Machine Learning: Science and Technology

PAPER • OPEN ACCESS

Amortized simulation-based frequentist inference for tractable and intractable likelihoods

To cite this article: Ali Al Kadhim *et al* 2024 *Mach. Learn.: Sci. Technol.* **5** 015020

View the article online for updates and enhancements.

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 015020 https://doi.org/10.1088/2632-2153/ad218e

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
30 October 2023

REVISED
15 January 2024

ACCEPTED FOR PUBLICATION
23 January 2024

PUBLISHED
2 February 2024

PAPER

Amortized simulation-based frequentist inference for tractable and intractable likelihoods

Ali Al Kadhim[1,*], **Harrison B Prosper**[1] **and Olivia F Prosper**[2]

[1] Department of Physics, Florida State University, Tallahassee, FL 32306-4350, United States of America
[2] Department of Mathematics, University of Tennessee, Knoxville, TN 37996-1320, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: aa18dg@fsu.edu

Keywords: confidence sets, likelihood-free inference, simulation-based inference, frequentist inference, likelihood ratio test, armotized, SBI

Abstract

High-fidelity simulators that connect theoretical models with observations are indispensable tools in many sciences. If the likelihood is known, inference can proceed using standard techniques. However, when the likelihood is intractable or unknown, a simulator makes it possible to infer the parameters of a theoretical model directly from real and simulated observations when coupled with machine learning. We introduce an extension of the recently proposed likelihood-free frequentist inference (`LF2I`) approach that makes it possible to construct confidence sets with the p-value function and to use the same function to check the coverage explicitly at any given parameter point. Like `LF2I`, this extension yields provably valid confidence sets in parameter inference problems for which a high-fidelity simulator is available. The utility of our algorithm is illustrated by applying it to three pedagogically interesting examples: the first is from cosmology, the second from high-energy physics and astronomy, both with tractable likelihoods, while the third, with an intractable likelihood, is from epidemiology[3].

1. Introduction

Simulation-based, or likelihood-free, inference is now ubiquitous in the sciences (see, for example, [1, 2]). Recently, Dalmasso *et al* introduced likelihood-free frequentist inference [3] (`LF2I`), a simulation-based method featuring provable frequentist [4] guarantees. Let $\mathcal{D} = \{X_i \mid i = 1, \ldots, N\}$ be the set of *observable* data sampled from a simulator F_θ and $D = \{x_i \mid i = 1, \ldots, N\}$ be the set of *observed* data.

Consider a large, in principle infinite, collection of data-driven statements of the form $\theta \in R(D)$ with parameter θ a point in the parameter space of a theoretical model, and $R(D)$ is a data-dependent subset of that space. Given a function of the data $\lambda(\mathcal{D};\theta)$, called a *test statistic*, the `LF2I` approach constructs data-driven statements based on the test statistic that are either true or false with the guarantee that over a large collection of such statements, a minimum fraction τ of them will be true. The fraction τ is called the *confidence level*, while the fraction $p \geqslant \tau$ of true statements, which may vary over the parameter space, is called the *coverage probability*, or coverage for short. The interesting aspect of `LF2I` is that the guarantee, $p \geqslant \tau$, holds for any data sample size N. Parameter subsets that satisfy this condition are called *confidence sets* $R(D)$. Ideally, the function $\lambda(\mathcal{D};\theta)$ compresses the data such that all relevant information about the parameter θ is preserved. We define the observed test statistic, λ_D, to be the test statistic evaluated at the observed data, $\lambda_D \equiv \lambda(\mathcal{D} = D;\theta)$.

Confidence sets and classical hypothesis tests are closely related. A classical hypothesis test [4] is a procedure for deciding between two hypotheses: a null hypothesis H_0 and an alternative hypothesis H_1. For example, we may wish to perform the following test:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0. \tag{1}$$

[3] Code to reproduce all of our results is available on https://github.com/AliAlkadhim/ALFFI.

The hypothesis test in equation (1) is equivalent to

$$H_0 : \theta \in \Theta_0 \text{versus } H_1 : \theta \in \Theta_1, \tag{2}$$

where $\Theta_0 \cap \Theta_1 = \emptyset$, and Θ_0 could be a single parameter point (which defines a *simple hypothesis*) or more than one point (thereby defining a *composite hypothesis*).

When the two hypotheses are simple the likelihood ratio test statistic is known to be optimal [5], while experience indicates that the likelihood ratio, or a function thereof, works well even when one or both hypotheses are composite [6]. The null hypothesis will either be rejected or will fail to be rejected depending on the value of the test statistic evaluated on a set of observed data $\mathcal{D} = D$. If the null hypothesis is rejected, we may choose to accept the alternative.

Assuming that large values of $\lambda = \lambda(\mathcal{D};\theta)$ cast doubt on the null hypothesis $\theta = \theta_0$, the latter is rejected if the p-value, $\mathbb{P}(\lambda > \lambda_D|\theta)$, is less than a given threshold α, called the *size* or *significance level* of the test[4] [7]. For example, if the test statistic is a χ^2 function, large values of χ^2 disfavor the hypothesis $\theta = \theta_0$. If the parameter point θ_0 is not rejected it is added to the confidence set $R(D)$. A confidence set is, therefore, the set of parameter points that have not been rejected at level α. The key idea of LF2I is to approximate the p-value with a neural network in such a way that the resulting confidence sets satisfy $p \geqslant \tau$ or, more realistically, $p \approx \tau$, given that the neural network provides an approximation to the p-value.

LF2I provides two methods to construct confidence sets: one requires approximating the value of λ_D, called the critical value C_α, for which the p-value $= \mathbb{P}(\lambda > \lambda_D|\theta)$ is equal to α,

$$\widehat{R}(D) = \left\{\theta_0 \in \Theta \mid \lambda(D;\theta_0) \geqslant \widehat{C}_{\alpha,\theta_0}\right\}, \tag{3}$$

where $\widehat{C}_{\alpha,\theta_0}$ is the estimated critical value at level α for the hypothesis $\theta = \theta_0$, and a second method that requires an approximation, $\widehat{p}(D;\theta_0)$, of the p-value,

$$\widehat{R}(D) = \left\{\theta_0 \in \Theta \mid \widehat{p}(D;\theta_0) > \alpha\right\}. \tag{4}$$

The algorithm proposed in this paper generalizes the second method, i.e. equation (4) so that the coverage can be explicitly checked using the p-value $\widehat{p}(D;\theta_0)$, while the explicit checking of the coverage in LF2I can only be done via equation (3), and then only for a particular value of α.

In practice, we choose to approximate the cumulative distribution function (cdf), $\mathbb{P}(\lambda \leqslant \lambda_D|\theta)$, with a neural network, in contrast to LF2I where the p-value, $\mathbb{P}(\lambda > \lambda_D|\theta)$, is approximated. Our key idea is to make the neural network a function of both the parameter point θ and the "observed" test statistic $\lambda_D \equiv \lambda(D;\theta)$. This simple extension makes it possible to apply the approximated cdf to any data set that is sampled from the same underlying distribution as the observed data D. Moreover, the cost of approximating the cdf is *amortized* over its subsequent use in constructing confidence sets and its use in explicitly checking the coverage of these sets without the need to retrain the network. To distinguish the modified LF2I from the original, we refer to the former as amortized likelihood-free frequentist inference (ALFFI). The ALFFI approach is illustrated in three pedagogical examples chosen from diverse areas of the sciences. The first two examples feature likelihoods that are tractable, while for the third example the real power of LF2I and ALFFI is illustrated with a problem in which the likelihood is intractable.

The paper is organized as follows. In section 2, we describe the ALFFI approach. This is followed, in section 3, with the three examples. Section 3.1 uses ALFFI to infer the parameters of a simple cosmological model that is fitted to Type 1a supernova data, while section 3.2 applies ALFFI to the prototypical signal/background problem in high-energy physics, which in astronomy is known as the On/Off problem [8]. For both of these problems, the likelihood is tractable. Section 3.3 illustrates the application of ALFFI to a well-known epidemiological model, which, though simple, has an intractable likelihood. The paper ends with a brief discussion in section 4 and our conclusions in section 5.

2. ALFFI

2.1. From cdf to confidence sets

A classical hypothesis test is designed to have power in distinguishing between the null and the alternative hypothesis. Consider the α-level hypothesis $\theta = \theta_0$, as in equation (1). This hypothesis is to be rejected if $\mathbb{P}(\lambda > \lambda_D \mid \theta_0) < \alpha$. The corollary is that θ_0 is *not* to be rejected if $\mathbb{P}(\lambda > \lambda_D \mid \theta_0) \geqslant \alpha$, that is, if $\mathbb{P}(\lambda \leqslant \lambda_D \mid \theta_0) \leqslant 1 - \alpha \equiv \tau$. The set of points θ_0 that have not been rejected at level α, and therefore remain

[4] α is also the threshold at which one is willing to commit a Type I error, the probability of erroneously rejecting the null hypothesis when it is in fact true.

as potentially viable hypotheses for the true value of θ, is by definition a confidence set $R(D)$ at $100\tau\%$ confidence level (CL). The boundary of the confidence set $R(D)$ is determined by the equation

$$\mathbb{P}(\lambda \leqslant \lambda_D \mid \theta_0) = \tau. \tag{5}$$

By construction, the coverage probability of such sets is

$$\mathbb{P}(\theta \in R(\mathcal{D}) \mid \theta) \geqslant \tau. \tag{6}$$

Note that $R(\mathcal{D})$ is a *random* set. Under repeated observations, the frequentist principle requires that the relative frequency with which these sets include the true value of θ never falls below the stated CL $\tau = 1 - \alpha$ regardless of the true value of θ. To the degree that the probabilities $\mathbb{P}(\lambda > \lambda_D \mid \theta)$ and $\mathbb{P}(\lambda \leqslant \lambda_D \mid \theta)$ are accurately modeled, the LF2I and ALFFI approaches yield random sets that satisfy the frequentist principle.

2.2. Data preparation for ALFFI

The LF2I and ALFFI approaches are applicable to any test statistic $\lambda(\mathcal{D};\theta)$ that is monotonic in the following sense: the test statistic is constructed so that a particular direction in its 1-dimensional space corresponds to hypotheses that are increasingly disfavored. In ALFFI, we consider test statistics for which large values correspond to disfavored hypotheses, or equivalently, small values of the p-value $= \mathbb{P}(\lambda > \lambda_D | \theta = \theta_0)$, where $\lambda_D \equiv \lambda(\mathcal{D} = D; \theta_0)$ is the observed value of the test statistic for the specified hypothesis. ALFFI (see appendix A) approximates the probability

$$\mathbb{P}(\lambda \leqslant \lambda_D \mid \theta) = \int^{\lambda_D} \mathrm{d}Y \int \mathrm{d}\mathcal{D}\, \delta\left(Y - \lambda(\mathcal{D}, \theta)\right) p(\mathcal{D} \mid \theta), \tag{7}$$

where $\delta(.)$ is the Dirac delta function [9] and $p(\mathcal{D}|\theta)$ is a statistical model, which may or may not be tractable. However, as in the LF2I approach [3], it is assumed that one has access to a large collection $\mathbb{S}$ of simulated pairs $(\mathcal{D}, \theta) \in \mathbb{S}$ where for every point θ sampled from any convenient prior π_θ a *single* instance of a data set $\mathcal{D}$ is simulated with the same characteristics, including sample size, as the real data D. In contrast to LF2I, in ALFFI a second collection of data sets $\mathbb{S}'$ is created from $\mathbb{S}$ by randomizing the order of the data sets $\{\mathcal{D}\}$ to form new pairs $(\mathcal{D}', \theta) \in \mathbb{S}'$ in which the parameters and the data sets $\mathcal{D}'$ are statistically independent. The data sets in $\mathbb{S}'$ serve as instances of 'observed' data sets.

For every parameter θ in $\mathbb{S}$, we compute two values of the test statistic, namely, $\lambda(\mathcal{D}, \theta)$ and $\lambda_D = \lambda(\mathcal{D}', \theta)$, as well as the discrete variable Z which is unity if $\lambda(\mathcal{D}, \theta) \leqslant \lambda(\mathcal{D}', \theta)$ and zero otherwise. This procedure results in a large collection of triplets of size B, $\mathcal{T} = \{(Z_i, \lambda_{D,i}, \theta_i)\}_{i=1}^{B}$, which constitute the training data. In LF2I, the observed test statistic under the null that $\theta = \theta_0$—the second component of the triplet—is computed using a *fixed* data set $\mathcal{D}' = D$, namely, the one actually observed, while ALFFI uses the data sets $\mathcal{D}' \in \mathbb{S}'$.

2.3. Approximating the cdf

The cdf $\mathbb{P}(\lambda \leqslant \lambda_D | \theta)$ is the expectation value $\mathbb{E}(Z | \lambda_D, \theta)$ of the discrete variable Z, a fact that suggests a straightforward way to approximate the cdf for a *fixed* data set D: histogram the parameter points θ, thereby yielding the histogram $\mathbb{H}_1$, and using the same bins as $\mathbb{H}_1$ histogram the parameter points again, but this time weighted by Z, yielding the histogram $\mathbb{H}_Z$. The ratio $\mathbb{H}_Z / \mathbb{H}_1$ provides a piece-wise-constant approximation of $\mathbb{E}(Z | \lambda_D, \theta)$.

Following LF2I, a smooth approximation of $\mathbb{E}(Z | \lambda_D, \theta)$ is created with a deep neural network (DNN) trained (that is, fitted) by minimizing the empirical risk

$$E(\boldsymbol{\omega}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(t_i, f_i), \tag{8}$$

where $\mathcal{L}(t, f)$ is a loss function, $f(\mathbf{x}; \boldsymbol{\omega})$ is a DNN with inputs $\mathbf{x}$ and free parameters $\boldsymbol{\omega}$, and t denotes known *targets*. The empirical risk or average loss, equation (8), is a Monte Carlo approximation of the risk functional

$$E[f] = \int \int \mathcal{L}(t, f)\, p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t, \tag{9}$$

where $p(\mathbf{x}, t) = p(t|\mathbf{x})\, p(\mathbf{x})$ is the (typically unknown) probability distribution of the data $\mathbf{x}, t$. From the calculus of variations, the function f that minimizes equation (9) is the solution of

$$\int \frac{\partial \mathcal{L}}{\partial f} p(t|\mathbf{x})\, \mathrm{d}t = 0, \tag{10}$$

Table 1. *Cosmological Model*: $x_i \pm \sigma_i$ and z_i are the measured distance moduli and redshifts, respectively, while $\mathcal{H}_0$ and n are the model parameters. *Signal/Background*: N and M are the observed counts for signal and background, respectively, while n and m are the expected signal and background counts, respectively. μ and ν are the unknown signal mean (parameter of interest) and unknown background mean (nuisance parameter), respectively. *SIR Model*: $\{x_i\}$ are the 13 observed counts of infected children on the 13 days of observation, and α and β are the model parameters. CTMC is a continuous time Markov chain model of the epidemic.

Example	Observed data (D)	θ	Priors π_θ	$\mathcal{D} \sim F_\theta$	Tractable $\mathcal{L}$?	λ
Cosmological Model	$x_i \pm \sigma_i, z_i$	$\mathcal{H}_0, n$	$\frac{\mathcal{H}_0}{100} \sim$ Unif$(0.66, 0.76)$, $n \sim$ Unif$(0.05, 0.65)$	$x_i \sim \mathcal{N}(\mu, \sigma_i)$	Yes	equation (16)
Signal/ background	$N = 3$, $M = 7$	μ, ν	$\mu \sim$ Unif$(0, 20)$, $\nu \sim$ Unif$(0, 20)$	$n \sim$ Poisson$(\theta + \nu)$, $m \sim$ Poisson(ν)	Yes	equation (21)
SIR Model	x_i	α, β	$\alpha \sim$ Unif$(0.1, 0.9)$, $\frac{\beta \times 10^3}{5} \sim$ Unif$(0.25, 0.65)$	$\{x_i\} \sim$ CTMC(θ)	No	equation (25)

assuming that $p(\mathbf{x}) > 0 \, \forall \mathbf{x}$. In the examples below, and following `LF2I`, we use the quadratic loss $\mathcal{L}(t, f) = (t - f)^2$, which, using equation (10), leads to the well-known result [10, 11]

$$f(\mathbf{x}; \boldsymbol{\omega}^*) = \int t\, p\,(t|\mathbf{x})\, \mathrm{d}t \equiv \mathbb{E}_t\,[t \mid \mathbf{x}]\,, \tag{11}$$

where $\boldsymbol{\omega}^*$ are the best-fit parameters of the neural network model. Setting $\mathbf{x} = \{\lambda_D, \theta\}$ and the targets $t = Z$ in equation (11) yields

$$\begin{aligned} f(\lambda_D, \theta; \boldsymbol{\omega}^*) &\approx \mathbb{E}\,[Z \mid \lambda_D, \theta]\,, \\ &= \mathbb{P}\,(\lambda \leqslant \lambda_D \mid \theta)\,, \end{aligned} \tag{12}$$

that is, it yields the quantity that we wish to approximate.

One of the key virtues of the `LF2I` and `ALFFI` approaches, as is evident in the result in equation (12), is that the neural network f is conditioned on θ, which implies that it is independent of the prior π_θ from which the parameter points are sampled. The form of the prior affects only the accuracy of the approximation: the accuracy of the approximation will be greatest where the density of the prior is greatest.

3. Results

The `ALFFI` approach is illustrated in the following three diverse examples: the first is from cosmology, the second from high-energy physics and astronomy, and the third from epidemiology. Our choice of the particular problem in each field highlights typical statistical inference problems that are encountered in each field. We demonstrate that the `ALFFI` method yields valid[5] multi-parameter confidence sets for all three examples, both in cases where the likelihood is tractable (the first two examples), and where the likelihood is intractable (the third example). We also demonstrate the use of different test statistics, demonstrating the compatibility with binned and un-binned analyses. Table 1 summarizes key attributes of each example.

3.1. Example 1: cosmological model

In the late 1990 s, fits of cosmological models to Type 1a supernova data led to the conclusion that the expansion of the Universe is accelerating [12, 13]. The fits then, as now, were performed using tractable likelihoods, typically a multivariate normal. In this example, we fit a cosmological model to the Union 2.1 data compilation of the Supernova Cosmology Project [14] via maximum likelihood and also with `ALFFI`. The Union 2.1 data set comprises measured distance moduli, $x \pm \sigma$, and redshifts, z, for 580 Type 1a supernovae. Given the size of the data sample, it is expected that accurate confidence sets for the cosmological parameters can be constructed using standard methods such as maximum likelihood. `ALFFI` is therefore not needed for this problem; it is simply used to showcase the algorithm.

Our cosmological model is defined by the equation of state

$$\mathcal{P} = -n a^n \Omega / 3, \tag{13}$$

[5] By valid we mean that the confidence set has the nominal type I error or CL.

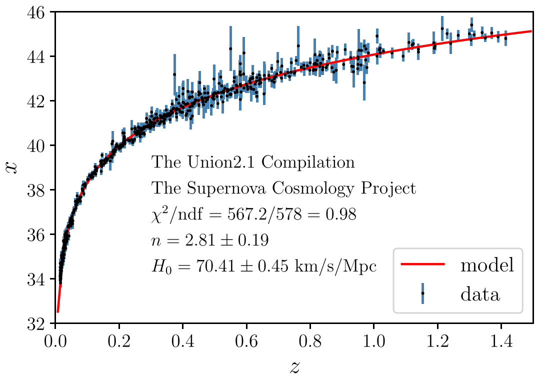

Figure 1. Example 1 (cosmology): maximum likelihood fit of 2-parameter cosmological model to Union 2.1 Type 1a data set. Note the excellent fit, which with a χ^2 per number of degrees of freedom (ndf) of 0.98 is on par with that of the ΛCDM model.

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. Together with the Hubble constant $\mathcal{H}_0$ this is a 2-parameter problem. (Details of the model are given in appendix B). If the small correlations between the 580 Typa 1a data points are neglected, the likelihood for these data is a diagonal multivariate normal. Maximizing this likelihood with respect to its parameters is equivalent to minimizing the function

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{x_i - \mu(z_i, \theta)}{\sigma_i} \right)^2, \tag{14}$$

where $\mu(z, \theta)$, the *distance modulus*[15]—an astronomical measure of distance, is given by

$$\mu(z, \theta) = 5\log_{10}\left[(1+z)\sin\left(\sqrt{-\Omega_K}\,u(z,\theta)\right)/\sqrt{-\Omega_K}\right] + 5\log_{10}\left(c/\mathcal{H}_0/10^{-5}\text{Mpc}\right). \tag{15}$$

The quantity c is the speed of light in vacuum in km/s, Ω_K is the curvature parameter, which we set to zero, and

$$u(z, \theta) = \int_{1/(1+z)}^{1} \frac{da}{a^2\sqrt{\Omega(a)}}, = 2^{\frac{1}{2n}}\left[\gamma\left(\frac{1}{2n}, \frac{1}{2}\right) - \gamma\left(\frac{1}{2n}, \frac{(1+z)^{-n}}{2}\right)\right]\sqrt{e}/n,$$

is a dimensionless function. When the model is fitted to the Union 2.1 data set by minimizing equation (14) an excellent fit is obtained, as shown in figure 1.

We now apply ALFFI to the same problem using the test statistic

$$\lambda(D, \theta) = \sqrt{\frac{\chi^2}{N}}, \tag{16}$$

where χ^2 is defined in equation (14), and satisfies the requirement by ALFFI that large values of the test statistic cast doubt on the null hypothesis. The form of the test statistic is chosen so that it is $\mathcal{O}(1)$ as it is an input to the neural network. The boundary of the associated 100τ% confidence set is given by equation (5), which requires a good approximation to the cdf $\mathbb{P}(\lambda \leqslant \lambda_D|\theta)$. The latter is approximated in two ways: with histograms and with a deep neural network (DNN) as described in section 2.3. The 2D histograms have 10 bins in both dimensions n and $\mathcal{H}_0$ with the parameter n scaled down by a factor 10 and $\mathcal{H}_0$ scaled down by a factor 100, so that both input parameters are of $\mathcal{O}(1)$.

The DNN is 1781-parameter fully-connected feed-forward neural network, with 3 input features, $\boldsymbol{x} = \{\lambda_D, n, \mathcal{H}_0\}$, 5 hidden layers with 20 nodes each, and a single output. We use a ReLU [16] activation function, the output node is a sigmoid that constrains the output to lie within the unit interval, and the DNN is trained as described in section 2 using PyTorch [17].

We use a batch size $K = 50 \ll N$ randomly sampled from $N = 250\,000$ data sets of 580 simulated distance moduli per data set, sampled at the same redshifts as the observed data. The Adam [18] optimizer with a fixed learning rate of 10^{-3} is used to train the network. As the training proceeds, the network with the smallest average loss is saved. The average loss is computed using a validation data set of size 5000 that is not used by the optimizer. We employ the following early stopping [19] criterion: the training stops if after

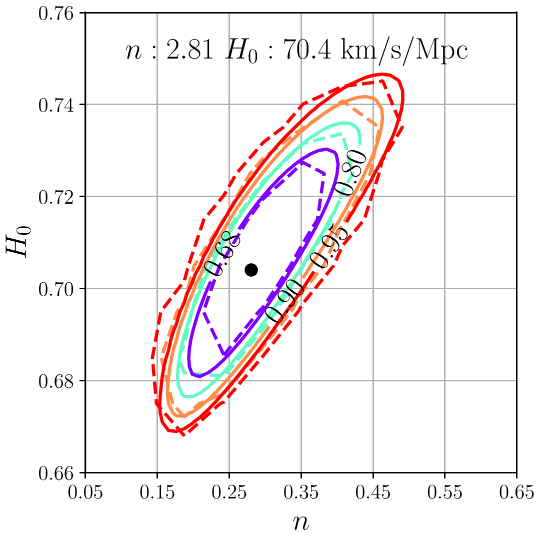

Figure 2. Example 1 (cosmology): confidence sets $R(D)$ for $\tau = 0.68, 0.80, 0.90$, and 0.95. (*dashed lines*) Boundaries of confidence sets, $R(D)$, defined by $\mathbb{P}(\lambda \leqslant \lambda_D|\theta) = \tau$ computed using the histogram-based approximation of the cdf. (*solid lines*) Boundaries of confidence sets computed using the DNN-based approximation of the cdf. (*black dot*) Location of the minimum of $\mathbb{P}(\lambda \leqslant \lambda_D|\theta)$, computed with the DNN approximation, which is taken to be the best-fit point and which agrees with the maximum likelihood results in figure 1.

50 000 steps no network is found with a smaller validation loss than the saved network. The best network as defined by this training protocol is found after about 100 000 iterations. If one defines an epoch to be N/K iterations, which for this example is 5000, then 100 000 iterations corresponds to 20 epochs of training. The simulated data sets of 580 distance moduli are sampled from 580 independent normal distributions using the standard deviations taken from the observed data.

Approximating $\mathbb{P}(\lambda \leqslant \lambda_D|\theta)$ using histograms and using ALFFI leads to the confidence sets shown in figure 2.

The best-fit values of the cosmological parameters $\theta = \{n, \mathcal{H}_0\}$ are taken to be the location of the minimum of $\mathbb{P}(\lambda \leqslant \lambda_D|\theta)$, which as indicated in figure 2 agrees with the values obtained from the likelihood fit.

In the LF2I approach, the coverage over the parameter space is checked by modeling the coverage probability with another neural network as a function of the parameters of the theoretical model. There are pros and cons to that approach. It is certainly convenient to have a functional approximation of the coverage probability as a function of the parameter space point because one can then estimate the coverage at any given point, not only at points for which there are sufficient simulated data. Unfortunately, however, as is true of most machine learning models, a reliable estimate of the accuracy of the trained machine learning model is not available. In ALFFI, the coverage is checked explicitly by direct enumeration at all points for which there are sufficient data. Since the problem consists of counting how often a particular statement is true, the problem is binomial; therefore, a reliable estimate of the accuracy of the coverage calculation is easy to compute. On the other hand, the coverage is available only at the parameter points for which there are sufficient simulated data. In this example, for a given parameter point, $T = 4,000$ sets of 580 Type1a supernovae data are simulated. For each data set, a test statistic, λ_D, is computed. If $\mathbb{P}(\lambda \leqslant \lambda_D|\theta) \leqslant \tau$ then, by definition, θ lies within the confidence set associated with λ_D. If S is the number of times this statement is true over the collection of T simulated data sets, then the coverage probability is $p \pm \sqrt{p(1-p)/T}$, where $p = S/T$. For continuous probability distributions and exact confidence sets, the coverage probability p is exactly equal to the CL τ. Therefore, if the confidence sets produced by ALFFI are accurate then we should find $p \approx \tau$ given that we are making an approximation. This calculation is performed at 500 randomly sampled points within the 95% CL set associated with the observed Type 1a data, as shown in the left panel of figure 3.

The right panel shows the coverage probabilities calculated for the 500 randomly sampled parameter points, shown in the left panel, compared with the desired $1 - \alpha = \tau$ CLs, depicted by the solid horizontal lines. Since the coverage probabilities are greater or equal to the CLs, we conclude that the coverage of the confidence sets computed using ALFFI satisfy the coverage condition in equation (6).

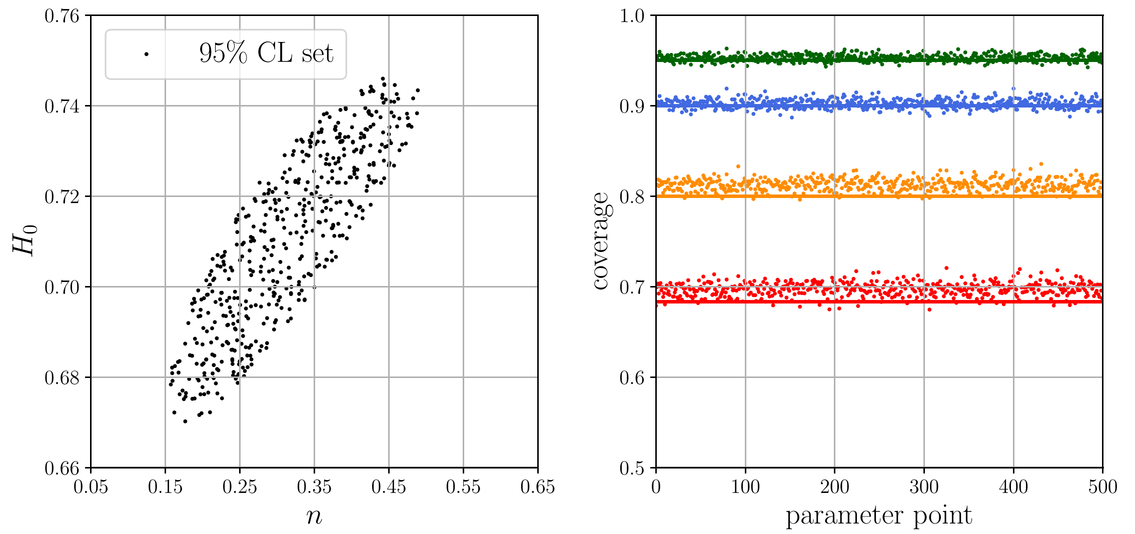

Figure 3. Example 1 (cosmology): (*left*) 500 points, from the 95% CL confidence set for the Type 1a data, at which the coverage has been computed. (*right*) the dots are the coverage probabilities for each of the 500 points in the cosmological model $n, \mathcal{H}_0$ parameter space and the horizontal lines are the values of the confidence levels τ.

3.2. Example 2: signal/background or ON/OFF model

The cosmological example served to illustrate the `ALFFI` algorithm, but, as noted, `ALFFI` is not really needed because the cosmological data are numerous, the likelihood function is tractable, and parameter estimation via maximum likelihood works well. Our second example addresses the signal/background problem in high-energy physics (see, for example, [20]), which in astronomy is referred to as the On/Off problem [8]. We choose an example in which the likelihood is tractable but the data are sparse and, consequently, asymptotic methods may not be reliable [7]. We demonstrate that the ALFFI method yields valid confidence sets even when the regularity conditions that underpin Wilks' theorem [21] and its variants [6] are violated.

The signal/background problem in high-energy physics and astronomy is as follows. An observation is made, for given period of time, which consists of counting N events: typically, photons in astronomy and particle collisions in high-energy physics. The count is potentially a sum of counts from signal and background sources. A second independent observation is made for the same duration (in the simplest case) where by design the background has the same characteristics as in the first observation but no signal is present. The second observation yields a count M. It is generally assumed that the likelihood function for the data $D = \{N, M\}$ is the product of two Poisson distributions,

$$\mathcal{L}(D;\mu,\nu) = \frac{(\mu+\nu)^N \exp(-(\mu+\nu))}{N!} \frac{\nu^M \exp(-\nu)}{M!}, \tag{17}$$

where μ and ν are the mean signal and background counts, respectively. In the signal/background problem the parameter of interest is μ, while ν is a *nuisance parameter*. We shall comment on how one might deal with such parameters in the discussion.

Our specific example is from the first experiment to search for neutron-antineutron oscillations using free neutrons [22], which took place in the 1980 s at the Institut Laue-Langevin (ILL) in Grenoble, France. Neutron-antineutron ($n\bar{n}$) oscillations are predicted by many proposed theories of physics beyond the Standard Model of particle physics [23]. For our purposes it suffices to note that if $n\bar{n}$ oscillations can occur then a pure neutron state, when observed at time $t \ll$ than the mean neutron lifetime, will be observed with probability

$$\begin{aligned} P_t &= \left(\frac{\epsilon^2}{\epsilon^2+\Delta E^2}\right)\sin^2\left(\left(\epsilon^2+\Delta E^2\right)^{1/2}t\right), \\ &\approx \left(\frac{\epsilon^2}{\Delta E^2}\right)\sin^2\left(\Delta E\, t\right), \end{aligned} \tag{18}$$

as an antineutron state, where $2\Delta E$ is the difference in neutron and antineutron energies in external fields and $\epsilon = \tau_{n\bar{n}}^{-1}$ is the energy characteristic of whatever new physics is responsible for the oscillations; $\tau_{n\bar{n}}$ is referred to as the oscillation time. We use units in which $\hbar = 1$. The experimental conditions are such that $\Delta E \gg \epsilon$, which justifies the approximation in equation (18). Furthermore, in the Grenoble experiment the quasi-free condition $\Delta Et \ll 1$ could be realized, leading to a transition probability,

$$P_t \approx \left(t/\tau_{n\bar{n}}\right)^2, \tag{19}$$

independent of the energy perturbation ΔE arising from the neutron and antineutron interactions with the ambient magnetic field. In the quasi-free condition $N = 3$ events were recorded in this experiment.

The background in the Grenoble experiment was directly measured by applying a magnetic field to suppress the transition probability P_t by making ΔE large enough. This condition yielded $M = 7$ events. The maximum likelihood estimate of the signal is $\hat{\mu} = N - M = -4$ events. However, since $\mu \geqslant 0$, we choose to take the best estimate of the signal in such experiments to be

$$\hat{\mu} = \begin{cases} N - M & \text{if} \quad N > M \\ 0 & \quad\text{otherwise.} \end{cases} \tag{20}$$

The sparsity of the data in the Grenoble experiment and our choice of signal estimate explicitly violate two of the regularity conditions for standard asymptotic results to hold [7]: the data should be sufficiently numerous and estimates must not lie on the boundary of the parameter space. The first condition is violated by the limited number of observations, and the second is violated by $\hat{\mu} = 0$, which lies on the boundary of the parameter space $\hat{\mu} \in [0, \infty)$. The violation of these regularity conditions, however, is not a problem for LF2I and ALFFI.

To construct confidence sets in the parameter space of $\theta = \{\mu, \nu\}$, we use the test statistic

$$\lambda\left(D;\theta\right) = -2\log\left[\frac{\mathcal{L}\left(D;\mu,\nu\right)}{\mathcal{L}\left(D;\hat{\mu},\hat{\nu}\right)}\right], \tag{21}$$

where $\hat{\mu}$ is given by equation (20) and $\hat{\nu}(\mu)$ by

$$\hat{\nu} = \begin{cases} M & \text{if} \quad \hat{\mu} = N - M \\ \left(M+N\right)/2 & \quad\text{otherwise.} \end{cases} \tag{22}$$

The cdf, $\mathbb{P}(\lambda \leqslant \lambda_D|\theta)$, was again approximated with a fully-connected feed-forward DNN with 3 input features $\mathbf{x} = \{\lambda_D, \mu, \nu\}$, 6 hidden layers with 12 nodes each, and a single output, estimating $\mathbb{E}\left[Z \mid \lambda_D, \mu, \nu\right]$. The activation function at each hidden node is a PReLU [24], and the network was trained with the Adam optimizer with a fixed learning rate of 6×10^{-4}. The training set is composed of 10^7 examples, which were used in batches of size 5×10^3, for the duration of 10^5 iterations, that is, for 50 epochs. A batch normalization [25] layer was added after every hidden layer, which was found to improve the results.

The DNN was used to compute the confidence sets shown in figure 4 and the associated coverage probabilities shown in figure 5. The fact that the coverage probabilities are close to the desired CLs confirms the accuracy of the confidence sets obtained with ALFFI.

3.3. Example 3: susceptible-Infected-Recovered (SIR) model

In the cosmology and signal/background examples, the likelihood functions are tractable. In our third example, from the field of epidemiology, the likelihood is intractable [26]. Therefore, this example is one for which LF2I and ALFFI are the most useful.

In this example, we fit the well-studied SIR epidemiological model to a classic data set, reproduced in figure 6, from a flu outbreak at an English Boarding School [27]. The SIR model has a nearly 100-year history, beginning with the work of Kermack and McKendrick in 1927 [28]. The simplest version of the model comprises coupled ordinary differential equations (ODEs) and assumes a population of individuals that is closed and well-mixed in which individuals fall into one of three classes or compartments: susceptible (S), infectious (I), and recovered or removed (R). The rate of change of susceptible individuals depends on the rate at which new infections arise as a result of contact between infectious and susceptible individuals. It is typically assumed that the contact rate (number of contacts per unit time) is proportional to the total

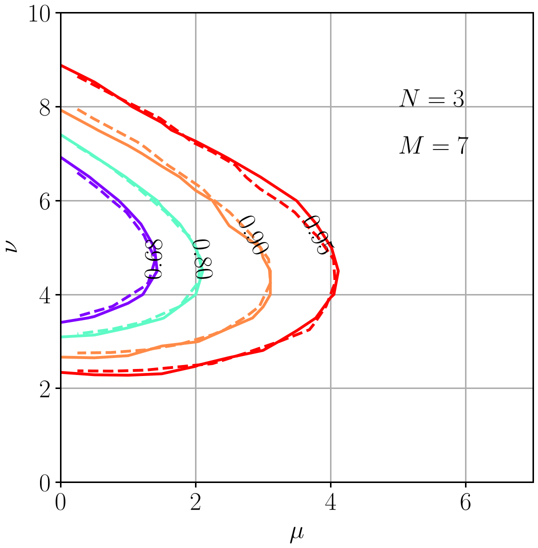

Figure 4. Example 2 (signal/background): confidence sets $R(D)$ for $\tau = 0.68, 0.80, 0.90$, and 0.95. (*dashed lines*) Boundaries of confidence sets, $R(D)$, defined by $\mathbb{P}(\lambda \leqslant \lambda_D|\theta) = \tau$ with the histogram-based approximation of the cdf. (*solid lines*) Boundaries of confidence sets computed using the DNN-based approximation of the cdf.

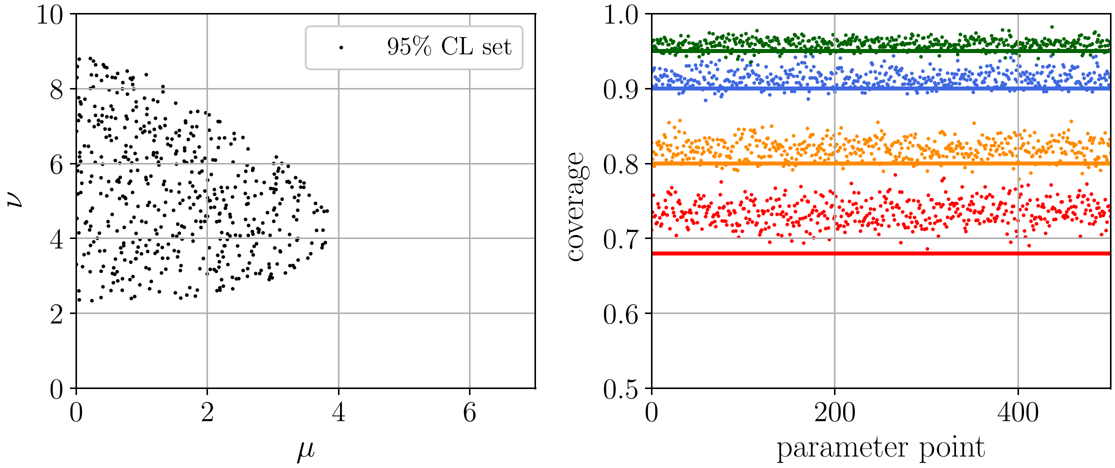

Figure 5. Example 2 (Signal/Background): (*left*) 500 points, from 95% CL confidence set for the Grenoble data [22], at which the coverage has been computed. (*right*) the coverage probabilities for each of the 500 points in the signal/background $\{\mu, \nu\}$ parameter space.

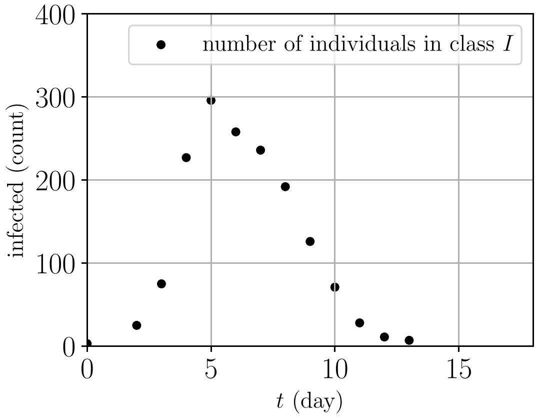

Figure 6. Example 3 (SIR): english boarding school data. The number of infected individuals at the reported times in days from the start of the flu outbreak.

population size N or constant. The assumption that contacts are proportional to the total population size leads to a mass action term βSI for the number of new infections per unit time (called the *incidence* in epidemiology). Infectious (I) individuals can leave the infectious class through recovery (or death) and progress to the the Recovered (or Removed) class, R. It is often assumed that the number of recoveries/removals per unit time is linear in I: αI. Under this assumption, the time spent in I is exponentially distributed with mean $1/\alpha$. The resulting system of ODEs is

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}t} &= -\beta SI,\\ \frac{\mathrm{d}I}{\mathrm{d}t} &= -\alpha I + \beta SI,\\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \alpha I.\end{aligned} \tag{23}$$

The qualitative dynamics of this model are governed by an epidemiological quantity called the basic reproduction number, R_0, which for this SIR model is $R_0 = \beta/\alpha$. If $R_0 > 1$, the model exhibits a single outbreak, with I increasing to a peak, then approaching zero as time tends to infinity. Under this scenario, S will decay and approach a positive, constant value, meaning that the epidemic does not infect the whole population. If $R_0 \leqslant 1$, I will decay and approach zero as time tends to infinity. This model has played an invaluable role in understanding infectious disease dynamics, informing control strategies, and serving as a building block for more complex compartmental epidemiological modeling frameworks.

The SIR model is often fitted to data by minimizing a weighted least squares function,

$$F(\theta) = \sum_{n=1}^{N} w_n (x_n - I_n)^2, \tag{24}$$

where the data $D = \{x_1, \ldots, x_N\}$ are the number of infected individuals at the corresponding reporting times $t_1, \ldots, t_N$, $I_n = I(t_n, \theta)$ is the predicted mean infection count at time t_n, found by solving equations (23) for the given θ, and w_n are weights. Following example 1, we choose a test statistic that is transformed and scaled

$$\lambda(D, \theta) = \sqrt{F(\theta)/N}/50, \tag{25}$$

so that the statistic is $\mathcal{O}(1)$. The weights are set to $w_n = I_n^{-1}$, which we hasten to add should not be taken to imply that the counts x_n are Poisson distributed. On the contrary, the counts x_n are correlated and their fluctuations are super-Poissonian.

We follow a similar training protocol as for example 1, except that the training sample size for this example is 750 000, a fully-connected DNN is used with 6 hidden layers and 25 nodes each, and the best model is found after about 350 000 iterations, that is, after 23 epochs. The confidence sets obtained are shown in figure 7 and the coverage probabilities are shown in figure 8. Again, we find that ALFFI produces accurate confidence sets.

4. Discussion

The approximated p-value function in LF2I is computed for a specific data set D, therefore, it cannot be used for other similar data sets with the same sample size. Consequently, the LF2I p-value function cannot be used to check the coverage explicitly at a given parameter point. In LF2i, an explicit coverage check can be performed using the critical value function $\widehat{C}_\alpha$, but only for a given α. LF2I provides an algorithm to train another neural network to approximate the coverage probability over the parameter space of the theoretical model. But, unfortunately, a reliable way to quantify the accuracy of the approximated coverage probability function is not available as is true of the neural network approximation of the p-value. Therefore, it is useful to devise methods that make the explicit calculation of coverage at any given parameter point, and for any level α, straightforward. Such methods make it possible to check the quality of the confidence sets by assessing the degree to which the coverage matches the CL $\tau = 1 - \alpha$, at any given point. If the coverage probabilities within the neighborhood of the estimated parameters agree with the desired CL, τ, then one may conclude that the confidence sets, $R(D)$, are satisfactory. The motivation for the extension introduced in this paper is the desire to have the p-value do double duty: 1) determine the confidence sets and 2) permit the explicit checking of the coverage using the same neural network.

A simple extension of the LF2I algorithm for the critical value function makes it possible to use the same approximation, $\widehat{C}_\alpha$, to construct confidence sets and check their coverage for any value of the CL τ. The LF2I algorithm is extended by including the CL τ as an input to the neural network and by using random

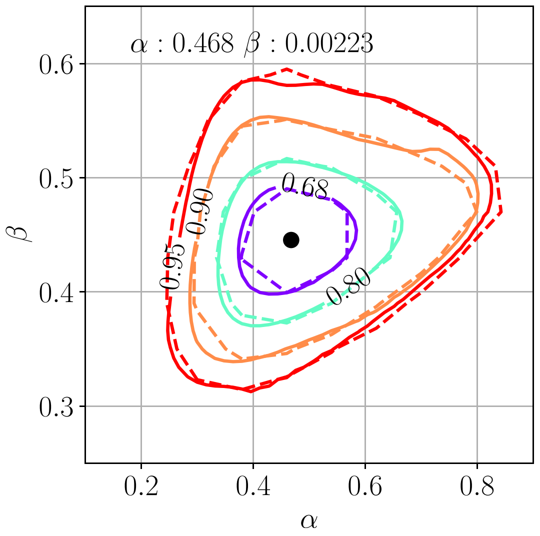

Figure 7. Example 3 (SIR): confidence sets $R(D)$ for $\tau = 0.68, 0.80, 0.90$, and 0.95. (dashed lines) Boundaries of confidence sets computed with the histogram-based approximation of the cdf. (solid lines) Boundaries of the confidence sets computed with the DNN-based approximation of the cdf. (*black dot*) Location of the minimum of the DNN-based cdf, which is taken to be the best-fit point.

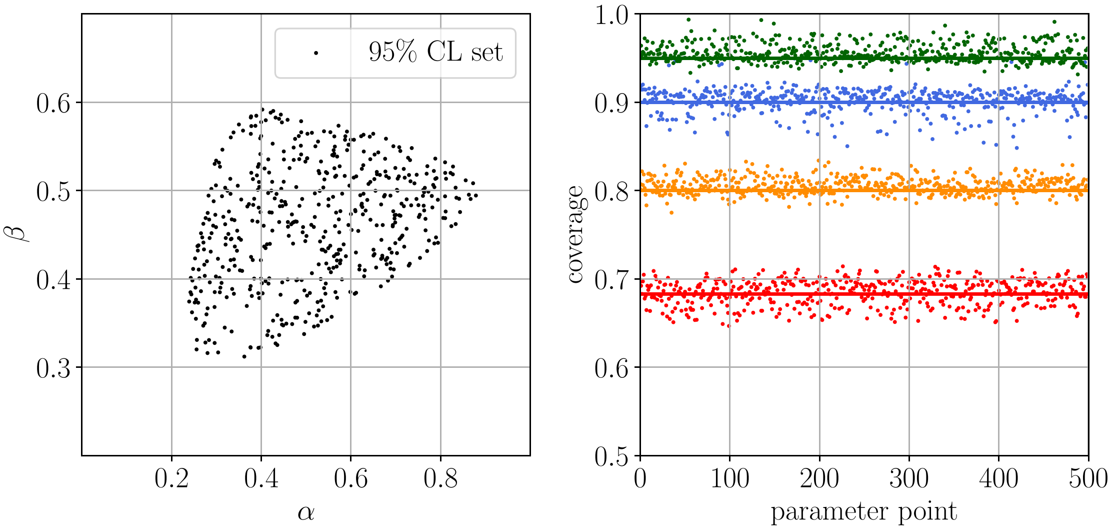

Figure 8. Example 3 (SIR): (*left*) 500 points, from the 95% CL confidence set for the Boarding School data, at which the coverage has been computed in the SIR α, β parameter space.(*right*) the coverage probabilities for each of the 500 points in the SIR parameter space for different confidence levls (solid horizontal lines).

values of τ during training. In practice, we augment the training data by assigning a random value of τ to every sampled point θ. During training the quantile loss associated with each point θ is calculated with the associated value of τ. Since the critical value network is a function of θ and τ, it can used in a way analogous to how the cdf is used in ALFFI. To illustrate the efficacy of this extension to LF2I and also provide a comparison with ALFFI, the extended algorithm was applied to the SIR example, but using a much smaller training sample size of 80 000 and a network with 5 hidden layers rather than 6, with coverage computed at 249 points. Figure 9 shows coverage results for confidence sets computed using the τ-dependent network. The results are comparable to those in figure 8 computed using ALFFI. But we advise caution in drawing firm conclusions from the comparison as the structures of the neural network models used have not been systematically optimized.

The computation of the coverage at a given point θ entails determining which confidence sets contain θ, a calculation that can be done extremely fast with ALFFI and the critical value function of LF2I. In both approaches, the computational burden required to compute the coverage at a given point θ is simply the burden of generating a sufficient number of simulations at that point. For example, suppose that one is

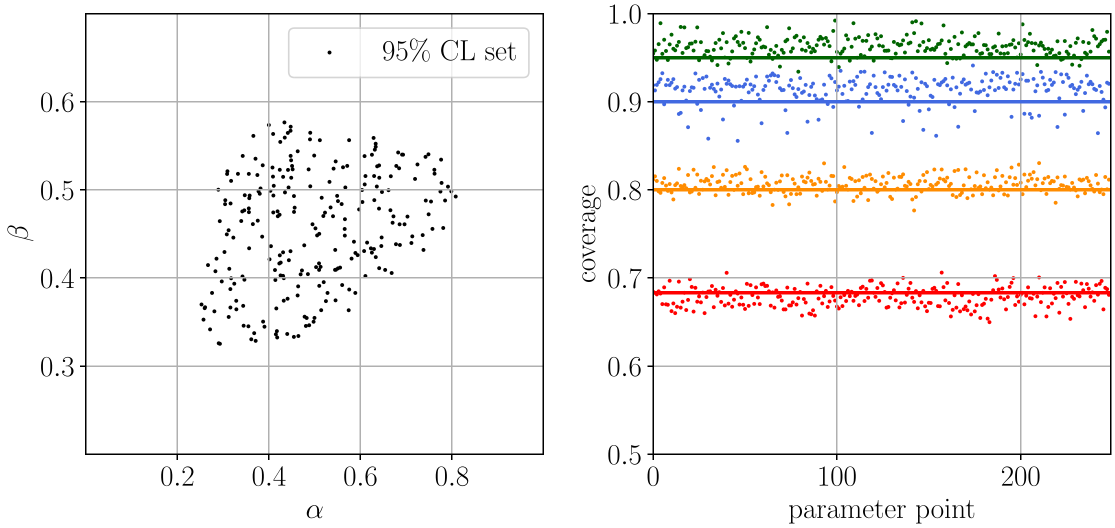

Figure 9. Example 3 (SIR with extended ALFFI critical value function): (*left*) 249 points, from the 95% CL confidence set for the Boarding School data, at which the coverage has been computed in the SIR α, β parameter space. (*right*) the coverage probabilities for each of the 249 points in the SIR parameter space for different confidence levels (solid horizontal lines).

interested in computing the coverage probability, $\mathbb{P}(\theta \in R(\mathcal{D}) \mid \theta)$, at the point θ to an accuracy of $\approx$5%, then 400 simulated data sets at that point would be sufficient.

Although we demonstrate the use of ALFFI for simultaneous inference on two parameters in each of the three examples, the method, in principle applies to problems with any number of parameters. However, it remains to be seen how, in practice, ALFFI's accuracy in computing confidence sets scales with the dimensionality of the theoretical model, and how that scaling compares with that of LF2I. Understanding this scaling would require a dedicated study.

In the cosmological and epidemiological models all parameters are of interest. But in the signal/background-On/Off problem one is typically interested only in the mean signal μ. The mean background, ν, is a *nuisance parameter*. Unfortunately, constructing confidence intervals for μ when there are nuisance parameters and when the data are sparse is challenging, though approximate methods exist (see, for example, [6]) including in LF2I. Given the success of LF2I and ALFFI in producing reasonably accurate confidence sets, it is of interest to explore whether the reasoning that underlies these approaches can be extended to an algorithm that can yield provably valid confidence intervals for individual parameters, ideally constructed from the associated multi-parameter confidence set.

One possible approach to construct confidence intervals from the confidence sets created with LF2I and ALFFI is to mimic the way that confidence intervals can be constructed from confidence ellipsoids for a multivariate normal density. In the two dimensional example, the objective would be to map a 2-dimensional confidence set to a one-dimensional confidence interval, as is done in the bivariate normal density. For example, in the signal/background problem, one can construct an interval for μ as follows: $I(D) = [\min(\{\mu_i\}), \max(\{\mu_i\})]$, where $\{(\mu_i, \nu_i)\}$ are points from the associated confidence set. It should be possible to use an algorithm like ALFFI to map from the CL of the set to that of the associated interval for the parameter of interest. If this can be done, then one would be able to determine what CL is needed for the confidence set to obtain the desired CL for the associated interval. To the best of our knowledge, if such a mapping could be devised in the general case it would constitute the first method in which valid multidimensional confidence sets can be mapped to valid one-dimensional confidence intervals without the need for explicit knowledge of the underlying statistical model. Ideas along these lines are under investigation.

In LF2I and ALFFI, the approximated probabilities do not depend on the prior π_θ since the network is conditional on the parameters θ. However, the accuracy of the approximation depends on the prior. Greater accuracy is expected where the density of sampled parameters is greater, just as the accuracy of any machine learning model typically varies across the space of inputs reflecting the distribution of training data over that space. Therefore, it makes sense to choose a prior that places the parameter points when they are needed most. For example, if one knew approxinately where the 68% CL sets are located in the parameter space, it would be advantageous to choose a prior that places the points in the neighborhood of those sets. Alternatively, the parameter points could sampled using an active learning approach.

5. Conclusions

The LF2I and ALFFI approaches are useful when asymptotic results [6, 7] may not be applicable, and are particularly useful when the likelihood function is intractable. The ALFFI approach extends the p-value approximation of LF2I by approximating one minus the p-value (that is, the cdf) of a test statistic with a neural network in which the test statistic is an input. This makes it possible to check the coverage probability of confidence sets, for any desired CL, at any point in the parameter space of a theoretical model using the *same* neural network. In LF2I one can check the coverage explicitly with the approximation to the critical value function, but not with the approximation to the p-value function. Direct calculation of the coverage probability at any given point provides an *a posteriori* assessment of the accuracy of the confidence sets constructed with ALFFI. In addition, by directly binning the point cloud of theoretical model parameters, it is possible to cross-check the neural network approximation.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/AliAlkadhim/ALFFI.

Acknowledgments

This work was performed in part at the Aspen Center for Physics, which is supported by National Science Foundation Grant PHY-1607611. This work is supported in part by the Department of Energy under Award No. DE-SC0010102 (HP) and supported in part by National Science Foundation Grant DMS-2045 843 (OP). We thank Prof. Ann Lee for fruitful discussions at the Aspen Physics Center in summer 2022.

Appendix A. ALFFI algorithm

Algorithm 1. Estimate the CDF $\mathbb{C}(\lambda_D \mid \theta_0) = \mathbb{P}(\lambda < \lambda_D \mid \theta_0)$, given the observed value λ_D of a test statistic λ.

```
Ensure: estimated CDF $\widehat{\mathbb{C}}(\lambda_D \mid \theta)$ for all $\theta = \theta_0 \in \Theta$
Set $\mathcal{T} \leftarrow \emptyset$
for $i$ in $\{1, \ldots, B\}$ do
    Draw parameter $\theta_i \sim \pi_\theta$
    Simulate $\mathcal{D}_i \leftarrow \{X_1, \ldots, X_n\}_i \sim F_\theta$
    Compute test statistic $\lambda_i \leftarrow \lambda(\mathcal{D}_i, \theta_i)$
    Simulate $\mathcal{D}_i' \leftarrow \{X_1, \ldots, X_n\}_i' \sim F_\theta$
    Compute observed test statistic under the null $\theta = \theta_0$, $\lambda_i' \leftarrow \lambda(\mathcal{D}_i', \theta_i)$
    Compute discrete indicator variable $Z_i \leftarrow 1(\lambda_i < \lambda_i')$
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{(Z_i, \theta_i, \lambda_i')\}$
end for
Use $\mathcal{T}$ to learn the function $\widehat{\mathbb{C}}(\lambda_D \mid \theta)$
return $\widehat{\mathbb{C}}(\lambda \mid \theta)$.
```

Appendix B. Cosmological model: details

The Equation of state of our cosmological model is

$$\mathcal{P} = -ba^n\Omega, \tag{B.1}$$

where n and b are free parameters, and $\mathcal{P}$, $a(t)$, and $\Omega(a)$ are the dimensionless pressure, the dimensionless universal scale factor, and the dimensionless energy density, respectively, and t is the elapsed time since the Big Bang. For $n > 1$ and $a \ll 1$, the equation of state is that of a pressureless dust of particles as in the ΛCDM model [15]. However, at later times the energy density becomes dominated by so-called phantom energy [29]. Our model is consistent with the Friedmann-Lemaître-Robertson-Walker metric with zero curvature, and the Friedmann equations [15],

$$\left(\frac{1}{a}\frac{\mathrm{d}a}{\mathrm{d}t}\right)^2 = \mathcal{H}_0^2\Omega(a) \tag{B.2}$$

$$\text{and } a\frac{\mathrm{d}\Omega}{\mathrm{d}a} = -3\left(\Omega + \mathcal{P}\right), \tag{B.3}$$

are assumed to hold, where $\mathcal{H}_0$ is the Hubble constant. The first Friedmann equation, equation (B.2), incorporates the convention $a(t_0) = 1$ at the present epoch t_0. By definition, the Hubble constant is the present value of the Hubble parameter $\mathcal{H}(a) = a^{-1}\mathrm{d}a/\mathrm{d}t$, which implies $\Omega(1) = 1$ for all cosmological models.

Combining equations (B.1) and (B.3), integrating the latter, and imposing the constraint $\Omega(1) = 1$, yields

$$\Omega\left(a\right) = \exp\left[3b\left(a^n - 1\right)/n\right]/a^3, \tag{B.4}$$

for the dimensionless energy density. This model is defined by the three parameters n, b, and $\mathcal{H}_0$, but we reduce it to a 2-parameter model by choosing $b = n/3$. From the first Friedmann equation, equation (B.2), the energy density $\Omega(a)$ and the dimensionless time $\mathcal{H}_0 t$ are related as follows

$$\begin{aligned}\mathcal{H}_0 t &= \int_0^a \frac{\mathrm{d}y}{y\sqrt{\Omega\left(y\right)}}, \\ &= \sqrt{e}\,2^{\frac{3}{2n}}\gamma\left(\frac{3}{2n}, \frac{a^n}{2}\right)/n,\end{aligned} \tag{B.5}$$

where $\gamma(a,x) = \int_0^x t^{a-1}e^{-t}\,\mathrm{d}t$ is the lower incomplete gamma function. Setting $a = 1$ yields the dimensionless age of the Universe

$$\mathcal{H}_0 t_0 = \sqrt{e}\,2^{\frac{3}{2n}}\gamma\left(\frac{3}{2n}, \frac{1}{2}\right)/n. \tag{B.6}$$

The *distance modulus* [15], which an astronomical measure of distance, is given by

$$\mu\left(z,\theta\right) = 5\log_{10}\left[\left(1+z\right)\sin\left(\sqrt{-\Omega_K}\,u\left(z,\theta\right)\right)/\sqrt{-\Omega_K}\right] + 5\log_{10}\left(c/\mathcal{H}_0/10^{-5}\mathrm{Mpc}\right), \tag{B.7}$$

where c is the speed of light in vacuum in km/s, Ω_K is the curvature parameter, and

$$u\left(z,\theta\right) = \int_{1/(1+z)}^1 \frac{da}{a^2\sqrt{\Omega\left(a\right)}}, = 2^{\frac{1}{2n}}\left[\gamma\left(\frac{1}{2n}, \frac{1}{2}\right) - \gamma\left(\frac{1}{2n}, \frac{\left(1+z\right)^{-n}}{2}\right)\right]\sqrt{e}/n,$$

is a dimensionless function. With $\Omega_K \to 0$, the distance modulus simplifies to

$$\mu\left(z,\theta\right) = 5\log_{10}\left[\left(1+z\right)c\,u\left(z,\theta\right)/\mathcal{H}_0\right] + 25. \tag{B.8}$$

For $t < t_0$, the time dependence of the scale factor $a(t)$ is similar to that of the standard ΛCDM model. But the model exhibits a future singularity (a Big Rip) characterized by the condition $a \to \infty$ at a *finite* time given by

$$\mathcal{H}_0 t_{\text{rip}} = \sqrt{e}\,2^{\frac{3}{2n}}\Gamma\left(\frac{3}{2n}\right)/n, \tag{B.9}$$

that is, at

$$t_{\text{rip}} = \frac{\Gamma\left(\frac{3}{2n}\right)}{\gamma\left(\frac{3}{2n}, \frac{1}{2}\right)}\,t_0.$$

ORCID iD

Ali Al Kadhim https://orcid.org/0000-0003-3490-8407

References

[1] Brehmer J 2021 *Nat. Rev. Phys.* **3** 305
[2] Cranmer K, Brehmer J and Louppe G 2020 *Proc. Nat. Acad. Sci.* **117** 30055–62
[3] Dalmasso N, Masserano L, Zhao D, Izbicki R and Lee A B 2023 Likelihood-free frequentist inference: confidence sets with correct conditional coverage (arXiv:2107.03920)
[4] Neyman J 1937 *Phil. Trans. R. Soc.* A **236** 333–80
[5] Neyman J and Pearson E S 1933 *Phil. Trans. R. Soc* A **231** 289–337
[6] Cowan G, Cranmer K, Gross E and Vitells O 2011 *Eur. Phys. J.* C **71** 1554
[7] Algeri S, Aalbers J, Dundas Mora K and Conrad J 2020 *Nat. Rev. Phys.* **2** 245–52

[8] Li T P and Ma Y Q 1983 *Astrophys. J.* **272** 317–24
[9] Gillespie D T 1983 *Am. J. Phys.* **51** 520–33
[10] Ruck D, Rogers S, Kabrisky M, Oxley M and Suter B 1990 *IEEE Trans. Neural Netw.* **1** 296–8
[11] Richard M D and Lippmann R P 1991 *Neural Comput.* **3** 461–83
[12] Riess A G *et al* 1998 *Astron. J.* **116** 1009
[13] Perlmutter S *et al* 1999 *Astrophys. J.* **517** 565–86
[14] Suzuki N *et al* Supernova Cosmology Project Team 2012 *Astrophys. J.* **746** 85
[15] Peebles P J E and Ratra B 2003 *Rev. Mod. Phys.* **75** 559–606
[16] Agarap A F 2018 Deep learning using rectified linear units (relu) (arXiv:1803.08375)
[17] Paszke A *et al* 2019 Pytorch: an imperative style, high-performance deep learning library (arXiv:1912.01703)
[18] Kingma D P and Ba J 2017 Adam: a method for stochastic optimization (arXiv:1412.6980)
[19] Raskutti G, Wainwright M J and Yu B 2014 *J. Mach. Learn. Res.* **15** 335–66
[20] Lyons L, Prosper H B and De Roeck A (eds) 2008 (ed) *Statistical Issues for LHC Physics. Proc., Workshop, PHYSTAT-LHC* (*Geneva, Switzerland, June 27-29, 2007*) (*CERN Yellow Reports: Conf. Proc.*)
[21] Wilks S S 1938 *Ann. Math. Statist.* **9** 60–62
[22] Fidecaro G *et al* (CERN-GRENOBLE-PADUA-RUTHERFORD-SUSSEX) 1985 *Phys. Lett.* B **156** 122–8
[23] Phillips D G I I *et al* 2016 *Phys. Rep.* **612** 1–45
[24] He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: surpassing human-level performance on imagenet classification (arXiv:1502.01852)
[25] Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift (arXiv:1502.03167)
[26] Andersson H and Britton T 2012 *Stochastic Epidemic Models and Their Statistical Analysis* vol 151 (Springer)
[27] Anon 1978 *Br. Med. J.* **1** 587
[28] Kermack W O and McKendrick A G 1927 *Proc. R. Soc.* A **115** 700–21
[29] Caldwell R R, Kamionkowski M and Weinberg N N 2003 *Phys. Rev. Lett.* **91** 071301