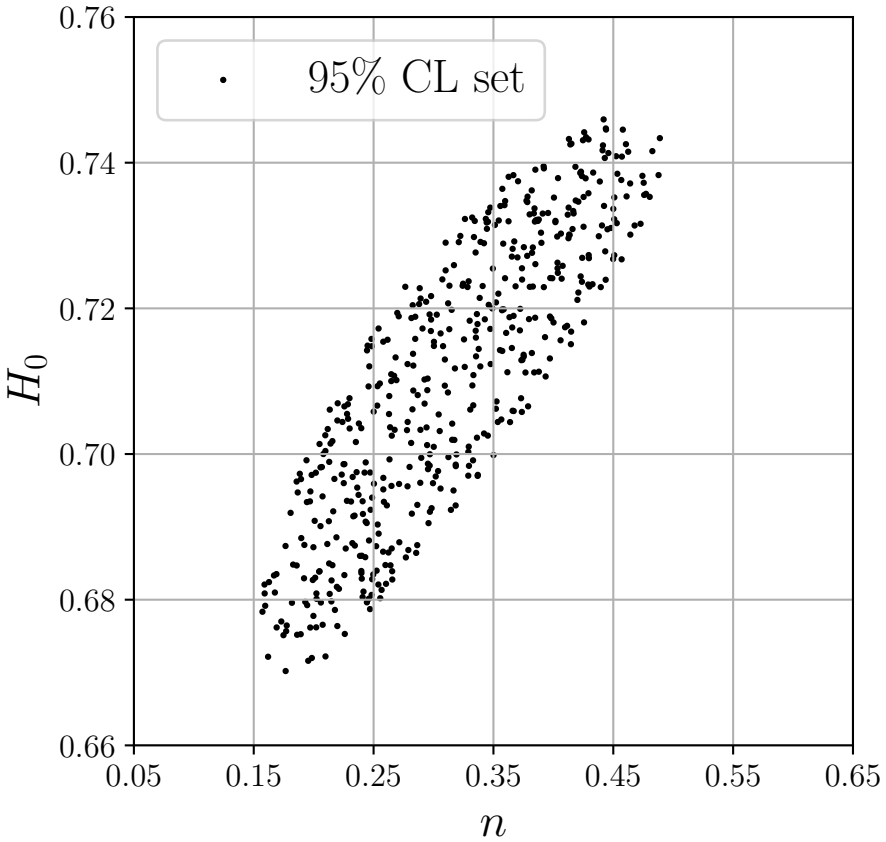
By the way, the model predicts that the universe will self-destruct in a Big Rip at about 1.4 times its current age!

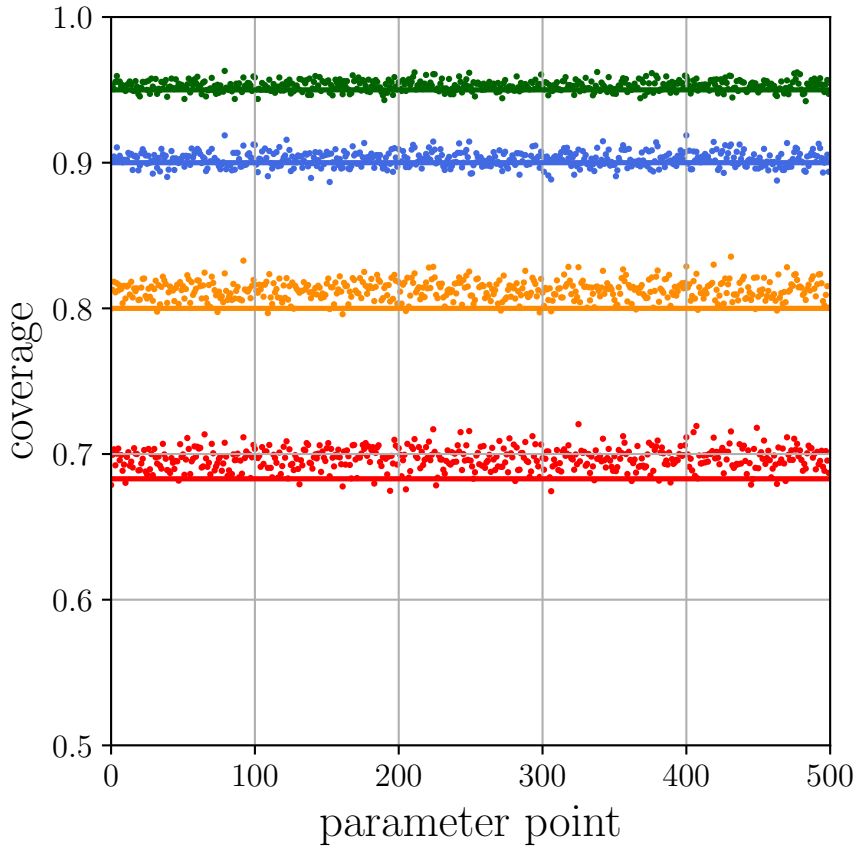
We approximate $\mathbb{P}(\lambda \leq \lambda_{\text{obs}})$ using a 5-layer fully-connected feed-forward neural network, with 20 nodes per layer, a single output, and `ReLU` non-linearities. The confidence sets are shown in the figure below.



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.





Our second example is the neutron-antineutron oscillations search we discussed earlier. We use the following test statistic

$$\lambda(D,\theta) = -2\log\left[\frac{p(D|\mu,\nu)}{p(D|\hat{\mu},\hat{\nu})}\right], \tag{10}$$

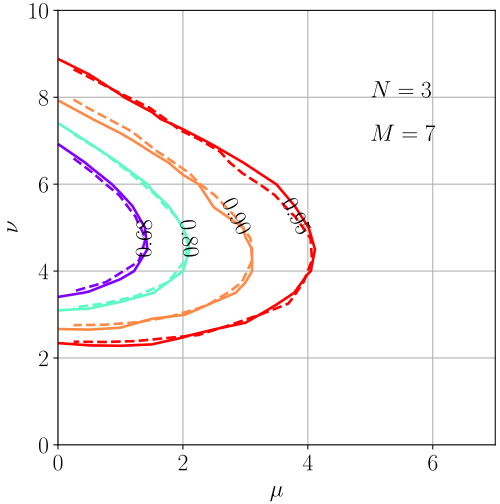
where $\hat{\mu}$ and $\hat{\nu}$ are the best-fit values of the parameters. Since $\mu \geq 0$, we take the estimate of the mean signal to be

$$\hat{\mu} = \begin{cases} N - M & \text{if} \quad N > M \\ 0 & \text{otherwise,} \end{cases}, \tag{11}$$

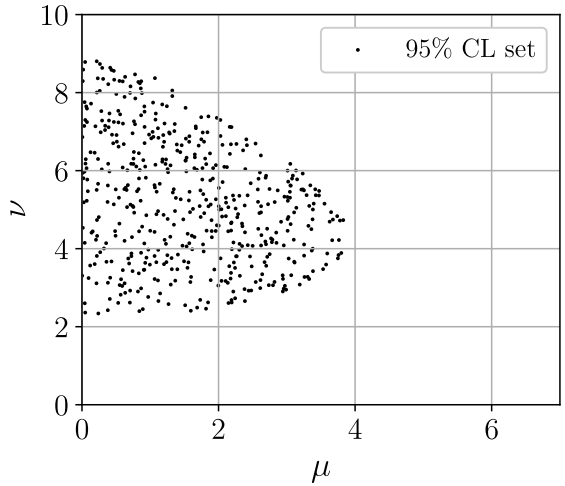
which explicitly violates one of the regularity conditions for the validity of Wilks' theorem, namely, that estimates must lie within the interior of the parameter space. For the estimate of the mean background, we take

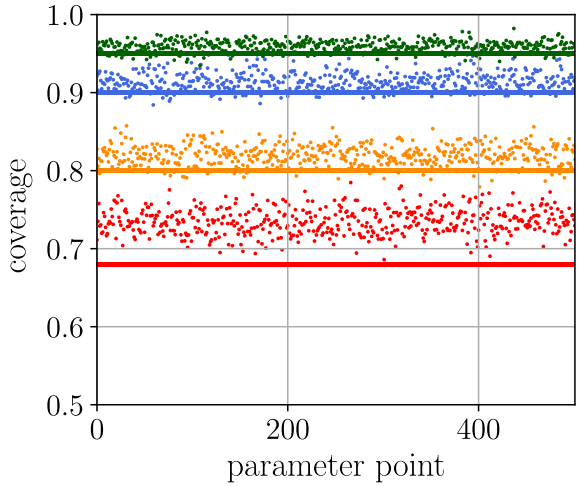
$$\hat{\nu} = \begin{cases} M & \text{if} \quad \hat{\mu} = N - M \\ (M+N)/2 & \text{otherwise.} \end{cases} \tag{12}$$

A similar neural network model is trained for the ON/OFF example and 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.

We end with an example in which the likelihood function is intractable and, therefore, where the full power of LF2I and ALFFI is needed.

The susceptible-infected-recovered (SIR) model is applied to a classic data set from a flu outbreak more than a century ago at an English Boarding School. In this model, individuals in the susceptible class, S, can migrate to the infected class, I, and from I to the recovered class, R.

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

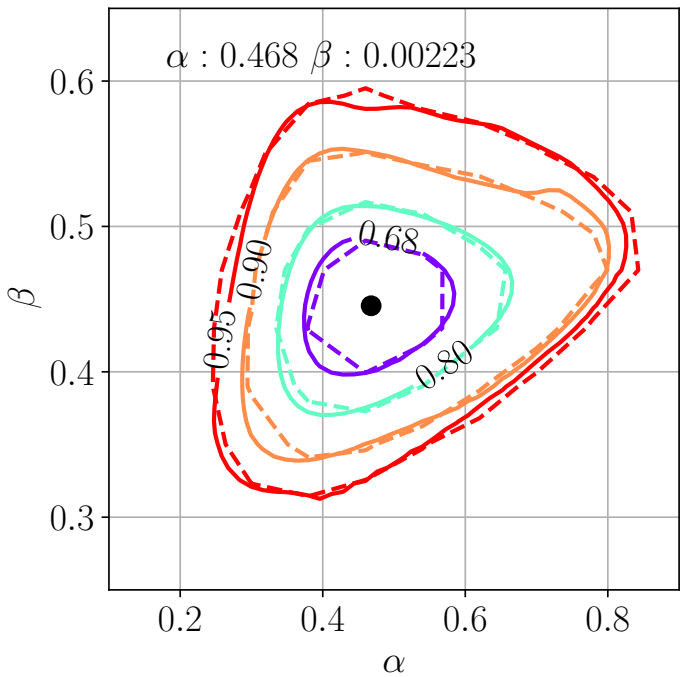
$$\begin{aligned} \frac{dS}{dt} &= -\beta SI, \\ \frac{dI}{dt} &= -\alpha I + \beta SI, \\ \frac{dR}{dt} &= \alpha I, \end{aligned} \tag{13}$$

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

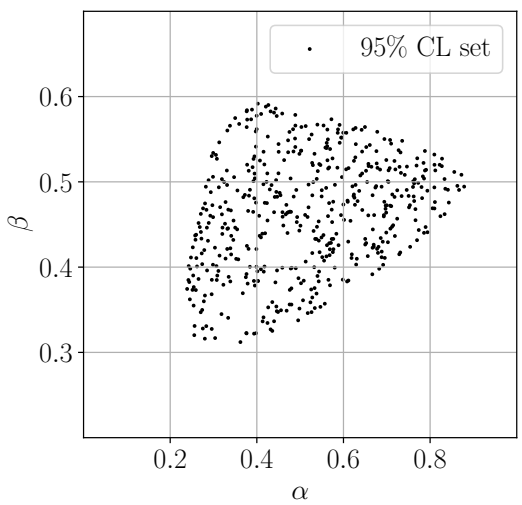
We choose a test statistic proportional to

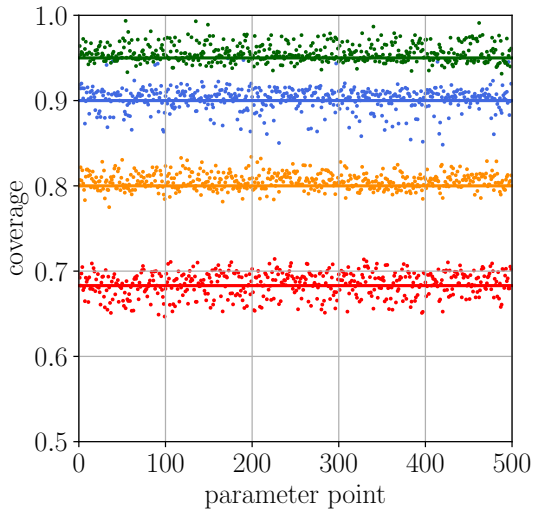
$$\lambda(D,\theta) = \sqrt{\sum_{n=1}^{N} \frac{(x_n - I_n(\theta))^2}{I_n(\theta)}}, \tag{14}$$

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.



We again train a relatively simple neural network to approximate $\mathbb{P}(\lambda \leq \lambda_{\text{obs}}|\theta)$ and use it 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 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.

Outline

- 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 makes it possible to both construct confidence sets and check their coverage explicitly using the same trained neural network model.
- The three simple examples illustrate the potential of likelihood-free frequentist inference, but, as always, more work is needed.
- The LF2I approach contains methods to compute confidence sets for subsets of the parameters, but, alas, without frequentist guarantees for small samples.

Thank you!

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

I wish to thank Prof. Ann Lee for fruitful discussions at the Aspen Physics Center during our workshop in summer 2022.

- 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)