Anonymized

Anonymous Authors ANONYMIZED@EMAIL.COM

Abstract

Many areas of science make extensive use of computer simulators that implicitly encode likelihood functions of complex systems. Classical statistical methods are poorly suited for these so-called likelihood-free inference (LFI) settings, particularly outside asymptotic and low-dimensional regimes. Although new machine learning methods, such as normalizing flows, have revolutionized the sample efficiency and capacity of LFI methods, it remains an open question whether they produce confidence sets with correct conditional coverage for small sample sizes. This paper unifies classical statistics with modern machine learning to present (i) a practical procedure for the Neyman construction of confidence sets with finite-sample guarantees of nominal coverage, and (ii) diagnostics that estimate conditional coverage over the entire parameter space. We refer to our framework as *likelihood-free frequentist inference* (LF2I). Any method that defines a test statistic, like the likelihood ratio, can leverage the LF2I machinery to create valid confidence sets and diagnostics without costly Monte Carlo samples at fixed parameter settings. We study the power of two test statistics (`ACORE` and `BFF`), which, respectively, maximize versus integrate an odds function over the parameter space. Our paper discusses the benefits and challenges of LF2I, with a breakdown of the sources of errors in LF2I confidence sets.

Keywords: likelihood-free inference, simulation-based inference, frequentist coverage, confidence sets, Neyman inversion

1. Introduction

Hypothesis testing and uncertainty quantification are the hallmarks of scientific inference. Methods that achieve good statistical performance (e.g., high power) often rely on being able to explicitly evaluate a likelihood function, which relates parameters of the data-generating process to observed data. However, in many areas of science and engineering, complex phenomena are modeled by forward simulators that *implicitly* define a likelihood function. For example,[1] given input parameters θ from some parameter space Θ, a stochastic model F_θ may encode the interaction of atoms or elementary particles, or the transport of radiation through the atmosphere or through matter in the Universe, by combining deterministic dynamics with random fluctuations and measurement errors to produce synthetic data $\mathbf{X}$.

Simulation-based inference without an explicit likelihood is commonly referred to as *likelihood-free inference* (LFI). The most well-known approach to LFI is Approximate Bayesian Computation (ABC; see Beaumont et al. 2002; Marin et al. 2012; Sisson et al. 2018 for a review). These methods use simulations sufficiently close to the observed data $D = \left\{\mathbf{x}_1^{\text{obs}}, \ldots, \mathbf{x}_n^{\text{obs}}\right\}$ to infer the underlying parameters, or more precisely, the posterior distri-

1. **Notation.** Let F_θ with density f_θ represent the stochastic forward model for a sample point $\mathbf{X} \in \mathcal{X}$ at parameter $\theta \in \Theta$. We refer to F_θ as a "simulator" as the assumption is that we can sample data from the model. We denote i.i.d "observable" data from F_θ by $\mathcal{D} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, and the actually observed or measured data by $D = \left\{\mathbf{x}_1^{\text{obs}}, \ldots, \mathbf{x}_n^{\text{obs}}\right\}$. The likelihood function $\mathcal{L}(D; \theta) = \prod_{i=1}^n f_\theta(\mathbf{x}_i^{\text{obs}})$.

bution $p(\theta|D)$. Recently, the arsenal of LFI methods has been expanded with new machine learning algorithms (such as neural density estimators) that instead use the output from simulators as training data; see Section 1.1, "Likelihood-free inference via machine learning". The objective here is to learn a "surrogate model" or *approximation* of the likelihood $p(D|\theta)$ or posterior $p(\theta|D)$. The surrogate model, rather than the simulations themselves, is then used for inference.

While the field of likelihood-free inference has undergone a revolution in terms of the complexity and dimensionality of problems that can be tackled (see Cranmer et al. 2020 for a recent review), the question remains (see, e.g., Hermans et al. 2021) whether machine-learning (ML) based confidence sets of θ, where θ are unknown parameters, indeed have correct conditional coverage. Suppose that we have a high-fidelity simulator F_θ and we observe data $\mathcal{D}$ of finite sample size n. Can we then, without restrictive assumptions on F_θ, construct a random set $R(\mathcal{D})$ that satisfies

$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in R(\mathcal{D}) \,|\, \theta\right) = 1 - \alpha \tag{1}$$

for a certain confidence level $\gamma = 1 - \alpha$, *no matter what the true parameter* $\theta \in \Theta$ *is*? Furthermore, can we find a procedure that is both computationally tractable, and that has theoretical guarantees of validity and power for small n? Finally, how do we in practice check performance of the constructed sets $R(\mathcal{D})$; that is, how do we check whether the actual coverage of $R(\mathcal{D})$ is indeed close to, and no smaller than, γ for *any* $\theta \in \Theta$ (without resorting to costly Monte Carlo simulations at fixed parameter settings on a "fine enough" grid in parameter space Θ (Cousins, 2018, Section 13)).

In this paper, we describe a new statistical framework for LFI which unifies classical statistics with modern machine learning (e.g., deep generative models, neural network classifiers, and quantile regression) to present practical procedures for (i) constructing finite-sample confidence sets with correct conditional coverage, and for (ii) computing diagnostics that estimate conditional coverage over the entire parameter space. We refer to our framework as *likelihood-free frequentist inference* (LF2I).

At the heart of the LF2I framework is the *Neyman construction of confidence sets* (Figure 2) — albeit applied to a setting where the distribution of the test statistic is unknown. The construction of frequentist confidence sets with correct conditional coverage has a long history in statistics (Fisher, 1925; Neyman, 1935), with the equivalence between tests and confidence sets formalized by Neyman 1937. Of particular note is the impact classical statistical procedures (including the Neyman construction) have had on scientific fields such as high energy physics (HEP); see Section 1.1, "Relation to Other Work". Most simulator-based methods, however, do not have theoretical guarantees on validity or power of Neyman confidence sets beyond low-dimensional data settings and large-sample theory assumptions (Feldman and Cousins, 1998).

What makes the Neyman construction of confidence sets difficult to implement for LFI is not only that one cannot evaluate the likelihood ratio (LR) statistic (Equation 5); one also needs to consider the hypothesis test $H_{0,\theta_0} : \theta = \theta_0$ for *every* $\theta_0 \in \Theta$. Simulation-based Monte Carlo and bootstrap approaches to hypothesis testing typically estimate critical values and significance probabilities (p-values) from a batch of simulations generated at the null value θ_0 (see, e.g., MacKinnon (2009) and Ventura (2010)). Such an approach is computationally inefficient and infeasible in higher-dimensional parameter spaces, because

the Neyman construction would then require a separate MC or bootstrap batch at each θ_0 on a fine grid in parameter space. Hence, in practice, Neyman inversions either rely on parametric model assumptions, or large sample asymptotic theory (Neyman and Pearson, 1928; Wilks, 1938). There are however at least two cases where one cannot assume that the LR statistic follows a χ^2 distribution: when the statistical model is irregular (see Section 7.1 for a Gaussian mixture model with intractable null distribution), and when the sample size n is small. Note that high-dimensional data and sample sizes as small as $n = 1$ are common in physics; recent examples include estimating the muon momentum (θ) from the energy deposited in a finely segmented calorimeter ($\mathbf{X} = \mathcal{D}$) by a particular muon (Kieseler et al., 2022), and estimating the mass of a galaxy cluster (θ) from velocities and projected radial distances ($\mathbf{X} = \mathcal{D}$) for a particular line-of-sight of the observer relative the galaxy cluster (Ho et al., 2021). In this work, we ask: How can we quickly and accurately estimate critical values and conditional coverage across the entire parameter space, when we do not know the distribution of the test statistic and cannot rely on large-sample approximations?

The main idea behind LF2I is that key quantities of interest in frequentist statistical inference — test statistics, critical values, p-values and confidence set coverage — are *conditional distribution functions of the (unknown) parameters* θ, and generally vary smoothly over the parameter space Θ. As a result, one can leverage machine learning methods and data simulated in the neighborhood of a parameter to improve estimates of quantities of interest with fewer total simulations. Figure 1 illustrates the general LF2I inference machinery. There are three main branches; each branch is modular with separate functionality:

- The main branch (Figure 1 center and Section 3.2) estimates a *test statistic* $\lambda(\mathcal{D};\theta)$ from a sample $\mathcal{T} = \{(\theta_i, \mathbf{X}_i, Y_i)\}_{i=1}^{B}$, which is generated by the simulator F_θ and a reference distribution G. In this paper, we study the theoretical and empirical performance of LF2I confidence sets derived from test statistics based on the odds function $\mathbb{O}(\mathbf{X};\theta)$ (Equation 7) for data $\mathbf{X}$ being generated from F_θ rather than from G for every $\theta \in \Theta$.

- The "calibration" branch (Figure 1 left and Section 3.3) generates a second sample $\mathcal{T}' = \{(\theta_i', \mathcal{D}_i')\}_{i=1}^{B'}$ to estimate the *critical value* C_{θ_0} for every level-α test of $H_{0,\theta_0} : \theta = \theta_0$ vs. $H_{1,\theta_0} : \theta \neq \theta_0$, via a quantile regression of the test statistic $\lambda(\mathcal{D};\theta_0)$ on $\theta_0 \in \Theta$. Once we have estimated the conditional quantile function $\widehat{C}_{\theta_0}$, we can directly construct Neyman confidence sets

$$\widehat{R}(\mathcal{D}) := \left\{ \theta \in \Theta \,\middle|\, \lambda(\mathcal{D};\theta) \geq \widehat{C}_\theta \right\} \tag{2}$$

  with approximate $(1-\alpha)$ finite-n conditional coverage. Alternatively, we can estimate a function of p-values, $\widehat{p}(D;\theta_0)$ for every test at $\theta = \theta_0$ with observed data D.

- The final "diagnostics" branch (Figure 1 right and Section 3.4) generates a third sample $\mathcal{T}'' = \{(\theta_i'', \mathcal{D}_i'')\}_{i=1}^{B''}$ to assess *conditional coverage* $\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D}) \mid \theta\right)$ of the constructed confidence sets $\widehat{R}(\mathcal{D})$ for every $\theta \in \Theta$, via a regression of the indicator variable $W := \mathbb{I}\left(\lambda(\mathcal{D};\theta) \geq \widehat{C}_\theta\right)$ on θ. That is, this procedure checks whether the constructed sets are indeed conditionally valid.

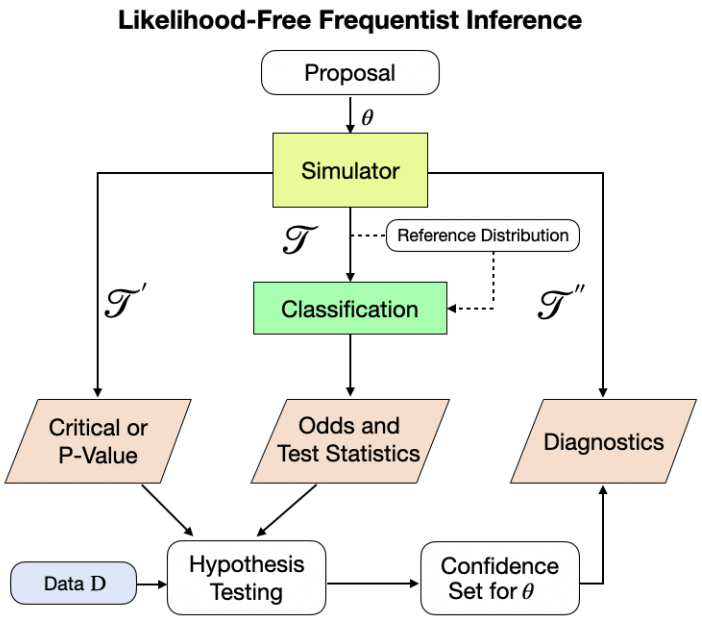


Figure 1: Schematic diagram of likelihood-free frequentist inference (LF2I). *Center:* The simulator generates a sample $\mathcal{T}$ of size B for learning the odds function $\mathbb{O}(\mathbf{X};\theta)$ and a test statistic $\lambda(\mathcal{D};\theta)$ via probabilistic classification. *Left:* The simulator generates a second sample $\mathcal{T}'$ of size B' for learning the critical values C_θ or p-values $p(\mathcal{D};\theta)$ for all $\theta \in \Theta$. Once data D are observed, then we can construct confidence sets $\widehat{R}(\mathcal{D})$ according to Equation 12. *Right:* The LF2I diagnostic branch checks whether the conditional coverage $\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D}) \mid \theta\right)$ of the confidence set is indeed correct across the entire parameter space. The three main parts of the inference machinery (critical or p-value estimation, odds and test statistic estimation, diagnostics) are separate modules, and only need to be trained once.

The three-branch LF2I machinery for constructing and assessing confidence sets with finite-n conditional coverage was first proposed by Anonymized 2020. The original version was called `ACORE` for "Approximate Computation via Odds Ratio Estimation", and is based on a test statistic that maximizes the odds over the parameter space. When the odds function is well-estimated, the `ACORE` statistic (Equation 8) is the same as the well-known likelihood ratio statistic.

In this paper, we study the theoretical and empirical properties of the more general LF2I framework. In particular,

- we highlight that there are LFI scenarios where one might benefit from targeting other statistics than the likelihood ratio statistic; we introduce the Bayes factor (Jeffreys, 1935, 1961) as a frequentist test statistic in LF2I, and refer to the approach as *Bayes Frequentist Factor* inference, or `BFF` for short. Because `BFF` averages (rather than maximizes) an estimate of $\mathbb{O}(\mathbf{X};\theta)$ over the parameter space, it can lead to statistical procedures that are more robust and powerful in practice (Section 3.2.2);

- we prove that LF2I leads to finite-n confidence sets with correct conditional coverage, if the quantile regression for estimating critical values or the probabilistic classifier for estimating p-values is consistent, and the number of simulations $B' \to \infty$; this result holds regardless of whether the test statistic is well-estimated (Theorem 8 and Corollary 1), and in practice we observe that correct coverage is achieved with relatively few simulations B' (Section 7);

- we compare our technique of estimating critical values to universal cross-fit LRT (Wasserman et al., 2020) for the setting in Dunn et al. (2021), where the likelihood can be evaluated (Section 7.2.1);

- we analyze the power of `BFF` and prove that (when there is no need for numerical integration) the power of the test based on `BFF` relative to that of the Bayes factor is determined by the convergence rate of the probabilistic classifier and the number of simulations B used to estimate the `BFF` statistic in Equation 10 (Section 4.3);

- we study scaling with increasing feature and parameter dimension in a true LFI setting, where both test statistics and critical values are estimated from simulations (Section 7.2.2);

- we provide a breakdown of the sources of errors in LF2I: when choosing a test statistic, one needs to consider both statistical errors (due to finite number of observations or simulations) and numerical errors (due to choice of optimization or integration algorithm) (Section 5);

- we propose practical hybrid procedures for handling nuisance parameters in LF2I; hybrid procedures are known to only control type I error approximately in hypothesis testing (Section 6);

- we demonstrate for a high-energy physics example that our method for assessing conditional coverage of constructed confidence sets across parameter space can provide practical guidance and insights as to the choice of LFI algorithm or hybrid technique for the problem at hand (Section 7.3).

In summary: this work presents a unified approach to likelihood-free frequentist inference, which goes beyond estimating key inferential quantities in LFI (such as likelihoods and posteriors) to constructing and assessing whether confidence sets indeed have correct conditional coverage. We make a minimum of distributional assumptions, and we are targeting challenging data settings (complex high-dimensional data $\mathbf{X}$, intractable likelihood models, small sample size n or unknown limiting distribution of test statistic), which commonly occur in the physical, engineering, and biological sciences.

1.1 Relation to Other Work

Our work draws on different ideas in the statistics, physics and machine learning (ML) literature:

Classical statistical inference in high-energy physics (HEP). Our LF2I approach is inspired by pioneering work in HEP that adopted classical hypothesis tests and Neyman confidence sets for the discovery of new physics (Feldman and Cousins, 1998; Cowan et al., 2011; Aad et al., 2012; Chatrchyan et al., 2012; Cranmer, 2015). In particular, our work grew from the discussion in HEP regarding theory and practice, and open problems such as how to efficiently construct Neyman confidence sets for general settings (Cowan et al., 2011), how to assess conditional coverage without costly Monte Carlo simulations (Cousins,

2018), and how to choose hybrid techniques in practice (Cousins, 2006).

Universal inference. Recently, Wasserman et al. (2020) proposed a "universal" inference test statistic for constructing valid confidence sets and hypothesis tests with finite-sample guarantees without regularity conditions. The assumptions are that the likelihood $\mathcal{L}(\mathcal{D};\theta)$ is known and that one can compute the maximum likelihood estimator (MLE). Our LFI framework does not require a tractable likelihood, but on the other hand our procedures implicitly assume that key quantities of interest vary smoothly in θ. Another main difference is that our approach estimates the critical value as a function of θ. The latter calibration step leads to more powerful tests than universal inference (see, for example, Section 7.2.1) but at the cost of having to simulate data from the likelihood.

Likelihood-free inference via machine learning. Recent LFI methods have been using the output from simulators as training data to learn a surrogate model for inference; see Cranmer et al. (2020) for a review. These LFI methods use synthetic data simulated across the parameter space to directly estimate key quantities, such as:

1. *posteriors* $p(\theta|\mathbf{x})$ (Marin et al., 2016; Papamakarios and Murray, 2016; Lueckmann et al., 2017; Greenberg et al., 2019; Chen and Gutmann, 2019; Izbicki et al., 2019; Radev et al., 2020),

2. *likelihoods* $p(\mathbf{x}|\theta)$ (Wood, 2010; Meeds and Welling, 2014; Wilkinson, 2014; Gutmann and Corander, 2016; Fasiolo et al., 2018; Lueckmann et al., 2019; Papamakarios et al., 2019; Picchini et al., 2020; Järvenpää et al., 2021), or

3. *density ratios*, such as the likelihood-to-marginal ratio $p(\mathbf{x}|\theta)/p(\mathbf{x})$ (Izbicki et al., 2014; Thomas et al., 2021; Hermans et al., 2020; Durkan et al., 2020), the likelihood ratio $p(\mathbf{x}|\theta_1)/p(\mathbf{x}|\theta_2)$ for $\theta_1, \theta_2 \in \Theta$ (Cranmer et al., 2015; Brehmer et al., 2020) or the profile-likelihood ratio (Heinrich, 2022).[2]

Undoubtedly, new ML methods, such as normalizing flows (Papamakarios et al., 2021) and other neural density estimators, have revolutionized the development in LFI in terms of sample efficiency and capacity, and will continue to do so. However, although the goal of LFI is inference on the unknown parameters θ, it remains an open question whether a given LFI algorithm produces reliable measures of uncertainty. A related question is how to objectively evaluate the power and coverage of various approaches when the true θ of interest is not known.

Our framework can lend good statistical properties and theoretical guarantees of nominal coverage to any LFI algorithm that computes a "test statistic"; that is, a measure of how well observed data fits the conjecture that the true parameter θ has a certain value θ_0. Furthermore, we provide diagnostic tools that assess whether constructed confidence sets and tests are valid for any value in the parameter space. Currently, no other LFI methodology can guarantee conditional validity and power in a finite-sample regime (without costly

2. `ACORE` and `BFF` are based on estimating the odds $\mathbb{O}(\mathbf{X};\theta)$ at $\theta \in \Theta$ (Equation 7); this is a "likelihood-to-marginal ratio" approach, which estimates a one-parameter function as in the original paper by Izbicki et al. (2014). The likelihood ratio $\mathbb{OR}(\mathbf{X};\theta_0,\theta_1)$ at $\theta_0, \theta_1 \in \Theta$ (Equation 9) is then computed from the odds function, without the need for an extra estimation step.

Monte Carlo samples at fixed parameter values), as well as provide practical diagnostics for assessing conditional coverage across the entire parameter space (when the true solution is not known).

2. Statistical Inference in a Traditional Setting

We review the Neyman construction of frequentist confidence sets, and the definitions of the likelihood ratio test statistic and the Bayes factor.

Equivalence of Tests and Confidence Sets. A classical approach to constructing a confidence set for an unknown parameter $\theta \in \Theta$ is to invert a series of hypothesis tests (Neyman, 1937): Suppose that for each possible value $\theta_0 \in \Theta$, there is a level α test δ_{θ_0} of

$$H_{0,\theta_0} : \theta = \theta_0 \quad \text{versus} \quad H_{1,\theta_0} : \theta \neq \theta_0; \tag{3}$$

that is, a test δ_{θ_0} where the type I error (the probability of erroneously rejecting a true null hypothesis H_{0,θ_0}) is no larger than α. For observed data $\mathcal{D} = D$, now define $R(D)$ as the set of all parameter values $\theta_0 \in \Theta$ for which the test δ_{θ_0} does not reject H_{0,θ_0}. Then, by construction, the random set $R(\mathcal{D})$ satisfies

$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in R(\mathcal{D}) \,|\, \theta\right) \geq 1 - \alpha,$$

for all $\theta \in \Theta$. That is, $R(\mathcal{D})$ defines a $(1-\alpha)$ *confidence set* for θ. Similarly, we can define tests with a desired significance level by inverting a confidence set with a certain coverage.

Likelihood Ratio Test. A general form of hypothesis tests that often leads to high power is the likelihood ratio test (LRT). Consider testing

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1, \tag{4}$$

where $\Theta_1 = \Theta \setminus \Theta_0$. For the *likelihood ratio (LR) statistic*,

$$\mathrm{LR}(\mathcal{D}; \Theta_0) = \log \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\mathcal{D}; \theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\mathcal{D}; \theta)}, \tag{5}$$

the LRT of hypotheses (4) rejects H_0 when $\mathrm{LR}(D; \Theta_0) < C$ for some constant C.

Figure 2 illustrates the construction of confidence sets for θ from level α likelihood ratio tests (3). The critical value for each such test δ_{θ_0} is $C_{\theta_0} = \sup\left\{C : \mathbb{P}\left(\mathrm{LR}(\mathcal{D}; \theta_0) < C \mid \theta = \theta_0\right) \leq \alpha\right\}$.

Bayes Factor. Let π be a probability measure over the parameter space Θ. The Bayes factor (Jeffreys, 1935, 1961) for comparing the hypothesis $H_0 : \theta \in \Theta_0$ to its complement, the alternative H_1, is a ratio of the marginal likelihood of the two hypotheses:

$$\mathrm{BF}(\mathcal{D}; \Theta_0) \equiv \frac{\mathbb{P}(\mathcal{D}|H_0)}{\mathbb{P}(\mathcal{D}|H_1)} = \frac{\int_{\Theta_0} \mathcal{L}(\mathcal{D}; \theta) d\pi_0(\theta)}{\int_{\Theta_1} \mathcal{L}(\mathcal{D}; \theta) d\pi_1(\theta)}, \tag{6}$$

where π_0 and π_1 are the restrictions of π to the parameter regions Θ_0 and $\Theta_1 = \Theta_0^c$, respectively. The Bayes factor is often used as a Bayesian alternative to significance testing, as it quantifies the change in the odds in favor of H_0 when going from the prior to the posterior: $\frac{\mathbb{P}(H_0|\mathcal{D})}{\mathbb{P}(H_1|\mathcal{D})} = \mathrm{BF}(\mathcal{D}; \Theta_0) \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}$.

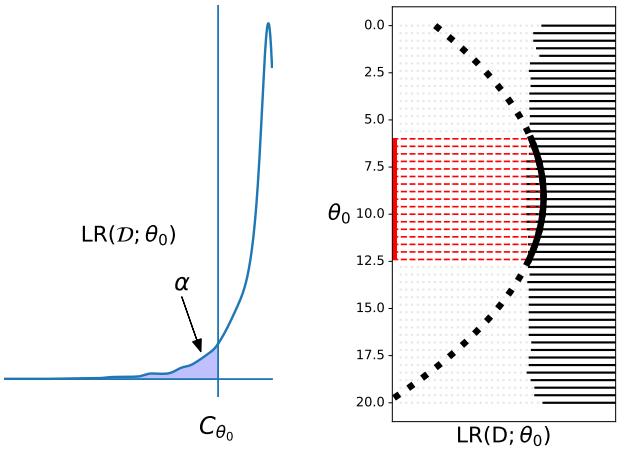


Figure 2: Constructing confidence intervals from hypothesis tests. *Left:* For each $\theta_0 \in \Theta$, we find the critical value C_{θ_0} that rejects the null hypothesis H_{0,θ_0} at level α; that is, C_{θ_0} is the α-quantile of the distribution of the likelihood ratio statistic $\text{LR}(\mathcal{D}; \theta_0)$ under the null. *Right:* The horizontal lines represent the acceptance region for each $\theta_0 \in \Theta$. Suppose we observe data $\mathcal{D} = D$. The confidence set for θ (indicated with the red line) consists of all θ_0-values for which the observed test statistic $\text{LR}(D; \theta_0)$ (indicated with the black curve) falls in the acceptance region.

3. Likelihood-Free Frequentist Inference

In the typical LFI setting, we cannot directly evaluate the likelihood ratio $\text{LR}(\mathcal{D}; \Theta_0)$, or even the likelihood $\mathcal{L}(\mathcal{D}; \theta)$. A simulator-based approach with, for example, `ACORE` or `BFF` can nevertheless lead to hypothesis tests and confidence sets with good frequentist properties. We assume we have access to: (i) a "high-fidelity" forward simulator, also denoted by F_θ, that can simulate observable data, (ii) a reference distribution G, which does not depend on θ, with larger support than F_θ for all $\theta \in \Theta$, and (iii) a probabilistic classifier that discriminates samples from F_θ and G.

3.1 Odds Function via Probabilistic Classification

We start by generating a labeled sample $\mathcal{T} = \{(\theta_i, \mathbf{X}_i, Y_i)\}_{i=1}^B$ to compare data from the simulator F_θ with data from the reference distribution G. Here, $\theta \sim \pi_\Theta$ (a fixed proposal distribution over Θ), the "label" $Y \sim \text{Ber}(p)$, $\mathbf{X}|\theta, Y = 1 \sim F_\theta$, and $\mathbf{X}|\theta, Y = 0 \sim G$. We then define the odds at θ and fixed $\mathbf{x}$ as

$$\mathbb{O}(\mathbf{x}; \theta) := \frac{\mathbb{P}(Y = 1|\theta, \mathbf{x})}{\mathbb{P}(Y = 0|\theta, \mathbf{x})}. \tag{7}$$

One way of interpreting the odds $\mathbb{O}(\mathbf{x}; \theta)$ is to regard it as a measure of the chance that $\mathbf{x}$ was generated from F_θ rather than from G. That is, a large odds $\mathbb{O}(\mathbf{x}; \theta)$ reflects the fact that it is plausible that $\mathbf{x}$ was generated from F_θ (instead of G). We call G a "reference distribution" as we are comparing F_θ for different θ with this distribution.

The odds function $\mathbb{O}(\mathbf{X};\theta)$ with $\theta \in \Theta$ as a parameter can be estimated with a probabilistic classifier, such as a neural network with a softmax layer, suitable for the data $\mathbf{X}$ at hand. Algorithm 3 in Appendix A summarizes our procedure for simulating a labeled sample $\mathcal{T}$ for estimating odds. For all experiments in this paper, we use p=1/2 and $G = F_{\mathbf{X}}$, where $F_{\mathbf{X}}$ is the marginal distribution of F_θ with respect to π_Θ.

3.2 Test Statistics Based on Odds Function

For testing $H_{0,\Theta_0} : \theta \in \Theta_0$ versus all alternatives $H_{1,\Theta_0} : \theta \notin \Theta_0$, we consider two test statistics: `ACORE` and `BFF`. Both statistics are based on $\mathbb{O}(\mathbf{X};\theta)$, but whereas `ACORE` eliminates the parameter θ by maximization, `BFF` averages over the parameter space.

3.2.1 `ACORE` by Maximization

The `ACORE` statistic (Anonymized, 2020) for testing Equation 3 is given by

$$\begin{aligned}\Lambda(\mathcal{D};\Theta_0) &:= \log \frac{\sup_{\theta\in\Theta_0}\prod_{i=1}^n \mathbb{O}(\mathbf{X}_i;\theta)}{\sup_{\theta\in\Theta}\prod_{i=1}^n \mathbb{O}(\mathbf{X}_i;\theta)} \\ &= \sup_{\theta_0\in\Theta_0}\inf_{\theta_1\in\Theta}\sum_{i=1}^n \log\left(\mathbb{OR}(\mathbf{X}_i;\theta_0,\theta_1)\right), \end{aligned} \tag{8}$$

where the odds ratio

$$\mathbb{OR}(\mathbf{x};\theta_0,\theta_1) := \frac{\mathbb{O}(\mathbf{x};\theta_0)}{\mathbb{O}(\mathbf{x};\theta_1)} \tag{9}$$

at $\theta_0, \theta_1 \in \Theta$ measures the plausibility that a fixed $\mathbf{x}$ was generated from θ_0 rather than θ_1.

We use $\widehat{\Lambda}(\mathcal{D};\Theta_0)$ to denote the `ACORE` statistic based on $\mathcal{T}$ and estimated odds $\widehat{\mathbb{O}}(\mathbf{X};\theta_0)$. When $\widehat{\mathbb{O}}(\mathbf{X};\theta_0)$ is well-estimated for every θ and $\mathbf{X}$, the estimated `ACORE` statistic $\widehat{\Lambda}(\mathcal{D};\Theta_0)$ is the same as the likelihood ratio statistic $\mathrm{LR}(\mathcal{D};\Theta_0)$ (Anonymized 2020; Proposition 3.1).

3.2.2 `BFF` by Averaging

Because the `ACORE` statistics in Equation 8 involves taking the supremum (or infimum) over Θ, it may not be practical in high dimensions. Hence, in this work, we propose an alternative statistic for testing (3) based on averaged odds:

$$\tau(\mathcal{D};\Theta_0) := \frac{\int_{\Theta_0}\prod_{i=1}^n \mathbb{O}(\mathbf{X}_i;\theta_0)d\pi_0(\theta)}{\int_{\Theta_0^c}\prod_{i=1}^n \mathbb{O}(\mathbf{X}_i;\theta)d\pi_1(\theta)}, \tag{10}$$

where π_0 and π_1 are the restrictions of the proposal distribution π to the parameter regions Θ_0 and Θ_0^c, respectively.

Let $\widehat{\tau}(\mathcal{D};\Theta_0)$ denote estimates based on $\mathcal{T}$ and $\widehat{\mathbb{O}}(\theta_0;\mathbf{x})$. If the probabilities learned by the classifier are well estimated, then the estimated averaged odds statistic $\widehat{\tau}(\mathcal{D};\Theta_0)$ is exactly the Bayes factor:

Proposition 1 (Fisher consistency)

Assume that, for every $\theta \in \Theta$, G dominates F_θ. If $\widehat{\mathbb{P}}(Y = 1|\theta,\mathbf{x}) = \mathbb{P}(Y = 1|\theta,\mathbf{x})$ for every θ and $\mathbf{x}$, then $\widehat{\tau}(\mathcal{D};\Theta_0)$ is the Bayes factor $BF(\mathcal{D};\Theta_0)$.

In this paper, we are using the Bayes factor as a frequentist test statistic. Hence, our term *Bayes Frequentist Factor (BFF)* statistic for τ and $\widehat{\tau}$.

3.3 Fast Construction of Neyman Confidence Set

Instead of a costly MC or bootstrap hypothesis test of $H_0 : \theta = \theta_0$ at each θ_0 on a fine grid (see, e.g., MacKinnon 2009 and Ventura 2010 for reference), we generate just one sample $\mathcal{T}'$ of moderate size B'. We then estimate either the critical value C_{θ_0} via quantile regression (Section 3.3.1), or the p-value $p(D;\theta_0)$ via regression (Section 3.3.2), for all $\theta_0 \in \Theta$ simultaneously. In Section 5, we propose a practical strategy for how to choose the number of simulations B' and regression method for the application at hand.

3.3.1 The Critical Value via Quantile Regression

Algorithm 1 describes how we can use quantile regression (e.g., Meinshausen 2006; Koenker et al. 2017) to estimate how the critical value C_{θ_0} for a level-α test of (3) varies with $\theta_0 \in \Theta$. To test a composite null hypothesis $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, we use the cutoff $\widehat{C}_{\Theta_0} := \inf_{\theta\in\Theta_0} \widehat{C}_\theta$. Although we originally proposed the calibration procedure for `ACORE`, the same scheme leads to a valid test (control of type I error as the number of simulations $B' \to \infty$) for *any* test statistic λ (Theorem 8). In particular, we control the type I error for `BFF` and `ACORE`, even if the estimated statistics are not good approximations of the Bayes Factor and the LR statistic, respectively.

Algorithm 1 Estimate the critical values C_{θ_0} for a level-α test of $H_{0,\theta_0} : \theta = \theta_0$ vs. $H_{1,\theta_0} : \theta \neq \theta_0$ for all $\theta_0 \in \Theta$ simultaneously

Require: stochastic forward simulator F_θ; sample size B' for training quantile regression estimator; π (a fixed proposal distribution over the parameter space Θ); test statistic λ; quantile regression estimator; desired level $\alpha \in (0,1)$

Ensure: estimated critical values $\widehat{C}_{\theta_0}$ for all $\theta_0 \in \Theta$

1: Set $\mathcal{T}' \leftarrow \emptyset$
2: **for** i in $\{1,\ldots,B'\}$ **do**
3: Draw parameter $\theta_i \sim \pi_\Theta$
4: Draw sample $\mathbf{X}_{i,1},\ldots,\mathbf{X}_{i,n} \overset{iid}{\sim} F_{\theta_i}$
5: Compute test statistic $\lambda_i \leftarrow \lambda((\mathbf{X}_{i,1},\ldots,\mathbf{X}_{i,n});\theta_i)$
6: $\mathcal{T}' \leftarrow \mathcal{T}' \cup \{(\theta_i,\lambda_i)\}$
7: **end for**
8: Use $\mathcal{T}'$ to learn the conditional quantile function $\widehat{C}_\theta := \widehat{F}^{-1}_{\lambda|\theta}(\alpha|\theta)$ via quantile regression of λ on θ
9: **return** $\widehat{C}_{\theta_0}$

3.3.2 The P-Value via Regression

If the data D are observed beforehand, then given any test statistic λ, we can alternatively compute p-values for each hypothesis $H_{0,\theta_0} : \theta = \theta_0$, that is,

$$p(D;\theta_0) := \mathbb{P}_{\mathcal{D}|\theta}\left(\lambda(\mathcal{D};\theta) < \lambda(D;\theta) \,|\, \theta = \theta_0\right). \tag{11}$$

The p-value $p(D;\theta_0)$ can be used to test hypothesis and create confidence sets for any desired level α. We can estimate these p-values by noticing that they are defined by a regression $\mathbb{E}[Z|\theta]$ of the random variable $Z := \mathbb{I}(\lambda(\mathcal{D};\theta) < \lambda(D;\theta))$ on θ, evaluated at θ_0. Thus, we can as in Algorithm 5 of Appendix B generate a training sample $\mathcal{T}' = \{(Z_1,\theta_1),\ldots,(Z_{B'},\theta_{B'})\}$ and then estimate p-values for all $\theta \in \Theta$ simultaneously. For testing the composite null hypothesis $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, we use

$$\widehat{p}(D;\Theta_0) := \sup_{\theta\in\Theta_0} \widehat{p}(D;\theta),$$

where $\widehat{p}(D;\theta)$ is defined as in Equation 11.

Note that there is a key computational difference between estimating p-values versus estimating critical values: the LF2I conditional quantile from Algorithm 1 is a function of the parameter θ only. Hence, we can construct confidence sets for different realizations D of $\mathcal{D}$ without retraining, as is evident from Equation 12 below. The p-value, on the other hand, is a function of both θ and of the observed sample D itself. As a result, Algorithm 5 has to be repeated for each observed D, making the computation of p-values more expensive.

3.3.3 Amortized Confidence Sets

Finally, we construct an approximate confidence region for θ by taking the set

$$\widehat{R}(D) = \left\{\theta \in \Theta \,\middle|\, \lambda(D;\theta) \geq \widehat{C}_\theta\right\}, \tag{12}$$

or alternatively,

$$\widehat{R}(D) = \{\theta \in \Theta \,|\, \widehat{p}(D;\theta) > \alpha\}\,; \tag{13}$$

see Algorithm 6 in Appendix C for the details. As shown in Anonymized (2020, Theorem 3.3), the random set $\widehat{R}(\mathcal{D})$ has nominal $1-\alpha$ coverage as $B' \to \infty$ regardless of the data sample size n.

As noted in Section 3.3.2, the confidence set in Equation 12 is fully *amortized*, meaning that once we have the test statistic $\lambda(\mathcal{D};\theta)$ and $\widehat{C}_\theta$ as a function of $\theta \in \Theta$, we can perform inference on new data without retraining.

3.4 Diagnostics: Conditional Coverage of Confidence Set via Regression

Our LF2I framework has a separate module ("Diagnostics" in Figure 1) for evaluating "local" goodness-of-fit in different regions of the parameter space Θ. This third module estimates the *conditional* coverage $\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D})\middle|\theta\right)$ of the confidence set $\widehat{R}(\mathcal{D})$ via regression. The procedure (see Algorithm 2) generates a new sample of size B'' from the simulator: $\mathcal{T}'' = \{(\theta'_1,\mathcal{D}'_1),\ldots,(\theta'_{B''},\mathcal{D}'_{B''})\}$, where $\theta'_i \sim \pi_\Theta$ and $\mathcal{D}'_i := \{\mathbf{X}_{i,1},\ldots,\mathbf{X}_{i,n}\} \overset{\text{iid}}{\sim} F_{\theta'_i}$. For each sample $\mathcal{D}'_i$, we check whether or not the test statistic λ_i is larger than the estimated critical value $\widehat{C}_{\theta'_i}$ (the output from Algorithm 1); this is equivalent to computing a binary variable W_i for whether or not the "true" value θ'_i falls within the confidence set $\widehat{R}(\mathcal{D}'_i)$ (Equation 12). Recall that the computations of the test statistic and the critical value are amortized, meaning that we do not retrain algorithms to estimate these two quantities. The

final step is to estimate empirical coverage as a function of θ by regressing W on $\theta^{'}$. This estimation requires a new fit, but after training the probabilistic classifier, we can evaluate the estimated coverage anywhere in parameter space Θ.

Our diagnostic procedure can locate regions in parameter space where estimated confidence sets might under- or over-cover; see Figure 3 and Figure 9 for examples. Note that standard goodness-of-fit techniques for conditional densities (Cook et al., 2006; Bordoloi et al., 2010; Talts et al., 2018; Schmidt et al., 2020) only check for marginal coverage over Θ.

Algorithm 2 Estimate the empirical coverage $\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D})\middle|\theta\right)$, for all $\theta \in \Theta$ simultaneously.

Require: stochastic forward simulator F_θ; π_Θ (a fixed proposal distribution over the full parameter space Θ); sample size B'' for estimating coverage; test statistic λ; nominal coverage α; estimated critical values $\widehat{C}_\theta$; regression estimator m

Ensure: estimated coverage $\widehat{\mathbb{P}}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D})\middle|\theta\right)$ for all $\theta \in \Theta$, where $\widehat{R}$ is the region derived from λ and $\widehat{C}_\theta$ (Equation 12)

Set $\mathcal{T}'' \leftarrow \emptyset$
for i in $\{1, \ldots, B^{''}\}$ **do**
  Draw parameter $\theta_i' \sim \pi_\Theta$
  Draw sample $\mathcal{D}_i^{'} := \{\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n}\} \overset{iid}{\sim} F_{\theta_i'}$
  Compute test statistic $\lambda_i \leftarrow \lambda(\mathcal{D}_i^{'}; \theta_i^{'})$
  Compute indicator variable $W_i \leftarrow \mathbb{I}\left(\lambda_i \geq \widehat{C}_{\theta_i'}\right)$
  $\mathcal{T}'' \leftarrow \mathcal{T}'' \cup \{(\theta_i', W_i)\}$
end for
Use $\mathcal{T}''$ to learn the conditional coverage $\widehat{\mathbb{P}}_{\mathcal{D}|\theta'}\left(\theta' \in \widehat{R}(\mathcal{D})\middle|\theta'\right)$ by regressing W on θ'

return $\widehat{\mathbb{P}}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D})\middle|\theta\right)$

3.5 Loss Functions

In this work, we use the cross-entropy loss to train probabilistic classifiers. Consider a sample point $\{\theta, \mathbf{x}, y\}$ generated according to Algorithm 3. Let p be a Ber(y) distribution, and q be a Ber $\left(\widehat{\mathbb{P}}(Y = 1|\theta, \mathbf{x})\right) = \text{Ber}\left(\frac{\widehat{\mathbb{O}}(\mathbf{x};\theta)}{1+\widehat{\mathbb{O}}(\mathbf{x};\theta)}\right)$ distribution. The *cross-entropy* between p and q is given by

$$\begin{aligned}\mathcal{L}_{\text{CE}}(\widehat{\mathbb{O}}; \{\theta, \mathbf{x}, y\}) &= -y \log\left(\frac{\widehat{\mathbb{O}}(\mathbf{x};\theta)}{1+\widehat{\mathbb{O}}(\mathbf{x};\theta)}\right) - (1-y)\log\left(\frac{1}{1+\widehat{\mathbb{O}}(\mathbf{x};\theta)}\right) \\ &= -y \log\left(\widehat{\mathbb{O}}(\mathbf{x};\theta)\right) + \log\left(1+\widehat{\mathbb{O}}(\mathbf{x};\theta)\right). \end{aligned} \tag{14}$$

For every $\mathbf{x}$ and θ, the expected cross-entropy $\mathbb{E}[L_{\text{CE}}(\widehat{\mathbb{O}}; \{\theta, \mathbf{x}, y\})]$ is minimized by $\widehat{\mathbb{O}}(\mathbf{x};\theta) = \mathbb{O}(\mathbf{x};\theta)$. If the probabilistic classifier attains the minimum of the cross-entropy loss, then as

shown in Anonymized 2020, the estimated `ACORE` statistic $\widehat{\Lambda}(\mathcal{D};\Theta_0)$ will be equal to the likelihood ratio statistic in Equation 5. Similarly, as stated in Proposition 1, at the minimum, the estimated `BFF` statistic $\widehat{\tau}(\mathcal{D};\Theta_0)$ is equal to the Bayes factor in Equation 6.

In the special case where G is the marginal distribution of $F_\theta(\mathbf{x})$ with respect to π, and when in addition $\mathbf{x}$ contains all observations (that is, $\mathbf{X} = \mathcal{D}$), the denominator of the `BFF` statistic in Equation 10 is equal to one, as shown in Equation 16. The `BFF` test statistic then simply becomes the integrated odds. Hence, in addition to the standard cross-entropy loss (Equation 14), we propose an *integrated odds loss* function which is directly related to the BFF (integrated odds) statistic:

$$\mathcal{L}(\widehat{\mathbb{O}},\mathbb{O}) := \int \left(\widehat{\mathbb{O}}(\mathbf{x};\theta) - \mathbb{O}(\mathbf{x};\theta)\right)^2 dG(\mathbf{x})d\pi(\theta). \tag{15}$$

In Section 4, Theorem 5, we show that the power of the `BFF` test statistic is bounded by the integrated odds loss.

4. Theoretical Guarantees

Next, we prove consistency of the critical value and the p-value estimation methods (Algorithms 1 and 5, respectively), and provide theoretical guarantees for the power of `BFF`. We refer the reader to Appendix D for a proof for finite Θ that the power of the `ACORE` test converges to the power of the LRT as B grows (Theorem 7).

In this section, $\mathbb{P}_{\mathcal{D},\mathcal{T}'|\theta}$ denotes the probability integrated over both $\mathcal{D} \sim F_\theta$ and $\mathcal{T}'$, whereas $\mathbb{P}_{\mathcal{D}|\theta}$ denotes integration over $\mathcal{D} \sim F_\theta$ only. For notational ease, we do not explicitly state again (inside the parentheses of the same expression) that we condition on θ.

4.1 Critical Value Estimation

We start by showing that our procedure for choosing critical values leads to valid hypothesis tests (that is, tests that control the type I error probability), as long as the number of simulations B' in Algorithm 1 is sufficiently large. We assume that the null hypothesis is simple, that is, $\Theta_0 = \{\theta_0\}$ — which is the relevant setting for the Neyman construction of confidence sets in the absence of nuisance parameters. See Theorem 8 in Appendix E for results for composite null hypotheses.

We assume that the quantile regression estimator described in Section 3.3.1 is consistent in the following sense:

Assumption 1 (Uniform consistency) *Let $F(\cdot|\theta)$ be the cumulative distribution function of the test statistic $\lambda(\mathcal{D};\theta_0)$ conditional on θ, where $\mathcal{D} \sim F_\theta$. Let $\widehat{F}_{B'}(\cdot|\theta)$ be the estimated conditional distribution function, implied by a quantile regression with a sample $\mathcal{T}'$ of B' simulations $\mathcal{D} \sim F_\theta$. Assume that the quantile regression estimator is such that*

$$\sup_{\lambda\in\mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta_0) - F(\lambda|\theta_0)| \xrightarrow[B'\longrightarrow\infty]{\mathbb{P}} 0.$$

Assumption 1 holds, for instance, for quantile regression forests (Meinshausen, 2006).

Next, we show that Algorithm 1 yields a valid hypothesis test as $B' \to \infty$.

Theorem 1 *Let $C_{B'} \in \mathbb{R}$ be the critical value of the test based on a strictly continuous statistic $\lambda(\mathcal{D}; \theta_0)$ chosen according to Algorithm 1 for a fixed $\alpha \in (0, 1)$. If the quantile estimator satisfies Assumption 1, then,*

$$\mathbb{P}_{\mathcal{D}|\theta_0, C_{B'}}(\lambda(\mathcal{D}; \theta_0) \leq C_{B'}) \xrightarrow[B' \longrightarrow \infty]{a.s.} \alpha,$$

where $\mathbb{P}_{\mathcal{D}|\theta_0, C_{B'}}$ denotes the probability integrated over $\mathcal{D} \sim F_{\theta_0}$ and conditional on the random variable $C_{B'}$.

If the convergence rate of the quantile regression estimator is known (Assumption 2), Theorem 2 provides a finite-B' guarantee on how far the type I error of the test will be from the nominal level.

Assumption 2 (Convergence rate of the quantile regression estimator) *Using the notation of Assumption 1, assume that the quantile regression estimator is such that*

$$\sup_{\lambda \in \mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta_0) - F(\lambda|\theta_0)| = O_P\left(\left(\frac{1}{B'}\right)^r\right)$$

for some $r > 0$.

Theorem 2 *With the notation and assumptions of Theorem 1, and if Assumption 2 also holds, then,*

$$|\mathbb{P}_{\mathcal{D}|\theta_0, C_{B'}}(\lambda(\mathcal{D}; \theta_0) \leq C_{B'}) - \alpha| = O_P\left(\left(\frac{1}{B'}\right)^r\right).$$

4.2 P-Value Estimation

We start by showing that the p-value estimation method described in Section 3.3.2 is consistent. The results shown here apply to any test statistic λ. That is, these results are not restricted to `BFF`.

We assume consistency in the sup norm of the regression method used to estimate the p-values:

Assumption 3 (Uniform consistency) *The regression estimator used in Equation 11 is such that*

$$\sup_{\theta} |\widehat{\mathbb{E}}_{B'}[Z|\theta] - \mathbb{E}[Z|\theta]| \xrightarrow[B' \longrightarrow \infty]{a.s.} 0.$$

Examples of estimators that satisfy Assumption 3 include Bierens (1983); Hardle et al. (1984); Liero (1989); Girard et al. (2014).

The next theorem shows that the p-values obtained according to Algorithm 5 converge to the true p-values. Moreover, the power of the tests obtained using the estimated p-values converges to the power one would obtain if the true p-values could be computed.

Theorem 3 *Under Assumption 3 and if $p(\mathcal{D};\Theta_0)$ is a strictly continuous random variable then, for every $\theta \in \Theta$,*

$$\widehat{p}(D;\Theta_0) \xrightarrow[B'\longrightarrow\infty]{a.s.} p(D;\Theta_0)$$

and

$$\mathbb{P}_{\mathcal{D},\mathcal{T}'|\theta}(\widehat{p}(\mathcal{D};\Theta_0) \leq \alpha) \xrightarrow{B'\longrightarrow\infty} \mathbb{P}_{\mathcal{D}|\theta}(p(\mathcal{D};\Theta_0) \leq \alpha).$$

The next corollary shows that as $B' \longrightarrow \infty$, the tests obtained using the p-values from Algorithm 5 have size α.

Corollary 1 *Under Assumption 3 and if F_θ is continuous for every $\theta \in \Theta$ and $p(\mathcal{D};\Theta_0)$ is a strictly continuous random variable, then*

$$\sup_{\theta\in\Theta_0} \mathbb{P}_{\mathcal{D},\mathcal{T}'|\theta}(\widehat{p}(\mathcal{D};\Theta_0) \leq \alpha) \xrightarrow{B'\longrightarrow\infty} \alpha.$$

Under stronger assumptions about the regression method, it is also possible to derive rates of convergence for the estimated p-values.

Assumption 4 (Convergence rate of the regression estimator) *The regression estimator is such that*

$$\sup_{\theta} |\widehat{\mathbb{E}}[Z|\theta] - \mathbb{E}[Z|\theta]| = O_P\left(\left(\frac{1}{B'}\right)^r\right).$$

for some $r > 0$.

Examples of regression estimators that satisfy Assumption 4 can be found in Stone (1982); Hardle et al. (1984); Donoho (1994); Yang et al. (2017).

Theorem 4 *Under Assumption 4,*

$$|p(D;\Theta_0) - \widehat{p}(D;\Theta_0)| = O_P\left(\left(\frac{1}{B'}\right)^r\right).$$

4.3 Power of `BFF`

In this section, we provide convergence rates for `BFF` and show that its power relates to the loss function of Equation 15. We assume we are testing a simple hypothesis $H_{0,\theta_0} : \theta = \theta_0$, where θ_0 is fixed.

We assume that $G(\mathbf{x})$ is the marginal distribution of $F_\theta(\mathbf{x})$ with respect to $\pi(\theta)$. We here also assume that $\mathbf{x}$ contains all observations; that is, $\mathbf{X} = \mathcal{D}$. In this case, the denominator of the average odds is

$$\int_\Theta \mathbb{O}(\mathbf{x},\theta)d\pi(\theta) = \int_{\Theta_1} \frac{f(\mathbf{x}|\theta)}{g(\mathbf{x})}d\pi(\theta) = \int_\Theta \frac{f(\mathbf{x}|\theta)}{\int_\Theta f(\mathbf{x}|\theta)d\pi(\theta)}d\pi(\theta) = 1, \tag{16}$$

and therefore there is no need to estimate the denominator in Equation 10.

We also make the following assumptions:

Assumption 5 (Bounded odds and estimated odds) *There exists* $0 < M, m < \infty$ *such that for every* $\theta \in \Theta$ *and* $\mathbf{x} \in \mathcal{X}$, $m \leq \mathbb{O}(\mathbf{x};\theta), \widehat{\mathbb{O}}(\mathbf{x};\theta) \leq M$.

Assumption 6 (Bounded second moment of odds estimation error) *Let*

$$h(\theta) = \int (\mathbb{O}(\mathbf{x};\theta) - \widehat{\mathbb{O}}(\mathbf{x};\theta))^2 dG(\mathbf{x}).$$

There exists $M', m' > 0$ *such that* $h(\theta) \leq M'$ *and* $\int h(\theta) d\pi(\theta) > m'$.

Assumption 5 states that the odds and estimated odds are both bounded away from 0 and infinity, for all choice of parameters θ and features $\mathbf{x}$. Assumption 6 states that the second moment of the difference between the true and estimated odds is bounded away from 0 and infinity.

Finally, we assume that the CDF of the power function of the test based on the `BFF` statistic τ in Equation 10 is smooth in a Lipschitz sense:

Assumption 7 (Smooth power function) *The cumulative distribution function of* $\tau(\mathcal{D};\theta_0)$, F_τ, *is Lipschitz with constant* C_L, *i.e., for every* $x_1, x_2 \in \mathbb{R}$, $|F_\tau(x_1) - F_\tau(x_2)| \leq C_L |x_1 - x_2|$.

With these assumptions, we can relate the odds loss with the probability that the outcome of `BFF` is different from the outcome of the test based on the Bayes factor:

Theorem 5 *Let* $\phi_\tau(\mathcal{D}) = \mathbb{I}(\tau(\mathcal{D};\theta_0) < c)$ *and* $\phi_{\widehat{\tau}_B}(\mathcal{D}) = \mathbb{I}(\widehat{\tau}_B(\mathcal{D};\theta_0) < c)$ *be the testing procedures for testing* $H_{0,\theta_0} : \theta = \theta_0$ *obtained using* τ *and* $\widehat{\tau}_B$. *Under Assumptions 5-7, there exists* $K' > 0$ *such that, for every* $0 < \epsilon < 1$,

$$\mathbb{P}_{\mathcal{D}|\theta,T}(\phi_\tau(\mathcal{D}) \neq \phi_{\widehat{\tau}_B}(\mathcal{D})) \leq \frac{K' \cdot \sqrt{L(\widehat{\mathbb{O}}, \mathbb{O})}}{\epsilon} + \epsilon,$$

where T *denotes the realized training sample* $\mathcal{T}$ *and* $\mathbb{P}_{\mathcal{D}|\theta,T}$ *is the probability measure integrated over the observable data* $\mathcal{D} \sim F_\theta$, *but conditional on the train sample used to create the test statistic. Notice that, by construction,* $\mathcal{D}$ *is independent of* $\mathcal{T}$.

Theorem 5 demonstrates that the probability that hypothesis tests based on the `BFF` statistic versus the Bayes factor lead to different conclusions is bounded by the integrated odds loss. This result is valuable because the integrated odds loss is easy to estimate in practice, and hence provides us with a practically useful metric. For instance, the integrated odds loss can serve as a natural criterion for selecting the "best" statistical model out of a set of candidate models with different classifiers, for tuning model hyperparameters, and for evaluating model fit.

Next, we provide rates of convergence of the test based on `BFF` to the test based on the Bayes factor. We assume that the chosen probabilistic classifier has the following rate of convergence:

Assumption 8 (Convergence rate of the probabilistic classifier) *The probabilistic classifier trained with* $\mathcal{T}$, $\widehat{\mathbb{P}}(Y = 1|\mathbf{x},\theta)$ *is such that*

$$\mathbb{E}_{\mathcal{T}}\left[\int \left(\widehat{\mathbb{P}}(Y = 1|\mathbf{x},\theta) - \mathbb{P}(Y = 1|\mathbf{x},\theta)\right)^2 dH(\mathbf{x},\theta)\right] = O\left(B^{-\kappa/(\kappa+d)}\right),$$

for some $\kappa > 0$ *and* $d > 0$, *where* $H(\mathbf{x},\theta)$ *is a measure over* $\mathcal{X} \times \Theta$.

Typically, κ relates to the smoothness of $\mathbb{P}$, while d relates to the number of covariates of the classifier — in our case, the number of parameters plus the number of features. Below, we provide some examples where Assumption 8 holds, using well-established results for the convergence rates of commonly used regression estimators:

- Kpotufe (2011) shows that kNN estimators are adaptive to the intrinsic dimension d under certain conditions. When $\widehat{\mathbb{P}}$ is a kNN estimator with $\mathbb{P}$ in a class of Lipschitz continuous functions, Assumption 8 holds with $\kappa = 2$. More generally, with $\mathbb{P}$ in a Hőlder space with parameter $0 < \beta \leq 1.5$, Assumption 8 holds with $\kappa = 2\beta$ (Győrfi et al. 2006; Ayano 2012).
- Kpotufe and Garg (2013) show that under certain conditions, when $\widehat{\mathbb{P}}$ is a kernel regression estimator with $\mathbb{P}$ in a class of Lipschitz continuous functions, Assumption 8 holds with $\kappa = 2$ and d the intrinsic dimension of the data. More generally, with $\mathbb{P}$ in a Hőlder space with parameter $0 < \beta \leq 1.5$, Assumption 8 holds with $\kappa = 2\beta$ (Győrfi et al. 2006).
- When $\widehat{\mathbb{P}}$ is a local polynomial regression estimator with $\mathbb{P}$ in a Sobolev space with smoothness β, Assumption 8 holds with $\kappa = \beta$, where d is the manifold dimension (Bickel and Li 2007).
- Biau (2012) shows that under certain conditions, when $\widehat{\mathbb{P}}$ is a random forest estimator with D covariates with $\mathbb{P}$ in a class of Lipschitz continuous functions, Assumption 8 holds with $\kappa = 2$ when the number of relevant features $d \leq D/2$.

More examples can be found in Győrfi et al. (2006), Tsybakov (2009) and Devroye et al. (2013).

We also assume that the density of the product measure $G \times \pi$ is bounded away from infinity.

Assumption 9 (Bounded density) *$H(\mathbf{x}, \theta)$ dominates $H' := G \times \pi$, and the density of H' with respect to H, denoted by h', is such that there exists $\gamma > 0$ with $h'(\mathbf{x}, \theta) < \gamma$, $\forall \mathbf{x} \in \mathcal{X}, \theta \in \Theta$.*

If the probabilistic classifier has the convergence rate given by Assumption 8, then the probability that hypothesis tests based on the `BFF` statistic versus the Bayes factor goes to zero has the rate given by the following theorem.

Theorem 6 *Let $\widehat{\tau}_B$ and τ be as in Theorem 5. Under Assumptions 5-9, there exists $K'' > 0$ such that, for any point in the alternative hypothesis (that is, $\theta \neq \theta_0$),*

$$\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_\tau(\mathcal{D}) \neq \phi_{\widehat{\tau}_B}(\mathcal{D})) \leq 2\sqrt{K''}B^{-\kappa/(4(\kappa+d))}.$$

Corollary 2 tells us that the power of the `BFF` test is close to the power of the exact Bayes factor test. This implies that `BFF` converges to the most powerful test in the Neyman-Person setting, where Bayes factor test is equivalent to the LRT.

Corollary 2 *Under Assumptions 5-9, there exists $K'' > 0$ such that, for any point in the alternative hypothesis (that is, $\theta \neq \theta_0$),*

$$\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\widehat{\tau}_B}(\mathcal{D}) = 1) \geq \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_\tau(\mathcal{D}) = 1) - 2\sqrt{K''}B^{-\kappa/(4(\kappa+d))}.$$

5. Sources of Error in LF2I Confidence Sets. Practical Strategy for Model Selection.

In traditional statistical inference, confidence sets depend on the choice of test statistic, the assumed distribution of the test statistic under the null, and the amount of available data. In LFI, however, there are additional sources of errors. For our LF2I inference machinery, we categorize these errors as follows:

e_1: Estimation error in learning the odds (Section 3.1);

e_2: Numerical error in evaluating the test statistic by maximization in `ACORE` (Equation 8) or by integration in `BFF` (Equation 10);

e_3: Estimation error in learning the critical values (Section 3.3.1) or the p-values (Section 3.3.2).

Validity and power. *Validity* is directly determined by e_3. As shown in Section 4, one can construct valid confidence sets regardless of how well the test statistic is estimated, as long as the quantile regressor (Algorithm 1) or probabilistic classifier for estimating p-values (Algorithm 5) is consistent and the training sample size B' is large enough. The *power or expected size* of the confidence set is, on the other hand, determined by both e_1 and e_2. The error e_1 depends on the capacity of the classifier for estimating odds and the training sample size B. The error e_2 is a purely numerical error and can be reduced by increasing the computational budget. Note that in the examples in Section 7, we are not employing any particular optimization or importance-weighted integration technique; we are simply generating a uniform grid over the parameter space and then computing the maximum or sum of relevant quantities over the grid points to evaluate the `ACORE` or `BFF` statistics, respectively.

Practical Strategy. To mitigate all sources of errors, we choose learning methods, number of simulations, and number of grid points as follows:

1. To estimate the odds function, select a probabilistic classifier and the number of simulations B based on the cross-entropy loss on held-out data;[3]
2. To compute the test statistic, choose the number of grid points for the numerical integration or optimization based on the available computational budget;
3. To ensure valid confidence sets, select the quantile regressor and the train sample size B' so that we achieve nominal coverage across the entire parameter space according to LF2I diagnostics (Section 3.4).

6. Handling Nuisance Parameters

In most applications, we only have a small number of parameters that are of primary interest. The other parameters in the model are usually referred to as nuisance parameters. In this

3. One can alternatively use the integrated odds loss (Equation 15). However, as shown in Appendix G, the odds loss is much more sensitive than the cross-entropy loss to the value of the estimated odds, which can lead to the odds loss wildly fluctuating for different values of B.

setting, we decompose the parameter space as $\Theta = \Phi \times \Psi$, where Φ contains the parameters of interest, and Ψ contains nuisance parameters. Our goal is to construct a confidence set for $\phi \in \Phi$. To guarantee frequentist coverage by Neyman's inversion technique, however, one needs to test null hypotheses of the form $H_{0,\phi_0} : \phi = \phi_0$ by comparing the test statistics to the cutoffs $\widehat{C}_{\phi_0} := \inf_{\psi \in \Psi} \widehat{C}_{(\phi_0,\psi)}$ (Section 3.3.1). That is, one needs to control the type I error at each ϕ_0 for *all* possible values of the nuisance parameters. Computing such infimum can be numerically unwieldy, especially if the number of nuisance parameters is large (van den Boom et al., 2020; Zhu et al., 2020). Below we propose approximate schemes for handling nuisance parameters:

In `ACORE`, we use a hybrid resampling or "likelihood profiling" method (Chuang and Lai, 2000; Feldman, 2000; Sen et al., 2009) to circumvent unwieldy numerical calculations as well as to reduce computational cost. For each ϕ (on a fine grid over Φ), we first compute the "profiled" value

$$\widehat{\psi}_\phi = \arg\max_{\psi \in \Psi} \prod_{i=1}^{n} \widehat{\mathbb{O}} \left(\mathbf{X}_i^{\text{obs}}; (\phi, \psi) \right),$$

which (because of the odds estimation) is an approximation of the maximum likelihood estimate of ψ at the parameter value ϕ for observed data D. By definition, the estimated `ACORE` test statistic for the hypothesis $H_{0,\phi_0} : \phi = \phi_0$ is exactly given by $\widehat{\Lambda}(\mathcal{D}; \phi_0) = \widehat{\Lambda}(\mathcal{D}; (\phi_0, \widehat{\psi}_\phi))$. However, rather than comparing this statistic to $\widehat{C}_{\phi_0}$, we use the hybrid cutoff

$$\widehat{C}'_{\phi_0} := \widehat{F}^{-1}_{\widehat{\Lambda}(\mathcal{D};\phi_0)\big|(\phi_0,\widehat{\psi}_{\phi_0})} \left(\alpha \,\middle|\, \phi_0, \widehat{\psi}_{\phi_0} \right), \tag{17}$$

where $\widehat{F}^{-1}$ is obtained via a quantile regression as in Algorithm 1, but using a training sample $\mathcal{T}'$ generated at *fixed* $\widehat{\psi}_{\phi_0}$ (that is, we run Algorithm 1 with the proposal distribution $\pi'((\phi, \psi)) \propto \pi(\phi) \times \delta_{\widehat{\psi}_\phi}(\psi)$, where $\delta_{\widehat{\psi}_\phi}(\psi)$ is a point mass distribution at $\widehat{\psi}_\phi$). Alternatively, one can compute the p-value

$$\widehat{p}(D; \phi_0) := \widehat{\mathbb{P}}_{\mathcal{D}\big|\phi_0,\widehat{\psi}_{\phi_0}} \left(\widehat{\Lambda}\left(\mathcal{D}; \phi_0\right) < \widehat{\Lambda}\left(D; \phi_0\right) \,\middle|\, \phi_0, \widehat{\psi}_{\phi_0} \right) \tag{18}$$

where $\widehat{\mathbb{P}}$ is obtained via a regression as in Algorithm 5, but with $\mathcal{T}'$ simulated at fixed $\widehat{\psi}_{\phi_0}$ (that is, we run Algorithm 5 with the proposal distribution $\pi'((\phi, \psi)) \propto \pi(\phi) \times \delta_{\widehat{\psi}_\phi}(\psi)$. Hybrid methods do not always control α, but they are often a good approximation that lead to robust results (Aad et al., 2012; Qian et al., 2016). We refer to `ACORE` approaches based on Equation 17 or Equation 18 as "`h-ACORE`" approaches.

In contrast to `ACORE`, the `BFF` test statistic averages (rather than maximizes) over nuisance parameters. Hence, instead of adopting a hybrid resampling scheme to handle nuisance parameters, we approximate p-values and critical values, in what we refer to as "`h-BFF`", by using the marginal model of the data $\mathbf{X}$ at parameter of interest ϕ:

$$\widetilde{f}(\mathbf{x}|\phi) = \int_{\nu \in \Psi} f_\theta(\mathbf{x}) \, d\pi(\nu).$$

We implement such a scheme by first drawing the train sample $\mathcal{T}'$ from the entire parameter space $\Theta = \Phi \times \Psi$, and then applying quantile regression (or probabilistic classification) using ϕ only.

Algorithm 7 details our construction of `ACORE` and `BFF` confidence sets when calibrating critical values under the presence of nuisance parameter (construction via p-value estimation is analogous). In Section 7.3, we demonstrate how our diagnostics branch can shed light on whether or not the final results have adequate frequentist coverage.

7. Examples

Next we analyze the empirical performance of our LF2I framework in different problem settings with: unknown null distribution (Section 7.1), high-dimensional feature and parameter dimension (Section 7.2), and nuisance parameters (Section 7.3).

7.1 Gaussian Mixture Model Example: Unknown Null Distribution

A common practice in LFI is to first estimate the likelihood and then assume that the LR statistic is approximately χ^2 distributed according to Wilks' theorem (Drton, 2009). However, in settings with small sample sizes or irregular statistical models, such approaches may lead to confidence sets with incorrect conditional coverage; it is often difficult to identify exactly when that happens, and then know how to recalibrate the confidence sets. See Algeri et al. (2019) for a discussion of all conditions needed for Wilks' theorem to apply, which are often not realized in practice.

The Gaussian mixture model (GMM) is a classical example where the LR statistic is known but its null distribution is unknown, even asymptotically. Indeed, the development of valid statistical methods for GMM is an active area of research (Redner, 1981; McLachlan, 1987; Dacunha-Castelle and Gassiat, 1997; Chen and Li, 2009; Wasserman et al., 2020). Here we consider a one-dimensional normal mixture with unknown mean but known unit variance:

$$X \sim 0.5N(\theta, 1) + 0.5N(-\theta, 1),$$

where the parameter of interest $\theta \in \Theta = [0, 5]$.

In this example, we consider three different approaches for estimating the critical value C_{θ_0} of a LRT at level α of $H_{0,\theta_0} : \theta = \theta_0$, for different $\theta_0 \in \Theta$:

- "LR with Monte Carlo samples", where we estimate critical values by drawing 1000 simulations at each point θ_0 on a fine grid over Θ; we take C_{θ_0} to be the $1-\alpha$ quantile of the distribution of the LR statistic, computed using the MC sample at each fixed θ_0. This approach is often just referred to as *MC hypothesis testing*.
- "Chi-square LRT", where we *assume* that $-2\text{LR}(\mathcal{D}; \theta_0) \sim \chi^2_1$, and hence take $-2C_{\theta_0}$ to be the same as the upper α quantile of a χ^2_1 distribution.
- "LR with C_{θ_0} via quantile regression", where we estimate C_{θ_0} via quantile regression (Algorithm 1) based on a total of $B' = 1000$ simulations of size n sampled uniformly on Θ.

We then construct Neyman confidence sets by inverting the hypothesis tests, and finally assess the conditional coverage of the constructed confidence sets with the diagnostic branch of the LF2I framework (Algorithm 2 with $B'' = 1000$).

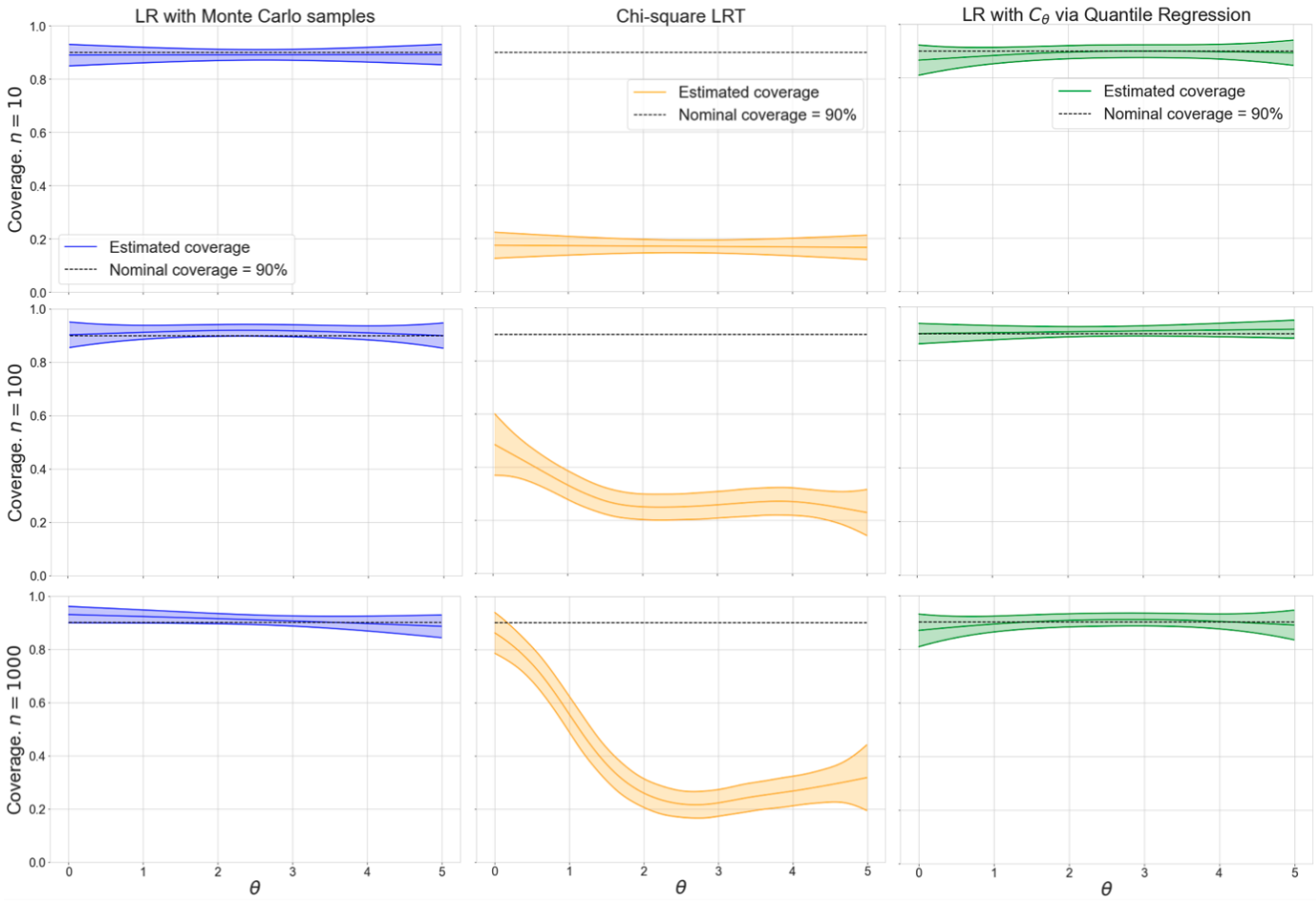


Figure 3: GMM example with sample size $n = 10$ (*top*), $n = 100$ (*center*) and $n = 1000$ (*bottom*). The curves show the estimated conditional coverage of 90% confidence sets for θ constructed with three different approaches; see text for details. The coverage of the confidence sets is estimated with the LF2I diagnostic branch (Algorithm 2), which outputs mean estimated conditional coverage with two-standard-deviation ($\pm 2\sigma$) prediction intervals. *Left:* "LR with Monte Carlo samples" achieves nominal conditional coverage but is computationally expensive, especially in higher dimensions. *Center:* "Chi-square LRT" clearly undercovers, i.e. confidence sets are not valid. *Right:* "LR with C_{θ_0} via quantile regression" returns finite-sample confidence sets with the nominal conditional coverage of 0.9, but using a number of simulations lower by a factor equal to the size of the grid over Θ (51 in this case).

Figure 3 shows LF2I diagnostics for the three different approaches when the data sample size is $n = 10, 100, 1000$. Confidence sets from "Chi-square LRT" are clearly not valid at any n, which shows that Wilks' theorem does not apply in this setting. The only exception arises when n is large enough and θ approaches 0, in which case the mixture reduces to a unimodal Gaussian (see bottom center panel of Figure 3). On the other hand, "LR with C_{θ_0} via quantile regression" returns valid finite-sample confidence sets with conditional coverage equivalent to "LR with Monte Carlo samples". A key difference between the LF2I and MC methods (which both achieve nominal coverage across the parameter space) is that the LF2I results are based on 1000 samples in total, whereas the MC results are based on 1000 MC samples at each θ_0 on a grid (in this example, we use 51 grid points; the gridding approach quickly becomes intractable in higher parameter dimensions and larger scales).

Appendix F gives details on the specific quantile regressor (for Algorithm 1) and probabilistic classifier (for Algorithm 2) used in Figure 3. The appendix also presents extensions of the above experiments to confidence sets via p-value estimation and asymmetric mixtures.

7.2 Multivariate Gaussian Example: Scaling with Dimension

In this section, we assess how our procedures scale with parameter and feature dimension for the (analytically solvable) problem of estimating the population mean of d-dimensional Gaussian data. (This is an example where we can analytically derive test statistics as well as the exact null distribution of the LR statistic.) In Section 7.2.1, we first assume that the LR statistic is known but not its null distribution, so that we can compare our calibrated confidence sets to universal inference sets and the exact (uniformly most powerful) LR confidence sets. Thereafter, in Section 7.2.2, we consider the standard LFI setting with a likelihood that is only implicitly encoded by the simulator.

For the multivariate Gaussian (MVG) example, suppose $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim N(\theta, I_d)$, where I_d is the d-dimensional identity matrix and $\theta \in \mathbb{R}^d$ is an unknown parameter. For this model, the sample mean $\overline{\mathbf{X}}_n \sim N(\theta, n^{-1} I_d)$ is a sufficient statistic, so we can express our test statistics in terms of $\overline{\mathbf{X}}_n$. The likelihood ratio statistic for testing $H_{0,\theta_0} : \theta = \theta_0$ versus $H_{1,\theta_0} : \theta \neq \theta_0$ is

$$\mathrm{LR}(\overline{\mathbf{X}}_n; \theta_0) = \log \frac{N(\overline{\mathbf{X}}_n; \theta_0, n^{-1} I_d)}{N(\overline{\mathbf{X}}_n; \overline{\mathbf{X}}_n, n^{-1} I_d)} = -\frac{n}{2} \|\overline{\mathbf{X}}_n - \theta_0\|^2. \tag{19}$$

For the MVG example, it holds exactly that $-2\mathrm{LR}(\overline{\mathbf{X}}_n; \theta_0) \sim \chi^2_d$. We refer to inference based on the above result as "exact LRT". For example, the exact LRT confidence set at level α is defined as

$$R^{\mathrm{LRT}}(\overline{\mathbf{X}}_n) = \{\theta_0 \in \Theta : n\|\bar{\mathbf{X}}_n - \theta_0\|^2 \leq c_{\alpha,d}\},$$

where $c_{\alpha,d}$ is the upper α quantile of a χ^2_d distribution.

For the Bayes factor, we assume a proposal distribution π that is uniform over an axis-aligned hyper-rectangle with corner points at $\mathbf{a} = (a, ..., a)$ and $\mathbf{b} = (b, ..., b) \in \mathbb{R}^d$ for $a < b$. The exact Bayes factor for testing $H_{0,\theta_0} : \theta = \theta_0$ versus $H_{1,\theta_0} : \theta \neq \theta_0$ is

$$\mathrm{BF}(\overline{\mathbf{X}}_n; \theta_0) = \frac{N(\overline{\mathbf{X}}_n; \theta_0, n^{-1} I_d)}{\left(\frac{1}{b-a}\right)^d \prod_{j=1}^d \left[\frac{1}{2}\mathrm{erf}\left(\frac{b-\overline{X}_{n,j}}{\sqrt{2n}}\right) - \frac{1}{2}\mathrm{erf}\left(\frac{a-\overline{X}_{n,j}}{\sqrt{2n}}\right)\right]}. \tag{20}$$

(See Appendix G for a derivation.) We refer to inference based on the above expression and high-resolution Monte Carlo sampling to compute critical values as "exact BF".

With the exact LRT and exact BF as benchmarks, we can assess the coverage and power of our LFI constructed confidence sets with increasing parameter and feature dimension d.

7.2.1 Finite-Sample Confidence Sets for Known Test Statistic

We start with an LFI setting where we assume the test statistic is known, but not its null distribution and critical values. Recently, Wasserman et al. (2020) proposed a general set of procedures for constructing confidence sets and hypothesis tests with finite-sample guarantees. One instance of universal inference uses the crossfit likelihood-ratio test (crossfit LRT), which averages the likelihood ratio statistic over two data splits; see also recent work by Dunn et al. (2021), which compares different universal inference schemes on MVG data. Our LFI approach can also produce valid finite-sample confidence sets for known test statistic by calibrating the critical value as in Algorithm 1.

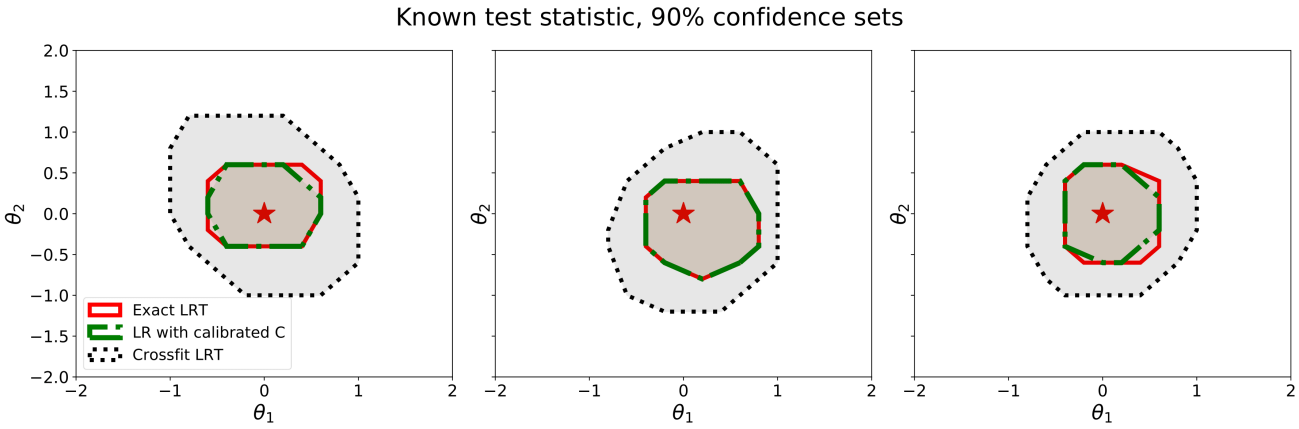


Figure 4: Confidence sets for known test statistics and bivariate Gaussian data. When $d = 2$, our method for estimating critical values with $B' = 500$ simulations ("LR with calibrated C"; green contour) returns 90% confidence sets that are close to the exact LRT confidence sets (red contour) and smaller than the more conservative universal via crossfit LRT sets (gray shading). The figures correspond to three random realizations of observed data with $n = 10$ drawn from the Gaussian model with true parameter $\theta = (0, 0)$ (indicated with a red star).

Figure 4 compares three "Exact LRT" sets with confidence sets constructed with our method for estimating the critical value ("LR with calibrated C"), and confidence sets via universal inference ("Crossfit LRT"). The dimension here is $d = 2$, the true (unknown) parameter is $\theta^* = (0, 0)$, and the sample size is $n = 10$. By calibrating the critical value, we can achieve valid confidence sets similar to exact LRT for a modest number of $B' = 500$ simulations. Universal inference does not adjust the critical values according to the value of θ, and pays a price for its generality in terms of larger confidence sets and lower power.

Finite-sample confidence sets for known test statistic

	d=10	d=20	d=50	d=100
Coverage of LR with calibrated C	0.91 ± 0.03	0.91 ± 0.03	0.88 ± 0.03	0.88 ± 0.03
Coverage of crossfit LRT	0.993 ± 0.008	0.997 ± 0.005	1.000 ± 0.000	1.000 ± 0.000

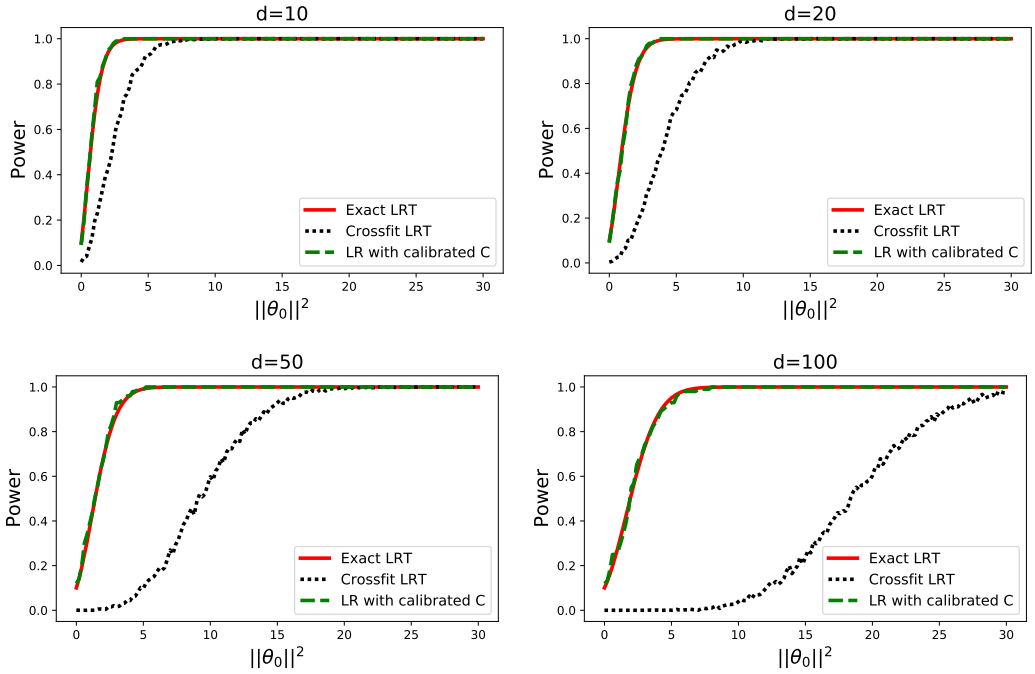


Figure 5: Confidence sets for known test statistic and d-dimensional Gaussian data. Coverage and power of finite-sample confidence sets constructed via exact LRT, LR with calibrated C, and universal inference via crossfit LRT (see text for details). All methods achieve the nominal coverage of 0.9. When the likelihood ratio statistic is known, our construction with $B' = 5000$ simulations yields the same power as the exact LRT, even in high dimensions. By calibrating the critical values, one can achieve more precise confidence sets and higher power than universal inference. See Figure 4 for example confidence sets in dimension $d = 2$. The difference in precision and power between the two methods increase with dimension d.

Figure 5 extends the comparison to coverage and power in higher dimensions d. As before, we observe a sample of size $n = 10$ from a MVG centered at $\theta^* = \mathbf{0}$. We construct confidence sets using exact LRT, LR with calibrated C, and crossfit LRT for 100 draws from the MVG. We then test $H_{0,\theta_0} : \theta = \theta_0$ versus $H_{1,\theta_0} : \theta \neq \theta_0$ for different values of θ_0 at increasing distance $\|\theta_0\|$ from the origin. We reject H_{0,θ_0} if θ_0 is outside the constructed confidence set. In this example, coverage is measured by the proportion of times the parameter value $\theta_0 = \mathbf{0}$ is (correctly) included in the confidence set over 100 such repetitions. Similarly, power is measured by the proportion of times a parameter value $\theta_0 \neq \mathbf{0}$ is (correctly) outside the constructed confidence set. For better visualization, we have chosen the test points θ_0 so that we have roughly an equal number of test points at each squared distance $\|\theta_0\|^2$.

The table at the top of the figure shows that both "LR with calibrated C" and "Crossfit LRT" control the type I error at level $\alpha = 0.1$ for dimensions d between 10 to 100. Crossfit LRT, however, tends to be overly conservative. As for the two-dimensional example, our method achieves almost the same power as the exact LR test, even for $d = 100$ and a modest budget of $B' = 5000$ simulations. Crossfit LRT has much lower power, as expected. The differences in power between the two methods grows with increasing dimension d.

7.2.2 Finite-Sample Confidence Sets in an LFI Setting

Next, we consider the more challenging LFI scenario where one is only able to sample data from a forward simulator F_θ, and hence needs to estimate *both* the test statistic and critical values. As before, we simulate observed data of sample size $n = 10$ from a d-dimensional Gaussian distribution with true mean $\theta^* = \mathbf{0}$, but now we estimate both the test statistics and the critical values for controlling the type I error. We use `ACORE` to approximate the LRT, and `BFF` to approximate tests based on the Bayes factor with a uniform prior over the hyper-rectangle $[-5, 5]^d$.

Following the strategy outlined in Section 5, we select a quadratic discriminant analysis (QDA) classifier to estimate the odds, and quantile regression with gradient boosted trees to estimate cutoffs at level $\alpha = 0.1$. Figure 6 compares `ACORE` and `BFF` confidence sets when $d = 2$ to the exact LRT and exact BF counterparts (achieved with computationally expensive MC sampling to estimate critical values). Both `ACORE` and `BFF` achieve similarly sized confidence sets as their exact counterparts, with modest budgets of $B = B' = 5000$ simulations and $M = 2500$ evaluation points for maximization or integration.

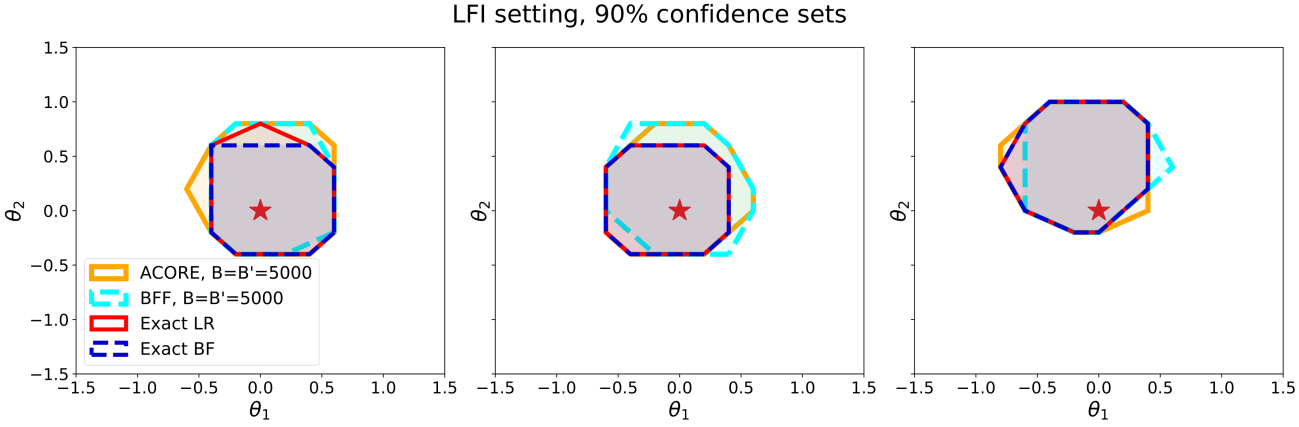


Figure 6: LFI setting: When $d = 2$, `BFF` and `ACORE` 90% confidence sets are of similar size to those constructed using the exact LR and BF. The true parameter $\theta = (0, 0)$ (indicated with a star), $n = 10$ observations, $B = B' = 5000$ and $M = 2500$ samples for `BFF` and `ACORE`. The figures show three random realizations of the observed data.

Figure 7 shows the coverage and power of these methods as the dimension d increases. We use the same approach as in Section 7.2.1 to compute the power over 100 repetitions. First, we observe that both `ACORE` and `BFF` confidence sets consistently achieve the nom-

Finite-sample confidence sets in a likelihood-free inference setting

	d=1	d=2	d=5	d=10
Coverage of ACORE	0.92 ± 0.03	0.92 ± 0.03	0.90 ± 0.03	0.90 ± 0.03
Coverage of BFF	0.94 ± 0.02	0.89 ± 0.03	0.96 ± 0.02	0.87 ± 0.03

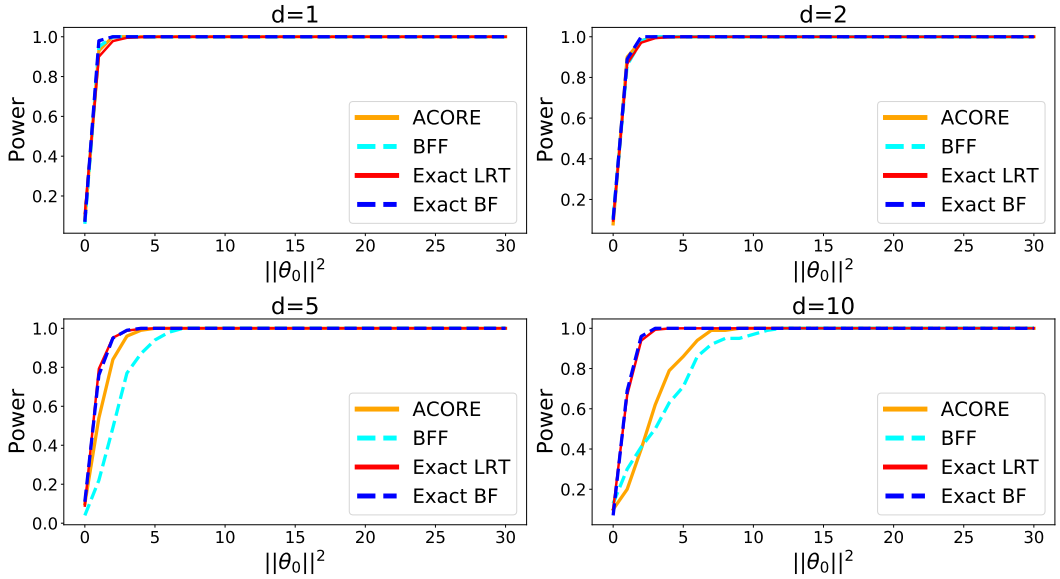


Figure 7: LFI setting: Coverage and power for ACORE and BFF confidence sets and their exact likelihood ratio test (LRT) and Bayes factor (BF) counterparts at dimension $d = 1, 2, 5$ and 10 across 100 repetitions. Both ACORE and BFF return valid confidence sets with coverage at or above the nominal confidence level $1 - \alpha = 0.9$. The loss in power relative the exact methods increases as d increases. (We use QDA to learn the odds, with sample size B guided by Figure 12, a computational budget for maximization and integration of $M = 10000$, and quantile regression gradient boosting trees with $B^{'} = 10000$.)

inal 0.90 confidence level,[4] even in higher dimensions. Next, we consider power. Loosely speaking, the exact LRT and BF power curves can be seen as upper bounds on the power of ACORE and BFF, respectively. The results indicate that ACORE and BFF confidence sets are precise in low dimensions, but their power drops as d increases.

A closer look (see Appendix G) indicates that the loss in power for $d \geq 5$ is primarily due to numerical error in the maximization or integration step (referred to as error e_2 in Section 5) of ACORE and BFF, respectively. Hence, we foresee that the current implementations of ACORE and BFF with uniformly spaced evaluation points would significantly benefit from more efficient numerical computation. For maximization, higher efficiency approaches have been suggested in the hyper-parameter search literature for machine learning algorithms, such as kernel-based Bayesian optimization (Kandasamy et al., 2015) and bandit-based approaches (Li et al., 2018) (see Feurer and Hutter (2019) for an overview). For integration, one could employ more efficient approaches that rely on, e.g., adaptive sampling (Peter

4. The coverage falls within or above expected variation for 100 repetitions, which is in the range [84, 95].

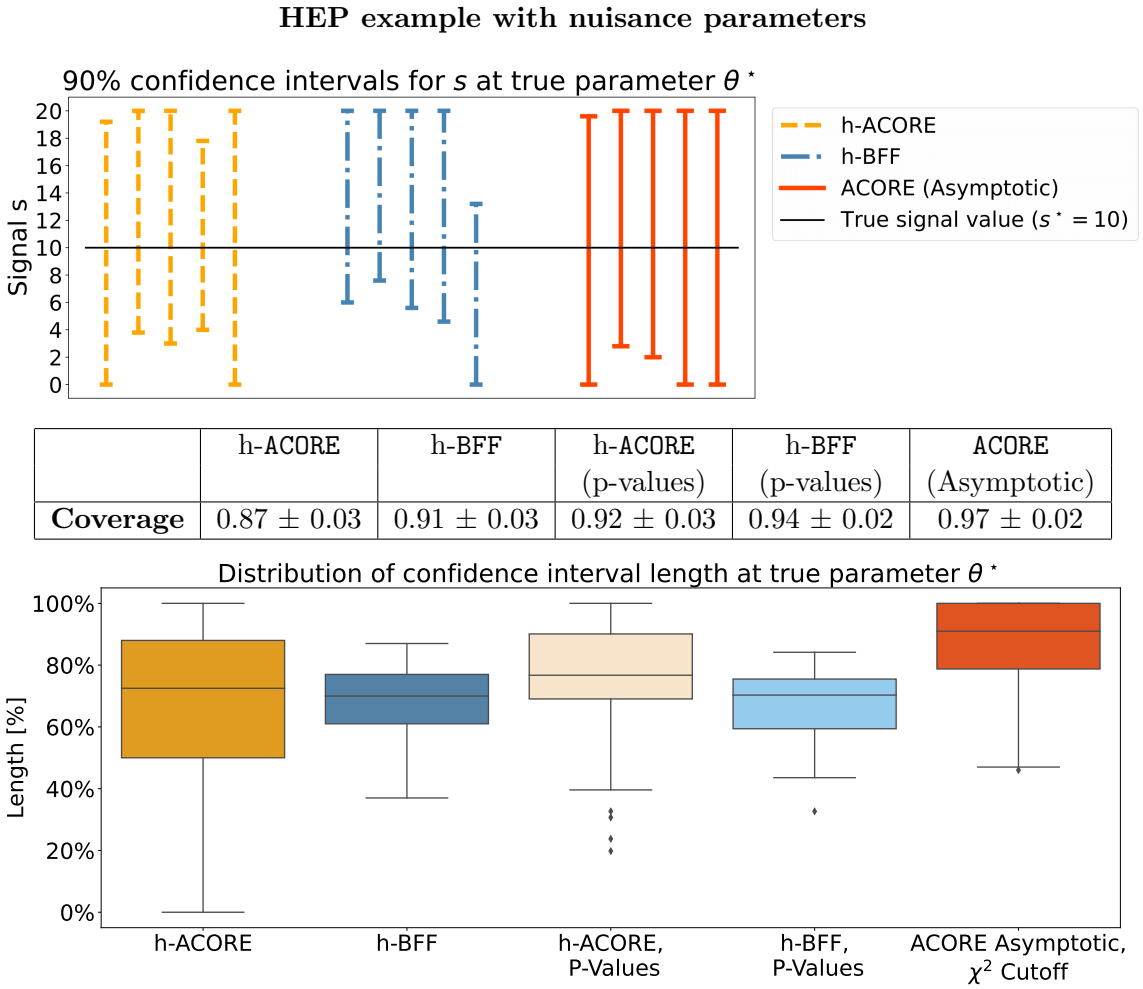
Lepage, 1978; Jadach, 2003), nested sampling (Feroz et al., 2009; Handley et al., 2015) or machine learning algorithms (Bendavid, 2017; Gao et al., 2020).

7.3 High-Energy Physics Example: Nuisance Parameters

Hybrid methods, which maximize or average over nuisance parameters, do not always control the type I error of statistical tests. For small sample sizes, there is no theorem as to whether profiling or marginalization of nuisance parameters will give better frequentist coverage for the parameter of interest (Cousins, 2018, Section 12.5.1). In addition, most practitioners consider a thorough check of frequentist coverage to be impractical (Cousins, 2018, Section 13). In this example, we apply the LFI hybrid schemes from Section 6 to a high-energy physics (HEP) counting experiment (Lyons, 2008; Cowan et al., 2011; Cowan, 2012) with nuisance parameters, which is a simplified version of a real particle physics experiment, where the true likelihood function is not known. We illustrate how our diagnostics can guide the analyst and provide insight into which method to choose for the problem at hand.

Consider a "Poisson counting experiment" where particle collision events are counted under the presence of both an uncertain background process and a (new) signal process. The goal is to estimate the signal strength. To avoid identifiablility issues, the background rate is estimated separately by counting the number of events in a control region where the signal is believed to be absent. Hence, the observable data $\mathbf{X} = (M, N)$ contain two measurements, where $M \sim \text{Pois}(\gamma b)$ is the number of events in the control region, and $N \sim \text{Pois}(b + \epsilon \cdot s)$ is the number of events in the signal region. Our parameter of interest is the signal strength s, and we will treat b and ϵ as nuisance parameters. (The parameter ϵ is a factor for the number of signal events expected in a particular experiment; it can, for instance, represent experimental inefficiencies or the experiment's running time (Lyons, 2008). The parameter γ is a scaling parameter that we take here as known.) In our example, we set $\gamma = 1$ and assume $n = 10$ observations from the Poisson model above, with parameters in the ranges $s \in [0, 20]$, $b \in [90, 110]$ and $\epsilon \in [0.5, 1.0]$.

Figure 8, top, shows examples of 90% confidence sets for the signal strength s for observed data with true parameter $\theta^* = (s^*, b^*, \epsilon^*) = (10, 100, 0.75)$. (We learn the odds using a QDA classifier with $B = 100000$; $n = 10$, $B^{'} = 10000$ and $M = 10000$.) The table lists the estimated coverage $\widehat{\mathbb{P}}_{\mathcal{D}|\theta^*}\left(\theta^* \in \widehat{R}(\mathcal{D})|\theta^*\right)$ for 100 repetitions at θ^*. The box plots show the distribution of confidence set lengths as a percentage of the signal parameter range $[0, 20]$ for the same 100 repetitions. In our comparison, we include h-`ACORE` and h-`BFF` confidence intervals, which we construct with the critical value and p-value hybrid schemes in Algorithm 7. We also include `ACORE` confidence intervals with cutoffs derived from the χ^2 distribution (which is the asymptotic distribution of the profiled likelihood ratio; Murphy and Van Der Vaart 2000). These results seem to imply that all five methods return valid confidence intervals, with the intervals based on calibrated critical values or p-values being shorter (more powerful) than those for `ACORE` based on asymptotic cutoff. Furthermore, h-`BFF` confidence sets appear to be shorter and less variable than h-`ACORE` confidence sets. Figure 8, however, only considers confidence sets for data at a particular θ-value. Next, we showcase how our diagnostic branch provides a thorough analysis of conditional coverage over the entire parameter space; this evaluation is essential in practice as the true θ^*-value is our quantity of interest, and not known in advance.



	h-ACORE	h-BFF	h-ACORE (p-values)	h-BFF (p-values)	ACORE (Asymptotic)
Coverage	0.87 ± 0.03	0.91 ± 0.03	0.92 ± 0.03	0.94 ± 0.02	0.97 ± 0.02

Figure 8: HEP example with nuisance parameters. *Top*: Examples of 90% confidence intervals of the signal strength s for observed data $X^{obs} = (X_1, ..., X_{10})$ from a model with true parameter $\theta^* = (s^*, b^*, \epsilon^*)$. *Center/bottom*: Repeat the construction of confidence intervals 100 times for data simulated at θ^*. The table lists the estimated coverage, and the boxplots graph the distribution of the length of constructed confidence intervals (as a percentage of the total parameter range for s). While all LFI methods appear to construct valid confidence intervals, calibration of critical values and p-values (with BFF, in particular) lead to smaller and more powerful confidence intervals.

With logistic regression and a total of $B'' = 500$ simulations, we compute the estimated coverage with a two-standard-deviation prediction band for all $\theta = (s, b, \epsilon)$ over a regular grid across the parameter space Θ. Parameter regions where the nominal coverage of $1-\alpha = 0.9$ falls within the prediction band are considered to have "correct coverage" (CC). Regions where the upper versus lower limit of the prediction band falls below versus above $1-\alpha$ are labeled as having "undercoverage" (UC) and "overcoverage" (OC), respectively. Figure 9, top, highlights that a very small portion (4.0%) of the parameter space has undercoverage for the hybrid version of BFF, whereas the ACORE approach based on asymptotic cutoff

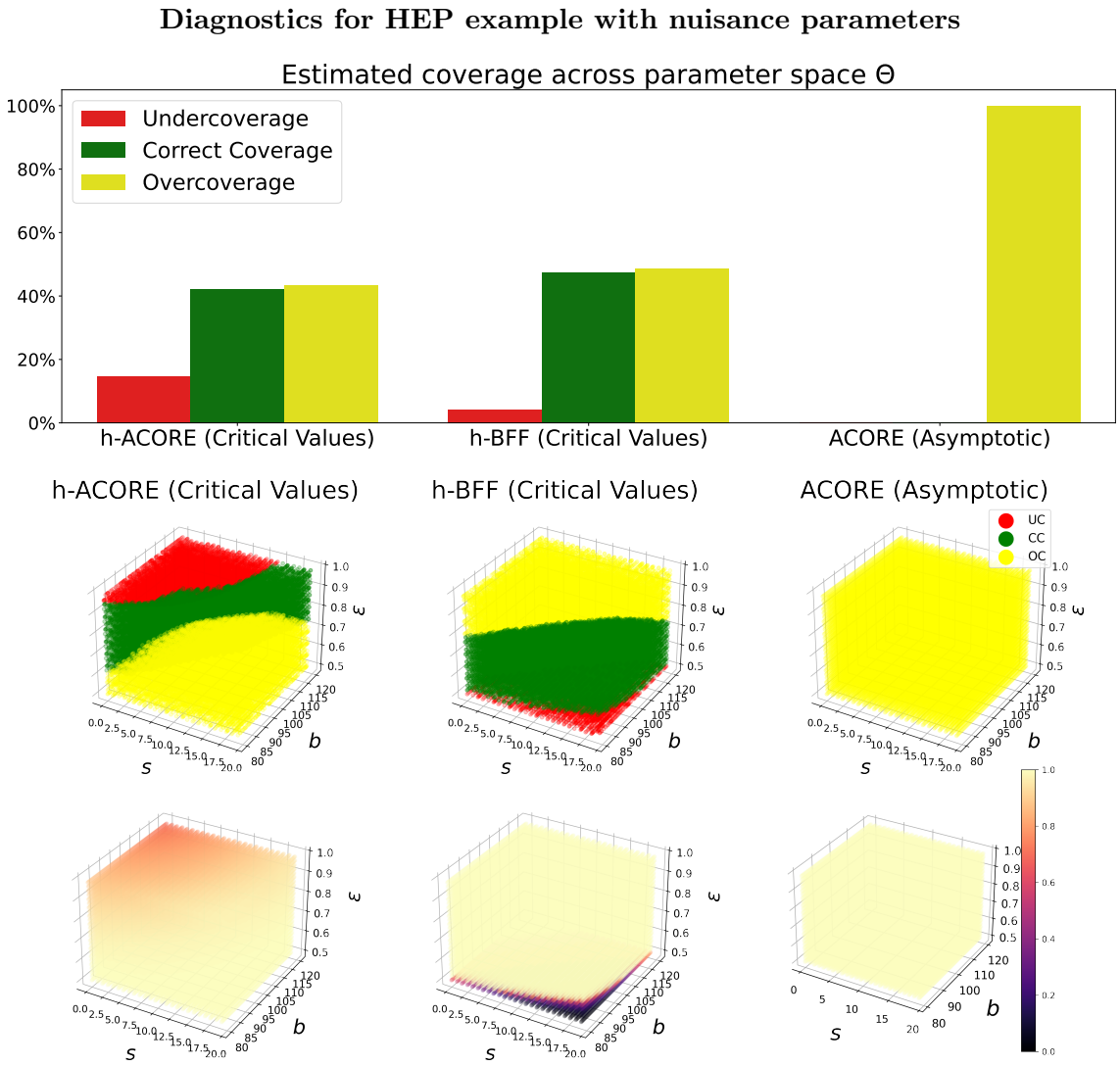


Figure 9: Hybrid approaches which maximize or average over nuisance parameters do not guarantee frequentist coverage. Our diagnostic tool (Section 3.4) can identify regions in parameter space with undercoverage (UC), correct coverage (CC), and overcoverage (OC), respectively. See text for details. *Top*: h-`BFF` performs the best in terms of having the largest proportion of the parameter space with correct coverage and only a small fraction of the parameter space with undercoverage. *Center*: Points in parameter space colored by label (UC, CC, OC). *Bottom*: Points in parameter space colored by the upper limit of the two-standard-deviation prediction interval for coverage. The latter results reveal that h-`BFF` with averaging over nuisance parameters tends to undercover more severely (though in a very small region of the parameter space) than h-`ACORE` with critical values via hybrid resampling/profiling; the UC regions of h-`BFF` here correspond to settings with low ϵ, high s and low b.

overcovers for all parameter values. The middle panel shows exactly where the CC, UC and OC regions fall in parameter space for the three different approaches; the bottom panel displays the values of the upper limit of the prediction band for coverage. Interestingly, these results illustrate that h-`BFF`, which averages over nuisance parameters, has the best overall performance, but it can severely undercover for a small region with low ϵ, low background b and high signal s. On the other hand, h-`ACORE` based on hybrid resampling undercovers (but only slightly) in a larger region of the parameter space with high ϵ, small signal s and high background b.

Often, coverage is evaluated using analyses similar to Figure 8 for best-fit or fiducial parameter values. Figure 9 is an example of how our LF2I framework can provide the analyst with a more complete and interpretable check of conditional validity.

8. Conclusions and Discussion

Validity. Our proposed LF2I methodology leads to frequentist confidence sets and hypothesis tests with finite-sample guarantees (when there are no nuisance parameters). *Any* existing or new test statistic — that is, not only estimates of the LR statistic — can be plugged into our framework to create tests that indeed control type I error. The implicit assumption is that the null distribution of the test statistic varies smoothly in parameter space. If that condition holds, then we can efficiently leverage quantile regression methods to construct valid confidence sets by a Neyman inversion of simple hypothesis tests, without having to rely on asymptotic results. In settings where the likelihood can be evaluated, our framework leads to more powerful tests and smaller confidence sets than universal inference, but at the cost of having to simulate data from the likelihood.

Nuisance parameters and diagnostics. For small sample sizes, there is no theorem to tell us whether profiling or marginalization of nuisance parameters will give better frequentist coverage for the parameter of interest (Cousins, 2018, Section 12.5.1). It is generally believed that hybrid resampling methods return approximately valid confidence sets, but that a rigorous check of validity is infeasible when the true solution is not known. Our diagnostic branch presents practical tools for assessing empirical coverage across the entire parameter space (including nuisance parameters). After the data scientist sees the results, he/she can decide which method is most appropriate for the application at hand. For example, in our HEP counting experiment (Section 7.3), our diagnostics revealed that h-`BFF` (which averages the estimated odds over nuisance parameters) returned less variable results and higher power in an LFI setting, but with small regions of the parameter space undercovering more severely than h-`ACORE` (which maximizes the estimated odds over nuisance parameters).

Power. Statistical power is the hardest property to achieve in practice in LFI. This is the area where we foresee that most statistical and computational advances will take place. As mentioned in Section 5 and then illustrated in Section 7.2.2, the power or size of the LFI confidence sets for θ depends not only on the theoretical properties of the (exact) test statistics, such as the LR statistic or the Bayes factor. Ultimately, the power may be decided by: (1) how well we are able to estimate the likelihood or odds (this is the e_1 statistical

estimation error), and (2) how accurately we, in practice, compute the target test statistic by integration or maximization (this is the e_2 numerical error). Machine learning offers exciting possibilities on both fronts. For example, with regards to (1), Brehmer et al. (2020) offer compelling evidence that one can can dramatically improve estimates of the likelihood $p(\mathbf{x}|\theta)$ for $\theta \in \Theta$, or the likelihood ratio $p(\mathbf{x}|\theta_1, \theta_2)$ for $\theta_1, \theta_2 \in \Theta$, by a "mining gold" approach that extracts additional information from the simulator about the latent process. Future work could incorporate a mining gold approach into our general LF2I framework, with the calibration and diagnostic branches as separate modules. This might improve estimates of the odds $\mathbb{O}(\mathbf{X}; \theta)$ (Equation 7), and thereby likelihood-based test statistics such as `ACORE` and `BFF`.

Other test statistics. Our work also presents another new direction for likelihood-free frequentist inference: So far frequentist LFI methods are estimating either likelihoods or likelihood ratios, and then often relying on asymptotic properties of the LR statistic. We note that there are settings where it may be easier to estimate the posterior $p(\theta|\mathbf{x})$ than the likelihood $p(\mathbf{x}|\theta)$; the same way there may be settings where the Bayes factor in practice leads to higher power than the LRT. Because our general LF2I framework (for creating valid finite-sample confidence sets and tests) is agnostic to whether we use Bayesian or classical algorithms to create the test statistic itself, we can potentially leverage prediction methods that estimate the conditional mean $\mathbb{E}[\theta|\mathbf{x}]$ and variance $\mathbb{V}[\theta|\mathbf{x}]$ to construct frequentist confidence sets and hypothesis tests for θ with finite-sample guarantees. We are currently exploring the test statistic $T = \frac{(\mathbb{E}[\theta|\mathbf{x}]-\theta_0)^2}{\mathbb{V}[\theta|\mathbf{x}]}$ (Anonymized (2021)), which in some scenarios corresponds to the Wald statistic for testing $H_{0,\theta_0} : \theta = \theta_0$ against $H_{1,\theta_0} : \theta \neq \theta_0$ Wald (1943), as a potentially more attractive alternative to the LR statistic.

References

G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A. Abdelalim, O. Abdinov, R. Aben, B. Abi, M. Abolins, et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1):1–29, Sep 2012. ISSN 0370-2693. doi: 10.1016/j.physletb.2012.08.020.

S. Algeri, J. Aalbers, K. D. Morå, and J. Conrad. Searching for new physics with profile likelihoods: Wilks and beyond. *arXiv preprint arXiv:1911.10237*, 2019.

Anonymized. Confidence sets and hypothesis testing in a likelihood-free inference setting. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2323–2334, Virtual, 13–18 Jul 2020. PMLR.

Anonymized. Advanced data analysis (ADA) report: Likelihood-free frequentist inference for calorimetric muon energy measurement in high-energy physics. Technical report, Anonymized, December 2021.

T. Ayano. Rates of convergence for the k-nearest neighbor estimators with smoother regression functions. *Journal of Statistical Planning and Inference*, 142(9):2530–2536, 2012.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

J. Bendavid. Efficient Monte Carlo integration using boosted decision trees and generative deep neural networks. *arXiv preprint arXiv:1707.00028*, 6 2017.

G. Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.

P. J. Bickel and B. Li. Local polynomial regression on unknown manifolds. *Complex datasets and inverse problems*, pages 177–186, 2007.

H. J. Bierens. Uniform consistency of kernel estimators of a regression function under generalized conditions. *Journal of the American Statistical Association*, 78(383):699–707, 1983.

R. Bordoloi, S. J. Lilly, and A. Amara. Photo-z performance for precision cosmology. *Monthly Notices of the Royal Astronomical Society*, 406(2):881–895, 2010.

J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117 (10):5242–5249, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1915980117.

S. Chatrchyan, V. Khachatryan, A. M. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, et al. Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc. *Physics Letters B*, 716(1):30–61, 2012.

J. Chen and P. Li. Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, 37(5A):2523–2542, 2009.

Y. Chen and M. U. Gutmann. Adaptive Gaussian copula ABC. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1584–1592. PMLR, 16–18 Apr 2019.

C.-S. Chuang and T. L. Lai. Hybrid resampling methods for confidence intervals. *Statistica Sinica*, 10(1):1–33, 2000. ISSN 10170405, 19968507.

S. R. Cook, A. Gelman, and D. B. Rubin. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006. doi: 10.1198/106186006X136976.

R. D. Cousins. Treatment of nuisance parameters in high energy physics, and possible justifications and improvements in the statistics literature. In *Statistical Problems In Particle Physics, Astrophysics And Cosmology*, pages 75–85. World Scientific, 2006.

R. D. Cousins. Lectures on statistics in theory: Prelude to statistics in practice, 2018.

G. Cowan. Discovery sensitivity for a counting experiment with back- ground uncertainty. *Technical Report*, 2012.

G. Cowan, K. Cranmer, E. Gross, and O. Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71(2), Feb 2011. ISSN 1434-6052. doi: 10.1140/epjc/s10052-011-1554-0.

K. Cranmer. Practical Statistics for the LHC. *arXiv e-prints*, art. arXiv:1503.07622, Mar 2015.

K. Cranmer, J. Pavez, and G. Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1912789117.

D. Dacunha-Castelle and E. Gassiat. Testing in locally conic models, and application to mixture models. *ESAIM: Probability and Statistics*, 1:285–317, 1997.

L. Devroye, L. Győrfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 2013.

D. L. Donoho. Asymptotic minimax risk for sup-norm loss: solution via optimal recovery. *Probability Theory and Related Fields*, 99(2):145–170, 1994.

M. Drton. Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2):979–1012, Apr 2009. ISSN 0090-5364. doi: 10.1214/07-aos571.

R. Dunn, A. Ramdas, S. Balakrishnan, and L. Wasserman. Gaussian universal likelihood ratio testing. *arXiv preprint arXiv:2104.14676*, 2021.

C. Durkan, I. Murray, and G. Papamakarios. On contrastive learning for likelihood-free inference. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2771–2781. PMLR, 13–18 Jul 2020.

M. Fasiolo, S. N. Wood, F. Hartig, and M. V. Bravington. An extended empirical saddle-point approximation for intractable likelihoods. *Electron. J. Statist.*, 12(1):1544–1578, 2018. doi: 10.1214/18-EJS1433.

G. Feldman. Multiple measurements and parameters in the unified approach. Technical report, Technical Report, Talk at the FermiLab Workshop on Confidence Limits, 2000.

G. J. Feldman and R. D. Cousins. Unified approach to the classical statistical analysis of small signals. *Physical Review D*, 57(7):3873–3889, Apr 1998. ISSN 1089-4918. doi: 10.1103/physrevd.57.3873.

F. Feroz, M. P. Hobson, and M. Bridges. Multinest: an efficient and robust Bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, 398(4):1601–1614, Oct 2009. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2009.14548.x.

M. Feurer and F. Hutter. *Hyperparameter Optimization*, pages 3–33. Springer International Publishing, Cham, 2019. ISBN 978-3-030-05318-5. doi: 10.1007/978-3-030-05318-5_1.

R. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd: Edinburgh, 11th ed. rev. edition, 1925.

C. Gao, J. Isaacson, and C. Krause. i- flow: High-dimensional integration and sampling with normalizing flows. *Machine Learning: Science and Technology*, 1(4):045023, Nov 2020. ISSN 2632-2153. doi: 10.1088/2632-2153/abab62.

S. Girard, A. Guillou, and G. Stupfler. Uniform strong consistency of a frontier estimator using kernel regression on high order moments. *ESAIM: Probability and Statistics*, 18: 642–666, 2014.

D. Greenberg, M. Nonnenmacher, and J. Macke. Automatic posterior transformation for likelihood-free inference. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2404–2414, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

M. U. Gutmann and J. Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125):1–47, 2016.

L. Győrfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media, 2006.

W. J. Handley, M. P. Hobson, and A. N. Lasenby. PolyChord: next-generation nested sampling. *Monthly Notices of the Royal Astronomical Society*, 453(4):4385–4399, Sep 2015. ISSN 1365-2966. doi: 10.1093/mnras/stv1911.

W. Hardle, S. Luckhaus, et al. Uniform consistency of a class of regression function estimators. *The Annals of Statistics*, 12(2):612–623, 1984.

L. Heinrich. Learning optimal test statistics in the presence of nuisance parameters. *arXiv preprint arXiv:2203.13079*, 2022.

J. Hermans, V. Begy, and G. Louppe. Likelihood-free MCMC with amortized approximate ratio estimators. *arXiv preprint arXiv:1903.04057*, 2020.

J. Hermans, A. Delaunoy, F. Rozet, A. Wehenkel, and G. Louppe. Averting a crisis in simulation-based inference. *arXiv preprint arXiv:2110.06581*, 2021.

M. Ho, A. Farahi, M. M. Rau, and H. Trac. Approximate bayesian uncertainties on deep learning dynamical mass estimates of galaxy clusters. *The Astrophysical Journal*, 908(2): 204, 2021.

R. Izbicki, A. Lee, and C. Schafer. High-Dimensional Density Ratio Estimation with Extensions to Approximate Likelihood Computation. In S. Kaski and J. Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 420–429, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

R. Izbicki, A. B. Lee, and T. Pospisil. ABC–CDE: Toward Approximate Bayesian Computation with complex high-dimensional data and limited simulations. *Journal of Computational and Graphical Statistics*, pages 1–20, 2019. doi: 10.1080/10618600.2018.1546594.

S. Jadach. Foam: A general-purpose cellular Monte Carlo event generator. *Computer Physics Communications*, 152(1):55 – 100, 2003. ISSN 0010-4655. doi: https://doi.org/10.1016/S0010-4655(02)00755-5.

H. Jeffreys. Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(2):203–222, 1935. doi: 10.1017/S030500410001330X.

H. Jeffreys. *Theory of probability*. Clarendon Press Oxford, 3rd ed. edition, 1961.

M. Järvenpää, M. U. Gutmann, A. Vehtari, and P. Marttinen. Parallel Gaussian process surrogate Bayesian inference with noisy likelihood evaluations. *Bayesian Anal.*, 16(1): 147–178, 2021. doi: 10.1214/20-BA1200.

K. Kandasamy, J. Schneider, and B. Poczos. High dimensional Bayesian optimisation and bandits via additive models. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 295–304, Lille, France, 07–09 Jul 2015. PMLR.

J. Kieseler, G. C. Strong, F. Chiandotto, T. Dorigo, and L. Layer. Calorimetric measurement of multi-tev muons via deep regression. *The European Physical Journal C*, 82(1):1–26, 2022.

R. Koenker, V. Chernozhukov, X. He, and L. Peng. *Handbook of quantile regression.* CRC press, 2017.

S. Kpotufe. k-NN Regression Adapts to Local Intrinsic Dimension. *Advances in Neural Information Processing Systems*, pages 729–737, 2011.

S. Kpotufe and V. Garg. Adaptivity to local smoothness and dimension in kernel regression. *Advances in Neural Information Processing Systems*, 26:3075–3083, 2013.

L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18-185:1–52, 2018.

H. Liero. Strong uniform consistency of nonparametric regression function estimates. *Probability theory and related fields*, 82(4):587–614, 1989.

J.-M. Lueckmann, P. J. Goncalves, G. Bassetto, K. Öcal, M. Nonnenmacher, and J. H. Macke. Flexible statistical inference for mechanistic models of neural dynamics. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1289–1299. Curran Associates, Inc., 2017.

J.-M. Lueckmann, G. Bassetto, T. Karaletsos, and J. H. Macke. Likelihood-free inference with emulator networks. In *Symposium on Advances in Approximate Bayesian Inference*, pages 32–53, 2019.

L. Lyons. Open statistical issues in Particle Physics. *The Annals of Applied Statistics*, 2 (3):887 – 915, 2008. doi: 10.1214/08-AOAS163.

J. G. MacKinnon. Bootstrap hypothesis testing. *Handbook of computational econometrics*, 183:213, 2009.

J.-M. Marin, P. Pudlo, C. Robert, and R. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22:1167–1180, 11 2012. doi: 10.1007/s11222-011-9288-2.

J.-M. Marin, L. Raynal, P. Pudlo, M. Ribatet, and C. Robert. ABC random forests for Bayesian parameter inference. *Bioinformatics (Oxford, England)*, 35, 05 2016. doi: 10.1093/bioinformatics/bty867.

G. J. McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):318–324, 1987.

E. Meeds and M. Welling. GPS-ABC: Gaussian process surrogate approximate Bayesian computation. *arXiv preprint arXiv:1401.2838*, 2014.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(35): 983–999, 2006.

S. A. Murphy and A. Van Der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000. doi: 10.1080/01621459.2000.10474219.

J. Neyman. On the problem of confidence intervals. *Ann. Math. Statist.*, 6(3):111–116, 09 1935. doi: 10.1214/aoms/1177732585.

J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380, 1937. ISSN 00804614.

J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 20A(1/2):175–240, 1928. ISSN 00063444.

G. Papamakarios and I. Murray. Fast ϵ-free inference of simulation models with Bayesian conditional density estimation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 1028–1036. Curran Associates, Inc., 2016.

G. Papamakarios, D. Sterratt, and I. Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848, 2019.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

G. Peter Lepage. A new algorithm for adaptive multidimensional integration. *Journal of Computational Physics*, 27(2):192 – 203, 1978. ISSN 0021-9991. doi: https://doi.org/10.1016/0021-9991(78)90004-9.

U. Picchini, U. Simola, and J. Corander. Adaptive MCMC for synthetic likelihoods and correlated synthetic likelihoods. *arXiv preprint arXiv:2004.04558*, 2020.

X. Qian, A. Tan, J. Ling, Y. Nakajima, and C. Zhang. The Gaussian CL_s method for searches of new physics. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 827(35):63–78, 2016.

S. T. Radev, U. K. Mertens, A. Voss, L. Ardizzone, and U. Köthe. Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2020. doi: 10.1109/TNNLS.2020.3042395.

R. Redner. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics*, 9(1):225–228, 1981.

S. J. Schmidt, A. I. Malz, J. Y. H. Soo, I. A. Almosallam, M. Brescia, S. Cavuoti, J. Cohen-Tanugi, A. J. Connolly, J. DeRose, P. E. Freeman, M. L. Graham, K. G. Iyer, M. J. Jarvis, J. B. Kalmbach, E. Kovacs, A. B. Lee, G. Longo, C. B. Morrison, J. A. Newman, E. Nourbakhsh, E. Nuss, T. Pospisil, H. Tranin, R. H. Wechsler, R. Zhou, R. Izbicki, and T. L. D. E. S. Collaboration). Evaluation of probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST). *Monthly Notices of the Royal Astronomical Society*, 499(2):1587–1606, 09 2020. ISSN 0035-8711. doi: 10.1093/mnras/staa2799.

B. Sen, M. Walker, and M. Woodroofe. On the unified method with nuisance parameters. *Statistica Sinica*, 19(1):301–314, 2009. ISSN 10170405, 19968507.

S. A. Sisson, Y. Fan, and M. Beaumont. *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC, 2018.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

O. Thomas, R. Dutta, J. Corander, S. Kaski, and M. U. Gutmann. Likelihood-free inference by ratio estimation. *Bayesian Anal.*, 2021. doi: 10.1214/20-BA1238. Advance publication.

A. B. Tsybakov. *Introduction to Nonparametric Estimation. Revised and Extended from the 2004 French Original. Translated by Vladimir Zaiats.* Springer Series in Statistics. New York: Springer, 2009.

W. van den Boom, G. Reeves, and D. B. Dunson. Approximating posteriors with high-dimensional nuisance parameters via integrated rotated Gaussian approximation. *Biometrika*, Aug 2020. ISSN 1464-3510. doi: 10.1093/biomet/asaa068.

V. Ventura. Bootstrap tests of hypotheses. In *Analysis of parallel spike trains*, pages 383–398. Springer, 2010.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482, 1943.

L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

R. Wilkinson. Accelerating ABC methods using Gaussian processes. In *Artificial Intelligence and Statistics*, pages 1015–1023, 2014.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 03 1938. doi: 10.1214/aoms/1177732360.

S. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466: 1102–4, 08 2010. doi: 10.1038/nature09319.

Y. Yang, A. Bhattacharya, and D. Pati. Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv preprint arXiv:1708.04753*, 2017.

Y. Zhu, X. Shen, and W. Pan. On high-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association*, 115(529):217–230, 2020.

A. Estimating Odds

Algorithm 3 provides details on how to create the training sample $\mathcal{T}$ for estimating odds. Out of the total number of simulations B, a proportion p is generated by the stochastic forward simulator F_θ at different parameter values θ, while the rest is sampled from a reference distribution G. Note that G can be any distribution that dominates F_θ. If G is the marginal distribution $F_{\mathbf{x}}$ and the observed sample size is $n = 1$, then the denominator of the `BFF` statistic is exactly equal to one, allowing us to bypass the calculation of the denominator.

Algorithm 3 Generate a labeled sample of size B for estimating odds

Require: stochastic forward simulator F_θ; reference distribution G; proposal distribution π_Θ over parameter space; number of simulations B for estimating the odds; parameter p of Bernoulli distribution

Ensure: labeled training sample

Set $\mathcal{T} \leftarrow \emptyset$
for i in $\{1, ..., B\}$ **do**
    Draw parameter value $\theta_i \sim \pi_\Theta$
    Draw $Y_i \sim \text{Ber}(p)$
    if $Y_i == 1$ **then**
        Draw sample $\mathbf{X}_i \sim F_{\theta_i}$
    else
        Draw sample $\mathbf{X}_i \sim G$
    end if
    $\mathcal{T} \leftarrow \mathcal{T} \cup (\theta_i, \mathbf{X}_i, Y_i)$
end for
return $\mathcal{T} = \{\theta_i, \mathbf{X}_i, Y_i\}_{i=1}^B$

Algorithm 4 Sample from the marginal distribution $F_{\mathbf{X}}$

Require: stochastic forward simulator F_θ; proposal distribution π_Θ over parameter space

Ensure: sample $\mathbf{X}_i$ from the marginal distribution $F_{\mathbf{X}}$

Draw parameter value $\theta_i \sim \pi_\Theta$
Draw sample $\mathbf{X}_i \sim F_{\theta_i}$
return $\mathbf{X}_i$

B. Estimating P-Values

Algorithm 5 Estimate the p-values $p(D;\theta_0)$, given observed data D, for a level-α test of $H_{0,\theta_0} : \theta = \theta_0$ vs. $H_{1,\theta_0} : \theta \neq \theta_0$ for all $\theta_0 \in \Theta$ simultaneously.

Require: observed data D; stochastic forward simulator F_θ; sample size B' for p-value estimation; π_Θ (a fixed proposal distribution over the parameter space Θ); test statistic λ; regression estimator
Ensure: estimated p-value $\widehat{p}(D;\theta)$ for all $\theta = \theta_0 \in \Theta$

Set $\mathcal{T}' \leftarrow \emptyset$
for i in $\{1, \ldots, B'\}$ **do**
    Draw parameter $\theta_i \sim \pi_\Theta$
    Draw sample $\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n} \overset{iid}{\sim} F_{\theta_i}$
    Compute test statistic $\lambda_i \leftarrow \lambda((\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n}); \theta_i)$
    Compute indicator $Z_i \leftarrow \mathbb{I}(\lambda_i < \lambda(D;\theta_i))$
    $\mathcal{T}' \leftarrow \mathcal{T}' \cup \{(\theta_i, Z_i)\}$
end for
Use $\mathcal{T}'$ to learn the p-value function $\widehat{p}(D;\theta) := \widehat{\mathbb{E}}[Z|\theta]$ via regression of Z on θ.
return $\widehat{p}(D;\theta_0)$

C. Constructing Confidence Sets

Algorithm 6 details the construction of the `ACORE` and `BFF` confidence sets defined in Section 3 (the algorithm based on p-value estimation is analogous). Then, Algorithm 7 details the construction of the (hybrid) `ACORE` and `BFF` confidence sets defined in Section 6 for the general setting with nuisance parameters. Note that the first chunk on estimating the odds and the last chunk with Neyman inversion are the same for `ACORE` and `BFF`. Furthermore, the test statistics are the same whether or not there are nuisance parameters.

Algorithm 6 Construct confidence set for θ with coverage $1-\alpha$ (no nuisance parameters; that is, $\Theta = \Phi$)

Require: stochastic forward simulator F_θ; proposal distribution π over Θ; parameter p of Bernoulli distribution; sample size B (for estimating odds ratios); sample size B' (for estimating critical values or p-values); probabilistic classifier; observed data $D = \left\{\mathbf{x}_1^{\text{obs}}, \ldots, \mathbf{x}_n^{\text{obs}}\right\}$; desired level $\alpha \in (0,1)$; number of parameter values at which to evaluate confidence set, n_{grid}; test statistic `test_stat` (`ACORE` or `BFF`)

Ensure: θ-values in confidence set

// Estimate odds
Generate labeled sample $\mathcal{T}$ according to Algorithm 3
Apply probabilistic classifier to $\mathcal{T}$ to learn $\widehat{\mathbb{P}}(Y = 1|\theta, \mathbf{X})$, for all $\theta \in \Theta$ and $\mathbf{X} \in \mathcal{X}$
Let the estimated odds $\widehat{\mathbb{O}}(\mathbf{X};\theta) \leftarrow \frac{\widehat{\mathbb{P}}(Y=1|\theta,\mathbf{X})}{\widehat{\mathbb{P}}(Y=0|\theta,\mathbf{X})}$

// Compute cut-offs for `ACORE` or `BFF`
if `test_stat` == `ACORE` **then**
    Define the test statistic $\lambda(\mathcal{D};\theta) \leftarrow \widehat{\Lambda}(\mathcal{D};\theta)$, for every θ, to be the `ACORE` statistic (Equation 8) with estimated odds
else if `test_stat` == `BFF` **then**
    Define the test statistic $\lambda(\mathcal{D};\theta) \leftarrow \widehat{\tau}(\mathcal{D};\theta)$, for every θ, to be the `BFF` statistic (Equation 10) with estimated odds
end if
Learn critical values $\widehat{C}_\theta$ according to Algorithm 1

// Confidence sets for θ via Neyman inversion
Initialize confidence set $\widehat{R}(D) \leftarrow \emptyset$
Let $\mathrm{L}_\Theta \leftarrow$ lattice over Θ with n_{grid} elements
for $\theta_0 \in \mathrm{L}_\Theta$ **do**
    if $\lambda(D;\theta_0) > \widehat{C}_{\theta_0}$ **then**
        $\widehat{R}(D) \leftarrow \widehat{R}(D) \cup \{\theta_0\}$
    end if
end for
return confidence set $\widehat{R}(D)$

Algorithm 7 Construct confidence set for ϕ under the presence of nuisance parameters with (approximate) coverage $1-\alpha$

Require: stochastic forward simulator F_θ; proposal distribution π over $\Theta = \Phi \times \Psi$; parameter p of Bernoulli distribution; sample size B (for estimating odds ratios); sample size B' (for estimating critical values or p-values); probabilistic classifier; observed data $D = \left\{\mathbf{x}_1^{\text{obs}}, \ldots, \mathbf{x}_n^{\text{obs}}\right\}$; desired level $\alpha \in (0,1)$; number of parameter values at which to evaluate confidence set, n_{grid}; test statistic `test_stat` (`ACORE` or `BFF`)

Ensure: ϕ-values in confidence set

// Estimate odds
Generate labeled sample $\mathcal{T}$ according to Algorithm 3
Apply probabilistic classifier to $\mathcal{T}$ to learn $\widehat{\mathbb{P}}(Y = 1|\theta, \mathbf{X})$, for all $\theta = (\phi, \psi) \in \Theta$ and $\mathbf{X} \in \mathcal{X}$
Let the estimated odds $\widehat{\mathbb{O}}(\mathbf{X}; \theta) \leftarrow \frac{\widehat{\mathbb{P}}(Y=1|\theta,\mathbf{X})}{\widehat{\mathbb{P}}(Y=0|\theta,\mathbf{X})}$

// Compute (hybrid) cut-offs for (h-)`ACORE` or (h-)`BFF`
if `test_stat` == `ACORE` **then**
    Define $\widehat{\psi}_\phi \leftarrow \arg\max_{\psi \in \Psi} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{x}_i^{\text{obs}}; (\phi, \psi))$ for every ϕ
    Define the test statistic $\lambda(\mathcal{D}; \phi) \leftarrow \widehat{\Lambda}\left(\mathcal{D}; (\phi, \widehat{\psi}_\phi)\right)$, for every ϕ, to be the `ACORE` statistic (Equation 8) with estimated odds
    Generate $\mathcal{T}'$ according to Algorithm 1 using the proposal distribution $\pi'((\phi, \psi)) \propto \pi(\phi) \times \delta_{\widehat{\psi}_\phi}(\psi)$
    Apply quantile regression to $\mathcal{T}'$ to learn $\widehat{C}_\phi = \widehat{F}^{-1}_{\lambda(\mathcal{D};\phi)\big|\left(\phi, \widehat{\psi}_\phi\right)}(\alpha)$ for every ϕ
else if `test_stat` == `BFF` **then**
    Define $\pi_\Psi(\psi) \leftarrow$ restriction of proposal distribution π over Ψ
    Define the test statistic $\lambda(\mathcal{D}; \phi) \leftarrow \widehat{\tau}(\mathcal{D}; \phi) := \frac{\int_\Psi \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; (\phi, \psi)) d\pi_\Psi(\psi)}{\int_\Theta \prod_{i=1}^n \widehat{O}(\mathbf{X}_i^{\text{obs}}; \theta) d\pi_1(\theta)}$, for every ϕ, to be the `BFF` statistic (Equation 10) with estimated odds
    Generate $\mathcal{T}'$ according to Algorithm 1
    Apply quantile regression to $\mathcal{T}'$ to learn $\widehat{C}_\phi = \widehat{F}^{-1}_{\lambda(\mathcal{D};\phi)\big|(\phi)}(\alpha)$
end if

// Confidence sets for ϕ via Neyman inversion
Initialize confidence set $\widehat{R}(D) \leftarrow \emptyset$
Let $\mathrm{L}_\Phi \leftarrow$ lattice over Φ with n_{grid} elements
for $\phi_0 \in \mathrm{L}_\Phi$ **do**
    if $\lambda(D; \phi_0) > \widehat{C}_{\phi_0}$ **then**
        $\widehat{R}(D) \leftarrow \widehat{R}(D) \cup \{\phi_0\}$
    end if
end for
return confidence set $\widehat{R}(D)$

D. Theoretical Guarantees of power for ACORE with Calibrated Critical Values

Next, we show, for finite Θ, that as long as the probabilistic classifier is consistent and the critical values are well estimated (which holds for large B' according to Theorem 8), the power of the ACORE test converges to the power of the LRT as B grows.

Theorem 7 *For each $C \in \mathbb{R}$, let $\widehat{\phi}_{B,C}(\mathcal{D})$ be the test based on the ACORE statistic $\widehat{\Lambda}_B$ with critical value C[5] for number of simulations B in Algorithm 3. Moreover, let $\phi_C(\mathcal{D})$ be the likelihood ratio test with critical value C. If, for every $\theta \in \Theta$, the probabilistic classifier is such that*

$$\widehat{\mathbb{P}}(Y = 1|\theta, \mathbf{X}) \xrightarrow[B\longrightarrow\infty]{\mathbb{P}} \mathbb{P}(Y = 1|\theta, \mathbf{X}),$$

where $|\Theta| < \infty$, and $\widehat{C}_B$ is chosen such that $\widehat{C}_B \xrightarrow[B\longrightarrow\infty]{Dist} C$ for a given $C \in \mathbb{R}$, then, for every $\theta \in \Theta$,

$$\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}\left(\widehat{\phi}_{B,\widehat{C}_B}(\mathcal{D}) = 1\right) \xrightarrow[B\longrightarrow\infty]{} \mathbb{P}_{\mathcal{D}|\theta}\left(\phi_C(\mathcal{D}) = 1\right).$$

Proof Because $\widehat{\mathbb{P}}(Y = 1|\theta, \mathbf{X}) \xrightarrow[B\longrightarrow\infty]{\mathbb{P}} \mathbb{P}(Y = 1|\theta, \mathbf{X})$, it follows directly from the properties of convergence in probability that for every $\theta_0, \theta_1 \in \Theta$

$$\sum_{i=1}^{n} \log\left(\widehat{\mathbb{OR}}(\mathbf{X}_i^{\text{obs}}; \theta_0, \theta_1)\right) \xrightarrow[B\longrightarrow\infty]{\mathbb{P}} \sum_{i=1}^{n} \log\left(\mathbb{OR}(\mathbf{X}_i^{\text{obs}}; \theta_0, \theta_1)\right).$$

The continuous mapping theorem implies that

$$\widehat{\Lambda}_B(\mathcal{D}; \Theta_0) \xrightarrow[B\longrightarrow\infty]{\mathbb{P}} \sup_{\theta_0\in\Theta_0} \inf_{\theta_1\in\Theta} \sum_{i=1}^{n} \log\left(\mathbb{OR}(\mathbf{X}_i^{\text{obs}}; \theta_0, \theta_1)\right),$$

and therefore $\widehat{\Lambda}_B(\mathcal{D}; \Theta_0)$ converges in distribution to $\sup_{\theta_0\in\Theta_0} \inf_{\theta_1\in\Theta} \sum_{i=1}^{n} \log\left(\mathbb{OR}(\mathbf{X}_i^{\text{obs}}; \theta_0, \theta_1)\right)$. Now, from Slutsky's theorem,

$$\begin{aligned}
&\widehat{\Lambda}_B(\mathcal{D}; \Theta_0) - \widehat{C}_B \\
&\quad\xrightarrow[B\longrightarrow\infty]{\text{Dist}} \sup_{\theta_0\in\Theta_0} \inf_{\theta_1\in\Theta} \sum_{i=1}^{n} \log\left(\mathbb{OR}(\mathbf{X}_i^{\text{obs}}; \theta_0, \theta_1)\right) - C.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}\left(\widehat{\phi}_{B,\widehat{C}_B}(\mathcal{D}) = 1\right) = \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}\left(\widehat{\Lambda}_B(\mathcal{D}; \Theta_0) - \widehat{C}_B \leq 0\right) \\
&\xrightarrow[B\longrightarrow\infty]{} \mathbb{P}_{\mathcal{D}|\theta}\Big(\sup_{\theta_0\in\Theta_0} \inf_{\theta_1\in\Theta} \sum_{i=1}^{n} \log\left(\mathbb{OR}(\mathbf{X}_i^{\text{obs}}; \theta_0, \theta_1)\right) - C \leq 0\Big) \\
&= \mathbb{P}_{\mathcal{D}|\theta}\left(\phi_C(\mathcal{D}) = 1\right),
\end{aligned}$$

where the last equality follows from Proposition 1. ■

5. That is, $\widehat{\phi}_{B,C}(\mathcal{D}) = 1 \iff \widehat{\Lambda}_B(\mathcal{D}; \Theta_0) < C$.

E. Additional Proofs

Proof [Proof of Proposition 1] Let $f(\mathbf{x}|\theta)$ the density of F_θ with respect to G. By Bayes rule,

$$\mathbb{O}(\mathbf{x};\theta) := \frac{\mathbb{P}(Y=1|\theta,\mathbf{x})}{\mathbb{P}(Y=0|\theta,\mathbf{x})} = \frac{f(\mathbf{x}|\theta)p}{(1-p)}.$$

If $\widehat{\mathbb{P}}(Y=1|\theta,\mathbf{x}) = \mathbb{P}(Y=1|\theta,\mathbf{x})$, then $\widehat{\mathbb{O}}(\mathbf{x};\theta_0) = \mathbb{O}(\mathbf{x};\theta_0)$. Therefore,

$$\begin{aligned}
\widehat{\tau}(\mathcal{D};\Theta_0) :=& \frac{\int_{\Theta_0}\prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta)d\pi_0(\theta)}{\int_{\Theta_1}\prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta)d\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0}\prod_{i=1}^n \mathbb{O}(\mathbf{X}_i^{\text{obs}};\theta)d\pi_0(\theta)}{\int_{\Theta_1}\prod_{i=1}^n \mathbb{O}(\mathbf{X}_i^{\text{obs}};\theta)d\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0}\prod_{i=1}^n \frac{f(\mathbf{X}_i^{\text{obs}}|\theta)p}{(1-p)}d\pi_0(\theta)}{\int_{\Theta_1}\prod_{i=1}^n \frac{f(\mathbf{X}_i^{\text{obs}}|\theta)p}{(1-p)}d\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0}\prod_{i=1}^n f(\mathbf{X}_i^{\text{obs}}|\theta)d\pi_0(\theta)}{\int_{\Theta_1}\prod_{i=1}^n f(\mathbf{X}_i^{\text{obs}}|\theta)d\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0}\mathcal{L}(\mathcal{D};\theta)d\pi_0(\theta)}{\int_{\Theta_1}\mathcal{L}(\mathcal{D};\theta)d\pi_1(\theta)} \\
=& \text{BF}(\mathcal{D};\Theta_0).
\end{aligned}$$

■

Proof [Proof of Theorem 2] The proof follows from applying the convergence rate to the last equation in the proof of Theorem 1. ■

Assumption 10 (Uniform consistency) *Let $\widehat{F}_{B'}(\cdot|\theta)$ be the estimated cumulative distribution function of the test statistic $\lambda(\mathcal{D};\Theta_0)$ conditional on θ based on a sample $\mathcal{T}'$ with size B' implied by the quantile regression, and let $F(\cdot|\theta)$ be its true conditional distribution. Assume that the quantile regression estimator is such that*

$$\sup_{\theta\in\Theta_0,\lambda\in\mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta) - F(\lambda|\theta)| \xrightarrow[B'\longrightarrow\infty]{\mathbb{P}} 0.$$

This assumption holds, for instance, for quantile regression forests (Meinshausen, 2006) under addional assumptions (see Proposition 3).

Proposition 2 *If, for every $\theta\in\Theta_0$, the quantile regression estimator is such that*

$$\sup_{\lambda\in\mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta) - F(\lambda|\theta)| \xrightarrow[B'\longrightarrow\infty]{\mathbb{P}} 0 \tag{21}$$

and either

- $|\Theta| < \infty$ *or,*
- Θ *is a compact subset of* $\mathbb{R}^d$*, and the function* $g_{B'}(\theta) = \sup_{t\in\mathbb{R}} |\widehat{F}_{B'}(t|\theta) - F(t|\theta)|$ *is almost surely continuous in* θ *and strictly decreasing in* B'*,*

then Assumption 10 holds.

Proof If $|\Theta| < \infty$, the union bound and Equation 26 imply that

$$\sup_{\theta\in\Theta_0} \sup_{\lambda\in\mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta) - F(\lambda|\theta)| \xrightarrow[B'\longrightarrow\infty]{\mathbb{P}} 0. \tag{22}$$

Similarly, by Dini's theorem, Equation 27 also holds if Θ is a compact subset of $\mathbb{R}^d$, and the function $g_{B'}(\theta)$ is continuous in θ and strictly decreasing in B'. ■

Theorem 8 *Let* $C_{B'} \in \mathbb{R}$ *be the critical value of the test based on a strictly continuous statistic* $\lambda(\mathcal{D}; \Theta_0)$ *chosen according to Algorithm 1 for a fixed* $\alpha \in (0, 1)$*. If the quantile estimator satisfies Assumption 10, then*

$$C_{B'} \xrightarrow[B'\longrightarrow\infty]{\mathbb{P}} C^*,$$

where C^* *is such that*

$$\sup_{\theta\in\Theta_0} \mathbb{P}_{\mathcal{D}|\theta}(\lambda(\mathcal{D}; \Theta_0) \leq C^*) = \alpha.$$

Proof Assumption 10 implies that

$$\sup_{\theta\in\Theta_0} |\widehat{F}^{-1}_{B'}(\alpha|\theta) - F^{-1}(\alpha|\theta)| \xrightarrow[B'\longrightarrow\infty]{\mathbb{P}} 0.$$

The result then follows from the fact that

$$\begin{aligned} 0 \leq |C_{B'} - C^*| &= |\sup_{\theta\in\Theta_0} \widehat{F}^{-1}_{B'}(\alpha|\theta) - \sup_{\theta\in\Theta_0} F^{-1}(\alpha|\theta)| \\ &\leq \sup_{\theta\in\Theta_0} |\widehat{F}^{-1}_{B'}(\alpha|\theta) - F^{-1}(\alpha|\theta)|, \end{aligned}$$

and thus

$$|C_{B'} - C^*| \xrightarrow[B'\longrightarrow\infty]{\mathbb{P}} 0.$$

■

Lemma 1 *Let* $g_1, g_2, \ldots$ *be a sequence of random functions such that* $g_i : \mathcal{Z} \longrightarrow \mathbb{R}$*, and let* Z *be a random quantity defined over* $\mathcal{Z}$*, independent of the random functions. Assume that* $g(Z)$ *is strictly continuous. If, for every* $z \in \mathcal{Z}$*,*

$$g_m(z) \xrightarrow[m\longrightarrow\infty]{a.s.} g(z),$$

then

$$g_m(Z) \xrightarrow[m\longrightarrow\infty]{\mathcal{L}} g(Z).$$

Proof Fix $y \in \mathbb{R}$ and let $A_y = \{z \in \mathcal{Z} : g(z) \neq y\}$. Notice that $\mathbb{P}(Z \in A_y) = 1$. Moreover, the almost sure convergence of $g_m(z)$ implies its convergence in distribution. It follows that for every $z \in A_y$,

$$\lim_m \mathbb{P}(g_m(z) \leq y) = \mathbb{P}\left(g(z) \leq y\right). \tag{23}$$

Now, using Equation 23 and Lebesgue's dominated convergence theorem, notice that

$$\begin{aligned}
\lim_m \mathbb{P}(g_m(Z) < y) &= \lim_m \int_{\mathcal{Z}} \mathbb{P}(g_m(Z) < y | Z = z) d\mathbb{P}_Z(z) = \int_{\mathcal{Z}} \lim_m \mathbb{P}(g_m(Z) < y | Z = z) d\mathbb{P}_Z(z) \\
&= \int_{A_z} \lim_m \mathbb{P}(g_m(z) < y) d\mathbb{P}_Z(z) = \int_{A_z} \mathbb{P}(g(z) < y) d\mathbb{P}_Z(z) \\
&= \int_{\mathcal{Z}} \mathbb{P}(g(Z) < y | Z = z) d\mathbb{P}_Z(z) = \mathbb{P}(g(Z) < y),
\end{aligned}$$

which concludes the proof. ■

Proof [Proof of Theorem 3] Assumption 3 implies that, for every D,

$$\begin{aligned}
0 \leq |\widehat{p}(D; \Theta_0) - p(D; \Theta_0)| &= |\sup_{\theta \in \Theta_0} \widehat{p}(D; \theta) - \sup_{\theta \in \Theta_0} p(D; \theta)| \\
&\leq \sup_{\theta \in \Theta_0} |\widehat{p}(D; \theta) - p(D; \theta)| \xrightarrow[B' \longrightarrow \infty]{\text{a.s.}} 0,
\end{aligned}$$

and therefore $\widehat{p}(D; \Theta_0)$ converges almost surely to $p(D; \Theta_0)$. It follows from Lemma 1 that $\widehat{p}(\mathcal{D}; \Theta_0)$ converges in distribution to $p(\mathcal{D}; \Theta_0)$. Conclude that

$$\mathbb{P}_{\mathcal{D}, \mathcal{T}' | \theta}(\widehat{p}(\mathcal{D}; \Theta_0) \leq \alpha) = F_{\widehat{p}(\mathcal{D}; \Theta_0) | \theta}(\alpha) \xrightarrow{B' \longrightarrow \infty} F_{p(\mathcal{D}; \Theta_0) | \theta}(\alpha) = \mathbb{P}_{\mathcal{D} | \theta}(p(D; \Theta_0) \leq \alpha),$$

where F_Z denotes the cumulative distribution function of the random variable Z. ■

Proof [Proof of Corollary 1] Fix $\theta \in \Theta$. Because F_θ is continuous, the definition of $p(\mathcal{D}; \theta)$ implies that its distribution is uniform under the null. Thus $\mathbb{P}_{\mathcal{D} | \theta}\left(p(\mathcal{D}; \theta) \leq \alpha\right) = \alpha$. Theorem 3 therefore implies that

$$\mathbb{P}_{\mathcal{D}, \mathcal{T}' | \theta}(\widehat{p}(\mathcal{D}; \theta) \leq \alpha) \xrightarrow{B' \longrightarrow \infty} \mathbb{P}_{\mathcal{D} | \theta}\left(p(\mathcal{D}; \theta) \leq \alpha\right) = \alpha. \tag{24}$$

Now, for any $\theta \in \Theta_0$, uniformity of the p-value implies that

$$\begin{aligned}
\mathbb{P}_{\mathcal{D} | \theta}(p(\mathcal{D}; \Theta_0) \leq \alpha) &= \mathbb{P}_{\mathcal{D} | \theta}\left(\sup_{\theta_0 \in \Theta_0} p(\mathcal{D}; \theta_0) \leq \alpha\right) \leq \mathbb{P}_{\mathcal{D} | \theta}\left(p(\mathcal{D}; \theta) \leq \alpha\right) \\
&= \alpha.
\end{aligned}$$

Conclude from Theorem 3 that

$$\mathbb{P}_{\mathcal{D}, \mathcal{T}' | \theta}(\widehat{p}(\mathcal{D}; \Theta_0) \leq \alpha) \xrightarrow{B' \longrightarrow \infty} \mathbb{P}_{\mathcal{D} | \theta}(p(\mathcal{D}; \Theta_0) \leq \alpha) \leq \alpha. \tag{25}$$

The conclusion follows from putting together Equations 24 and 25. ■

Proof [Proof of Theorem 4]

$$\begin{aligned}|\widehat{p}(D;\Theta_0) - p(D;\Theta_0)| &= |\sup_{\theta\in\Theta_0}\widehat{p}(D;\theta) - \sup_{\theta\in\Theta_0} p(D;\theta)| \\ &\leq \sup_{\theta\in\Theta_0}|\widehat{p}(D;\theta) - p(D;\theta)| \\ &= O_P\left(\left(\frac{1}{B'}\right)^r\right),\end{aligned}$$

where the last line follows from Assumption 4 ■

Proposition 3 *If, for every $\theta \in \Theta_0$, the quantile regression estimator is such that*

$$\sup_{\lambda\in\mathbb{R}}|\widehat{F}_{B'}(\lambda|\theta) - F(\lambda|\theta)| \xrightarrow[B'\longrightarrow\infty]{\mathbb{P}} 0 \tag{26}$$

and either

- $|\Theta| < \infty$ *or,*
- Θ *is a compact subset of* $\mathbb{R}^d$*, and the function* $g_{B'}(\theta) = \sup_{t\in\mathbb{R}}|\widehat{F}_{B'}(t|\theta) - F(t|\theta)|$ *is almost surely continuous in* θ *and strictly decreasing in* B'*,*

then Assumption 10 holds.

Proof If $|\Theta| < \infty$, the union bound and Equation 26 imply that

$$\sup_{\theta\in\Theta_0}\sup_{\lambda\in\mathbb{R}}|\widehat{F}_{B'}(\lambda|\theta) - F(\lambda|\theta)| \xrightarrow[B'\longrightarrow\infty]{\mathbb{P}} 0. \tag{27}$$

Similarly, by Dini's theorem, Equation 27 also holds if Θ is a compact subset of $\mathbb{R}^d$, and the function $g_{B'}(\theta)$ is continuous in θ and strictly decreasing in B'. ■

Lemma 2 *Under Assumption 6,* $\int(\mathbb{O}(\mathbf{x};\theta_0) - \widehat{\mathbb{O}}(\mathbf{x};\theta_0))^2 dG(\mathbf{x}) \leq \frac{M'}{m'}L(\widehat{\mathbb{O}},\mathbb{O})$.

Proof Let h be as in Assumption 6. Notice that

$$\begin{aligned}\int(\mathbb{O}(\mathbf{x};\theta_0) - \widehat{\mathbb{O}}(\mathbf{x};\theta_0))^2 dG(\mathbf{x}) &\leq \sup_{\theta\in\Theta}\int(\mathbb{O}(\mathbf{x};\theta) - \widehat{\mathbb{O}}(\mathbf{x};\theta))^2 dG(\mathbf{x}) \\ &\leq M' \leq \frac{M'}{m'}\int h(\theta)d\pi(\theta) \\ &= \frac{M'}{m'}L(\widehat{\mathbb{O}},\mathbb{O}),\end{aligned}$$

which concludes the proof.

■

Lemma 3 *Under Assumptions 5 and 6, there exists $K > 0$ such that*

$$\mathbb{E}_{\mathcal{D}|\theta,T}\left[|\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)|\right] \leq K\sqrt{L(\widehat{\mathbb{O}},\mathbb{O})}.$$

Proof For every $\theta \in \Theta$

$$\begin{aligned}
\mathbb{E}^2_{\mathcal{D}|\theta,T}[|\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)|] &= \left(\int |\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)|\ dF(\mathbf{x}|\theta)\right)^2 \\
&= \left(\int |\mathbb{O}(\mathbf{x};\theta_0) - \widehat{\mathbb{O}}(\mathbf{x};\theta_0)|\ dF(\mathbf{x}|\theta)\right)^2 \\
&= \left(\int |\mathbb{O}(\mathbf{x};\theta_0) - \widehat{\mathbb{O}}(\mathbf{x};\theta_0)|\mathbb{O}(\mathbf{x};\theta) dG(\mathbf{x})\right)^2 \\
&\leq \left(\int (\mathbb{O}(\mathbf{x};\theta_0) - \widehat{\mathbb{O}}(\mathbf{x};\theta_0)^2 dG(\mathbf{x})\right)\left(\int \mathbb{O}^2(\mathbf{x};\theta) dG(\mathbf{x})\right),
\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz. Assumption 5 implies that

$$\int \mathbb{O}^2(\mathbf{x};\theta) dG(\mathbf{x}) \leq M^2,$$

from which we conclude that

$$\mathbb{E}^2_{\mathcal{D}|\theta,T}[|\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)|] \leq M^2 \int (\mathbb{O}(\mathbf{x};\theta_0) - \widehat{\mathbb{O}}(\mathbf{x};\theta_0))^2 dG(\mathbf{x}).$$

Conclude from Lemma 2 that

$$\mathbb{E}^2_{\mathcal{D}|\theta,T}[|\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)|] \leq K^2 \cdot L(\widehat{\mathbb{O}},\mathbb{O}),$$

where $K = M\sqrt{\frac{M'}{m'}}$. ■

Lemma 4 *Under Assumptions 5-9, there exists $C > 0$ such that*

$$\mathbb{E}_{\mathcal{D},\mathcal{T}|\theta}\left[|\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)|\right] \leq CB^{-\kappa/(2(\kappa+d))}.$$

Proof Let $\widehat{p} = \widehat{\mathbb{P}}(Y = 1|\mathbf{x},\theta)$ and $p = \mathbb{P}(Y = 1|\mathbf{x},\theta)$ be the probabilistic classifier and true classification function, respectively, on the training sample T. Let $h(y) = \frac{y}{1-y}$ for $0 < y < 1$. A Taylor expansion of h implies that

$$(h(\widehat{p}) - h(p))^2 = (h(p) + R_1(\widehat{p}) - h(p))^2 = R_1(\widehat{p})^2,$$

where $R_1(\widehat{p}) = h'(\xi)(\widehat{p} - p)$ for some ξ between p and $\widehat{p}$. Also note that due to Assumption 5,

$$\exists a > 0 \text{ s.t. } p, \widehat{p} > a,\ \forall x \in \mathcal{X}, \theta \in \Theta.$$

Thus,

$$\begin{aligned}
\mathbb{E}_{\mathcal{T}}\left[\int \left(h(\widehat{p})-h(p)\right)^2 dG(\mathbf{x})d\pi(\theta)\right] &= \mathbb{E}_{\mathcal{T}}\left[\int \frac{1}{(1-\xi)^4}\left(\widehat{p}-p\right)^2 dG(\mathbf{x})d\pi(\theta)\right] \\
&\leq \frac{1}{(1-a)^4}\mathbb{E}_{\mathcal{T}}\left[\int \left(\widehat{p}-p\right)^2 dG(\mathbf{x})d\pi(\theta)\right] \\
&= \frac{1}{(1-a)^4}\mathbb{E}_{\mathcal{T}}\left[\int \left(\widehat{\mathbb{P}}(Y=1|\mathbf{x},\theta)-\mathbb{P}(Y=1|\mathbf{x},\theta)\right)^2 h'(\mathbf{x},\theta)dH(\mathbf{x},\theta)\right] \\
&\leq \frac{\gamma}{(1-a)^4}\mathbb{E}_{\mathcal{T}}\left[\int \left(\widehat{\mathbb{P}}(Y=1|\mathbf{x},\theta)-\mathbb{P}(Y=1|\mathbf{x},\theta)\right)^2 dH(\mathbf{x},\theta)\right] \\
&= O\left(B^{-\kappa/(\kappa+d)}\right)
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D},\mathcal{T}|\theta}\left[\left|\tau(\mathcal{D};\theta_0)-\widehat{\tau}_B(\mathcal{D};\theta_0)\right|\right] &= \mathbb{E}_T\left[\mathbb{E}_{\mathcal{D}|\theta,\mathcal{T}}\left[\left|\tau(\mathcal{D};\theta_0)-\widehat{\tau}_B(\mathcal{D};\theta_0)\right|\right]\right] \\
&\leq \mathbb{E}_{\mathcal{T}}\left[K\sqrt{L(\widehat{\mathbb{O}},\mathbb{O})}\right] \\
&\leq K\sqrt{\mathbb{E}_{\mathcal{T}}\left[L(\widehat{\mathbb{O}},\mathbb{O})\right]} \\
&= K\sqrt{\mathbb{E}_{\mathcal{T}}\left[\int \left(h(\widehat{p})-h(p)\right)^2 dG(\mathbf{x})d\pi(\theta)\right]} \\
&= O\left(B^{-\kappa/(2(\kappa+d))}\right),
\end{aligned}$$

where the second inequality follows from Lemma 3. ■

Proof [Proof of Theorem 5] It follows from Markov's inequality and Lemma 3 that with probability at least $1-\epsilon$, $\mathcal{D}$ is such that

$$\left|\tau(\mathcal{D};\theta_0)-\widehat{\tau}(\mathcal{D};\theta_0)\right| \leq \frac{K\cdot\sqrt{L(\widehat{\mathbb{O}},\mathbb{O})}}{\epsilon} \tag{28}$$

Now we upper bound $\mathbb{P}_{\mathcal{D}|\theta,T}(\phi_\tau(\mathcal{D})\neq\phi_{\widehat{\tau}}(\mathcal{D}))$. Define A as the event that Eq. 28 happens. Then:

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}|\theta,T}(\phi_\tau(\mathcal{D})\neq\phi_{\widehat{\tau}}(\mathcal{D})) &\leq \mathbb{P}_{\mathcal{D}|\theta,T}(\phi_\tau(\mathcal{D})\neq\phi_{\widehat{\tau}}(\mathcal{D}),A)+\mathbb{P}_\theta(A^c) \\
&\leq \mathbb{P}_{\mathcal{D}|\theta,T}\left(\mathbb{I}\left(\tau(\mathcal{D};\theta_0)<c\right)\neq\mathbb{I}\left(\widehat{\tau}(\mathcal{D};\theta_0)<c\right),A\right)+\epsilon \\
&\leq \mathbb{P}_{\mathcal{D}|\theta,T}\left(c-\frac{K\cdot\sqrt{L(\widehat{\mathbb{O}},\mathbb{O})}}{\epsilon}<\tau(\mathcal{D};\theta_0)<c+\frac{K\cdot\sqrt{L(\widehat{\mathbb{O}},\mathbb{O})}}{\epsilon}\right)+\epsilon
\end{aligned}$$

Assumption 7 then implies that

$$\mathbb{P}_{\mathcal{D}|\theta,T}(\phi_\tau(\mathcal{D})\neq\phi_{\widehat{\tau}}(\mathcal{D})) \leq \frac{K'\cdot\sqrt{L(\widehat{\mathbb{O}},\mathbb{O})}}{\epsilon}+\epsilon$$

where $K' = 2KC_L$, which concludes the proof. ■

Proof [Proof of Theorem 6] It follows from Markov's inequality and Lemma 4 that with probability at least $1 - \epsilon$, $\mathcal{D}$ is such that

$$|\tau(\mathcal{D}; \theta_0) - \widehat{\tau}(\mathcal{D}; \theta_0)| \leq \frac{CB^{-\kappa/(2(\kappa+d))}}{\epsilon} \tag{29}$$

Following the same reasoning as for Theorem 5, we obtain that

$$\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_\tau(\mathcal{D}) \neq \phi_{\widehat{\tau}}(\mathcal{D})) \leq \frac{K''B^{-\kappa/(2(\kappa+d))}}{\epsilon} + \epsilon$$

where $K'' = 2CC_L$. Notice that taking $\epsilon^* = \sqrt{K''}B^{-\kappa/(4(\kappa+d))}$ optimizes the bound and gives the result. ■

Proof [Proof of Corollary 2] The result follows from noticing that

$$\begin{aligned}\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\widehat{\tau}_B}(\mathcal{D}) = 1) &\geq \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_\tau(\mathcal{D}) = 1) - \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_\tau(\mathcal{D}) \neq \phi_{\widehat{\tau}_B}(\mathcal{D})) \\ &\geq \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_\tau(\mathcal{D}) = 1) - 2\sqrt{K''}B^{-\kappa/(4(\kappa+d))},\end{aligned}$$

where the last inequality follows from Theorem 6. ■

F. Gaussian Mixture Model Example

Here we (i) provide details on the algorithms used to estimate critical values and coverage in Figure 3, (ii) show what the estimated conditional quantile functions look like for the three methods described in the same example, (iii) discuss results of experiments which account for asymmetric mixtures, and (iv) include results for applying p-value estimation to the problem in Section 7.1.

Regarding (i): The quantile regressor used to estimate C_{θ_0} is a neural network, with two hidden layers and 32×32 neurons, which minimizes the quantile loss. Our experiments showed that using quantile boosted regression trees led to equivalent results, but we opted for NNs due to their inherent smoothing capabilities, which resulted in stabler estimates of the conditional quantile function. The algorithm used to estimate coverage is a binomial Generalized Additive Model (GAM) with logit link and a smoothing spline applied to the independent variable, which is θ in that setting (see Algorithm 2). The two-standard-deviation ($\pm 2\sigma$) prediction intervals are based on the Bayesian posterior variance of the parameters in the fitted GAM object. See documentation of the R package `MGCV` for more details.

Regarding (ii): Figure 10 shows the estimated conditional quantile function, both via Monte-Carlo (MC) and via quantile regression. The plot also includes the upper α quantile of a

χ^2_1 distribution. Here B' and the number of MC simulations are both 5000[6], but the latter is again repeated for every θ_0 on a fine grid. The size of each simulated sample is $n = 1000$. Connecting this plot with the central panel at the bottom of Figure 3, we can see that the "Chi-square LRT" only achieves nominal coverage in a neighborhood of $\theta = 0$, where the three curves in Figure 10 are close. As they diverge, Neyman Inversion for "Chi-square LRT" fails to include the true parameter most of the times.

Regarding (iii): So far the experiments have focused on symmetric mixtures, where both components have the same probability of being selected. We also repeated the above experiments with a mixing parameter equal to 0.8, i.e. when the mixture is strongly unbalanced towards one mixture component but is still bi-modal. In terms of coverage, the results were qualitatively the same as those obtained in the case of symmetric mixtures.

Regarding (iv): we conclude by showing that p-value estimation leads to confidence sets with correct conditional coverage, hence providing an alternative to critical value estimation via quantile regression. Figure 11 presents the results obtained on the symmetric Gaussian mixture model with samples of size $n = 10, 100, 1000$, which can be compared with the right panel in Figure 3. Although all examples achieve correct conditional coverage, it must be noted that p-values were estimated using $B' = 10000$ to train gradient boosted classification trees, instead of $B' = 1000$ used in Section 7.1 and above. In practice we have indeed observed that estimating p-values via Algorithm 5 requires more simulations than estimating critical values via Algorithm 1. Moreover, as already noted in Section 3.3.2, the procedure for p-value estimation has to be repeated for each observed sample D, while critical value estimation is amortized: once the quantile regressor is fitted, it can be used for any number of observed samples.

G. Multivariate Gaussian Example

Here we provide (i) the analytical derivations for the marginal distribution and Bayes factor in the multivariate Gaussian setting, and (ii) Section 7.2.2 details for the probabilistic classifier selection and the analysis of the drop in power for `ACORE` and `BFF` at $d = 5$ and $d = 10$.

G.1 Analytical Derivations

Given that the covariance matrix is $\Sigma = I_d$ in this setting, the marginal distribution $F_{\mathbf{X}}$ has a closed form solution for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, which can be expressed as follows:

6. Increased with respect to the value used in Section 7.1 just to make the MC and Quantile Regression curves smoother for visualization purposes. Coverage was achieved even at the previous $B' = 1000$.

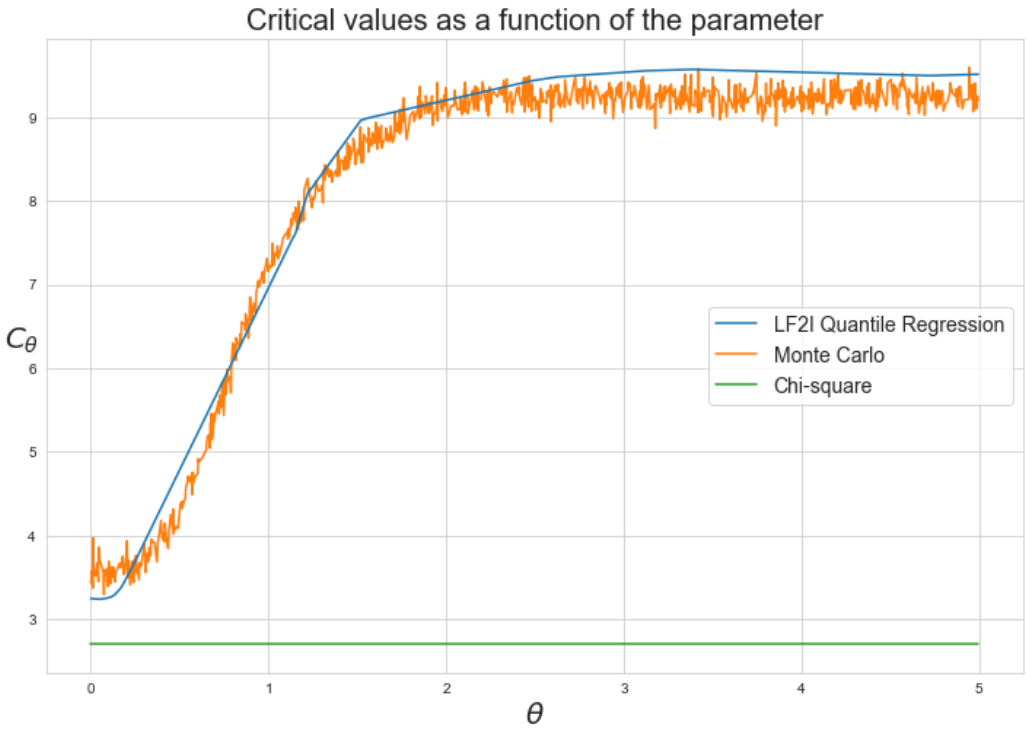


Figure 10: Conditional quantile functions estimated via Monte Carlo (*orange*) and quantile regression (*blue*). Both functions get closer to the upper α quantile of a χ^2_1 distribution (*green*) as $\theta \to 0$, but diverge as the mixture becomes bimodal. This is the reason why "Chi-square LRT" strongly undercovers in Section 7.1.

$$
\begin{aligned}
F_{\mathbf{X}}(\mathbf{x}) &= \int_{\mathbf{a}}^{\mathbf{b}} (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right) d\boldsymbol{\mu} \\
&= \int_{\mathbf{a}}^{\mathbf{b}} (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{d} x_i^2 - 2x_i\mu_i + \mu_i^2\right)\right) d\mu_1 d\mu_2 ... d\mu_d \\
&= \prod_{i=1}^{d} \left[\int_{a_i}^{b_i} (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}x_i^2 + x_i\mu_i - \frac{1}{2}\mu_i^2\right) d\mu_i\right] \\
&= \prod_{i=1}^{d} \frac{1}{2}\text{erf}\left(\frac{b_i - x_i}{\sqrt{2}}\right) - \frac{1}{2}\text{erf}\left(\frac{a_i - x_i}{\sqrt{2}}\right),
\end{aligned}
$$

In this setting, the proposal distribution π is uniform over an axis-aligned hyper-rectangle with extremes $\mathbf{a} = (a, ..., a)$ and $\mathbf{b} = (b, ..., b)$ for $a < b \in \mathbb{R}$. Since $\overline{\mathbf{X}}_n$ is a sufficient statistic, the exact Bayes factor for the Neyman construction when testing $H_{0,\theta_0} : \theta = \theta_0$ versus $H_{1,\theta_0} : \theta \neq \theta_0$ is equal to:

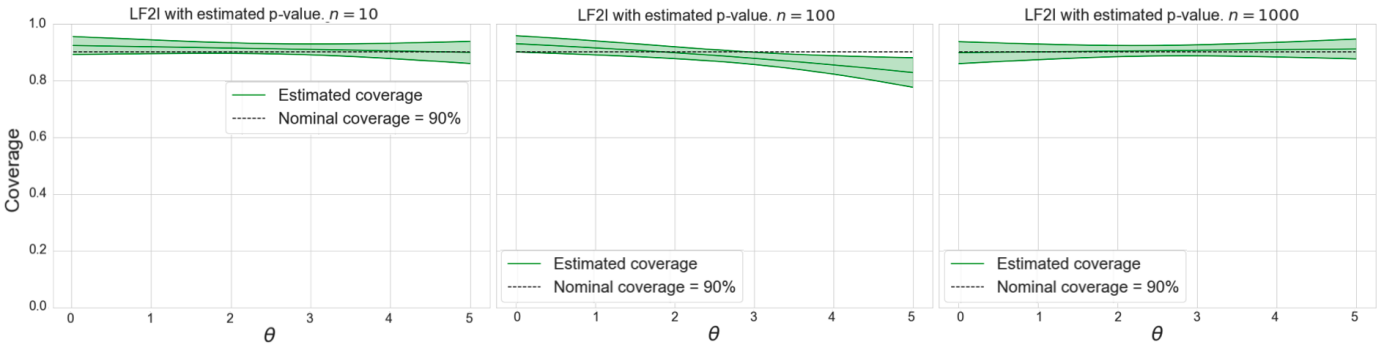


Figure 11: GMM example with sample size $n = 10$ (*left*), $n = 100$ (*center*) and $n = 1000$ (*right*) and confidence sets constructed using p-value estimation. The plots show the estimated coverage across Θ of 90% confidence sets for θ. As before, conditional coverage is estimated using the diagnostic branch of the LF2I framework.

$$\begin{aligned}
\mathrm{BF}(\mathcal{D};\theta_0) &= \frac{N(\overline{\mathbf{X}}_n;\theta_0,n^{-1}I_d)}{\int_{\mathbf{a}}^{\mathbf{b}} N(\overline{\mathbf{X}}_n;\theta,n^{-1}I_d)d\pi(\theta)} \\
&= \frac{N(\overline{\mathbf{X}}_n;\theta_0,n^{-1}I_d)}{\left(\frac{1}{b-a}\right)^d \int_{\mathbf{a}}^{\mathbf{b}} N(\overline{\mathbf{X}}_n;\theta,n^{-1}I_d)d\theta} \\
&= \frac{N(\overline{\mathbf{X}}_n;\theta_0,n^{-1}I_d)}{\left(\frac{1}{b-a}\right)^d \prod_{j=1}^d \left[\frac{1}{2}\mathrm{erf}\left(\frac{b-\overline{X}_{n,j}}{\sqrt{2n}}\right) - \frac{1}{2}\mathrm{erf}\left(\frac{a-\overline{X}_{n,j}}{\sqrt{2n}}\right)\right]},
\end{aligned}$$

where $\overline{X}_{n,j}$ is the j-th coordinate of $\overline{\mathbf{X}}_n$.

G.2 Section 7.2.2 Details

Figure 12 (left) compares cross-entropy loss curves for the QDA (the best classifier for the Gaussian likelihood model) and MLP classifiers. As we increase B, odds estimation becomes more accurate, and we expect to see a decrease in both cross-entropy loss and integrated odds loss, as shown in Figure 12 (right).

We showed in Section 4 that the power of `BFF` is bounded by the integrated odds loss. In practice, this loss may be more stably estimated for larger B, which would make it an attractive alternative to the cross-entropy loss. The performance difference in Figure 12 is reflected in Figure 13, highlighting the importance of choosing the best fitting classifier.

To pinpoint the cause of the degradation in power in high dimensions for `ACORE` and `BFF` in Section 7.2.2, we separate the error in estimating the odds from the numerical error in the maximization or integration step for the test statistic (errors e_1 and e_2 in Section 5). Figure 13 shows that the QDA estimation error is negligible at both $d = 5$ and $d = 10$ (as opposed to MLP estimation error). To isolate the numerical error, Figure 14 shows

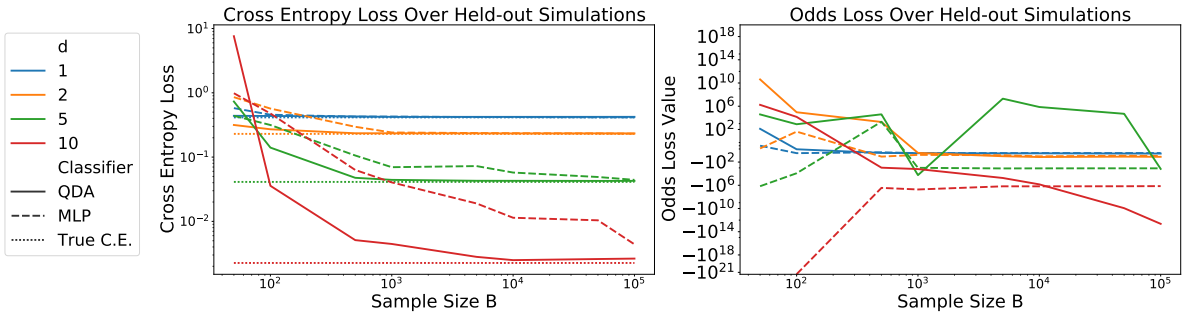


Figure 12: *Left*: Cross-entropy loss in learning the odds versus the sample size B (Algorithm 3) for a QDA and MLP classifier, as well as the true cross entropy, for the Gaussian likelihood model in dimensions $d = 1, 2, 5$ and 10. QDA has the lowest cross-entropy loss among the classifiers we considered (of which MLP is one example). The values B at which the cross entropy plateaus are used as the sample sizes for learning the odds at various dimensions. Right: The integrated odds loss generally decreases with increasing B, as expected, though it is noisier (the presence of small probabilities blows up the odds ratio). For larger values of B, the integrated odds loss should be more stable.

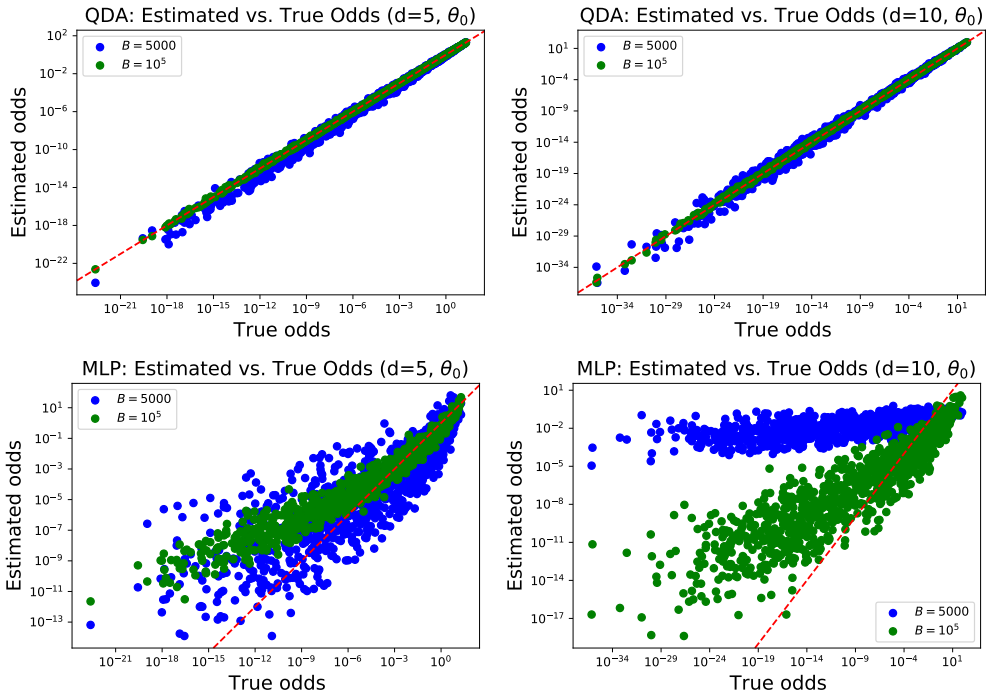


Figure 13: Odds classifiers trained on B samples, evaluated on 1000 test samples. QDA (top row) fits better than MLP (bottom row), and QDA with $B = 10^5$ fits well.

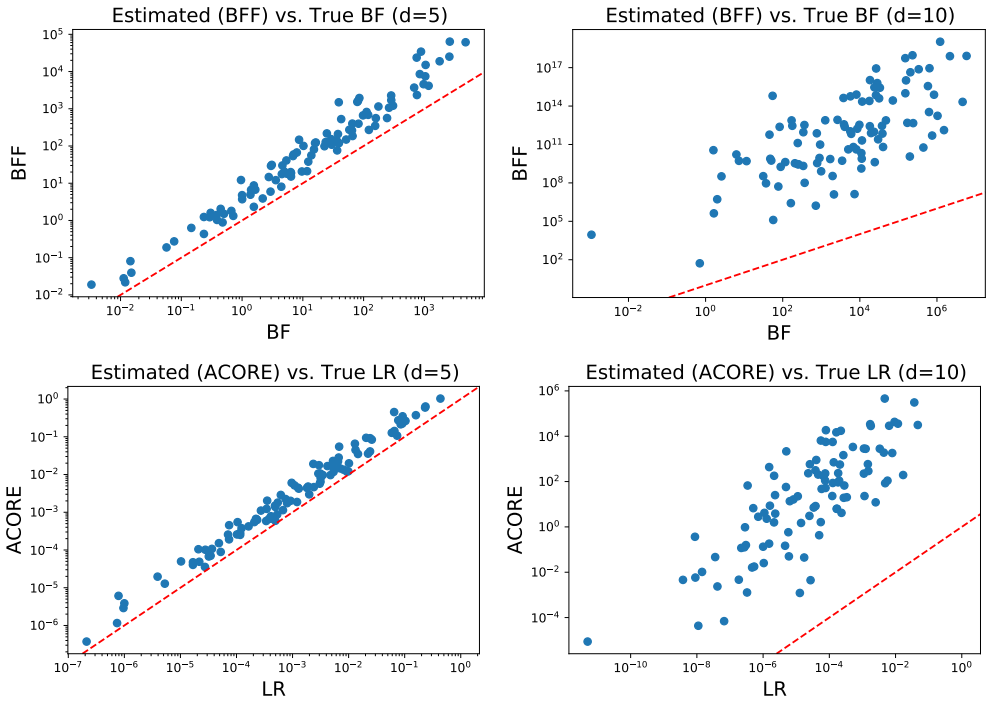


Figure 14: We estimate the `BFF` and `ACORE` test statistics using exact odds, so the only error is due to numerical estimation of the denominator with $N = 30000$ uniform samples. We see that as d grows, this numerical estimation quickly becomes imprecise, even for large values of N.

the estimated `ACORE` and `BFF` statistics using the analytical odds function. Even with a large budget of $M = 30000$, we underestimate both the odds maximum and the integrated odds across the parameter space, resulting in an over-estimation of the `ACORE` and `BFF` test statistics.

H. Computational Stability for `BFF`

When computing the `BFF` statistics for the Neyman construction hypothesis testing, the denominator is approximated by an average in the following way:

$$\tau(\mathcal{D};\theta_0) := \frac{\prod_{i=1}^n \mathbb{O}(\mathbf{X}_i;\theta_0)}{\int_\Theta \left(\prod_{i=1}^n \mathbb{O}(\mathbf{X}_i;\theta)\right) d\pi(\theta)} \approx \frac{\prod_{i=1}^n \mathbb{O}(\mathbf{X}_i;\theta_0)}{\frac{1}{m}\sum_{j=1}^m \prod_{i=1}^n \mathbb{O}(\mathbf{X}_i;\theta_j)},$$

where $\theta_j \sim \pi(\theta)$ for $j = 1, ..., m$. In practice, the product of odds can quickly run into overflow/underflow. If one assumes $m \leq \mathbb{O}(\mathbf{X}_i;\theta_j) \leq M$ for all X_i, θ_j, the product over n samples can range from $m^n \leq \prod_{i=1}^n \mathbb{O}(\mathbf{X}_i;\theta_j) \leq M^n$ which could be below or above machine precision depending on the values of m and M respectively. Running computations in log-space provides computationally stable calculations even for large samples. First, we can express the test statistic approximation in the following way:

$$\tau(\mathcal{D};\theta_0) \approx \frac{\prod_{i=1}^{n}\mathbb{O}(\mathbf{X}_i;\theta_0)}{\frac{1}{m}\sum_{j=1}^{m}\prod_{i=1}^{n}\mathbb{O}(\mathbf{X}_i;\theta_j)} = \frac{\exp^{\sum_{i=1}^{n}\log(\mathbb{O}(\mathbf{X}_i;\theta_0))}}{\frac{1}{m}\sum_{j=1}^{m}\exp^{\sum_{i=1}^{n}\log(\mathbb{O}(\mathbf{X}_i;\theta_j))}}.$$

Let $\psi^0 = \sum_{i=1}^{n}\log(\mathbb{O}(\mathbf{X}_i;\theta_0))$ and $\psi_j = \sum_{i=1}^{n}\log(\mathbb{O}(\mathbf{X}_i;\theta_j))$. Computing the log-space version of the `BFF` test statistics then leads to

$$\log(\tau(\mathcal{D};\theta_0)) = \psi^0 - \log\left(\frac{1}{m}\sum_{j=1}^{m}\exp^{\psi_j}\right) = \psi^0 + \log(m) - \log\left(\sum_{j=1}^{m}\exp^{\psi_j}\right).$$

The above can be made computationally stable by using any of the "log-sum-exp" implementations available (such as in `SciPy`, Virtanen et al. (2020)).