Validation of Approximate Likelihood and Emulator Models for Computationally Intensive Simulations

Niccolò Dalmasso[1], **Ann B. Lee**[1], **Rafael Izbicki**[2], **Taylor Pospisil**[1,3], **Ilmun Kim**[1], **Chieh-An Lin**[4]
[1] Department of Statistics & Data Science, Carnegie Mellon University University
[2] Department of Statistics, Federal University of São Carlos
[3] Google LLC, [4] Institute for Astronomy, University of Edinburgh

Abstract

Complex phenomena in engineering and the sciences are often modeled with computationally intensive feed-forward simulations for which a tractable analytic likelihood does not exist. In these cases, it is sometimes necessary to estimate an approximate likelihood or fit a fast emulator model for efficient statistical inference; such surrogate models include Gaussian synthetic likelihoods and more recently neural density estimators such as autoregressive models and normalizing flows. To date, however, there is no consistent way of quantifying the quality of such a fit. Here we propose a statistical framework that can distinguish any arbitrary misspecified model from the target likelihood, and that in addition can identify with statistical confidence the regions of parameter as well as feature space where the fit is inadequate. At the heart of our approach is a two-sample test that quantifies the quality of the fit at fixed parameter values, and a global test that assesses goodness-of-fit across simulation parameters. While our general framework can incorporate any test statistic or distance metric, we specifically argue for a new two-sample test that can leverage any regression method to attain high power and provide diagnostics in complex data settings. Software for our approach is available on GitHub in `Python` and `R`.

1 Introduction

The likelihood function $\mathcal{L}(\mathbf{x};\theta)$ links the unknown components θ of the data-generating mechanism with the observable data $\mathbf{x}$ and is a key component for performing statistical inference over parameters of interest. For complex phenomena, there is often no tractable analytical form for the likelihood; many times such phenomena are instead studied using numerical simulators derived from the underlying physical or biological processes, which encode, e.g, complex observational effects, selection biases, etc. In situations where the likelihood function cannot be easily evaluated, but a stochastic numerical simulator (which serves as the ground truth) is available, approximate inference of parameters of interest is possible. Tools that explore feed-forward simulations to infer θ without requiring explicit likelihoods are referred to as *likelihood-free inference* (LFI) methods, of which Approximate Bayesian Computation (ABC) (Beaumont et al., 2002; Marin et al., 2012) is the best known approach. Several variations of ABC methods exist and have resulted in many successful applications; see Sisson et al. (2018) for a review.

However, there is a growing number of disciplines where accurate analyses require highly realistic and computationally intensive simulations. In such cases, it may not be feasible to repeatedly generate new simulations at different parameter settings as generally required by ABC methods. Instead, a common practice is to run the simulator only for a few points in parameter space, in a format of batches or ensembles, where an *ensemble* is a collection of multiple realizations (e.g., corresponding to different initial conditions) of the same physical model (i.e. they all share the same θ). For example, modern climate and weather forecasting models (e.g., CESM (Hurrell et al., 2013)) often incorporate complex representations of the atmosphere, ocean, land, ice, etc, on fine spatial and temporal resolutions across the entire world. These models are commonly run as an ensemble of dynamical simulations with different initial conditions, where each simulation can take weeks to compile on supercomputer clusters (see Baker et al. (2015); Kay et al. (2015) and references within). Similarly, cosmological N-body simulations, which compute gravity between particle pairs, are equally costly and often either created at a fixed cosmology (parameter

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

value θ) (Abbott et al., 2016; Hildebrandt et al., 2017), or on a sparse grid of a few carefully chosen parameter values (Kacprzak et al., 2016; Gupta et al., 2018).

Given the above scenario, a solution to make inference feasible is to approximate the computationally expensive simulator with a faster *emulator* model that can speed up probabilistic modeling by several orders of magnitude. Some common models, which directly approximate the likelihood,[1] are Gaussian synthetic likelihoods (Wood, 2010; Price et al., 2018; Ong et al., 2018), density ratio estimators (Izbicki et al., 2014; Thomas et al., 2016; Dinev and Gutmann, 2018), and more recently neural density estimators (NDE), such as autoregressive models (e.g., Uria et al. 2014, 2016; van den Oord et al. 2016a,c,b) and normalizing flows (e.g., Dinh et al. 2014; Kingma et al. 2016; Papamakarios et al. 2017; see Papamakarios et al. (2019) for a recent review). Other related works estimate the likelihood ratio (Tong, 2013; Cranmer et al., 2015; Stoye et al., 2018; Brehmer et al., 2018).[2]

Typically, machine learning-based LFI models are assessed by computing built-in loss functions (e.g., Kullback-Leibler divergences in emulator networks). Such loss functions however only return a relative measure of performance rather than a goodness-of-fit to simulated data; they do not answer the question "Should we keep searching for better estimates for this problem or is our fit good enough?" (see Figure 3, left, for an example). Thus, an important challenge is that of *validation*: determining whether an approximate likelihood or emulator model reproduces to the extent possible the targeted simulations in distribution. If the model is deemed inadequate, then the question of *diagnostics* becomes relevant. That is, pinpointing "how" and "where" the emulator differs from the simulator in a potentially high-dimensional feature space across different parameters; thereby providing valuable information for further improvements of the emulator, and insights on which simulations to run given a fixed budget. Up to now, popular approaches to simulation-based validation (Cook et al., 2006; Prangle et al., 2014; Talts et al., 2018) are valuable as consistency checks, but cannot always identify likelihood models that are clearly misspecified (see Section 2.3 for an example). Furthermore, as these tools were originally designed for checking Bayesian posterior models, they do not capture all aspects of the estimated likelihood, and therefore provide limited information on how to improve the estimates.

In this paper, we propose general procedures for validating likelihood models. These procedures are inspired by classical hypothesis testing, but generalize to complex data in an LFI setting, and can identify statistically significant deviations from the simulated distribution. We use a new regression-based two-sample test Kim et al. (2016, 2018) to first compare the simulator and emulator models locally, i.e., at fixed parameters; these local tests are then aggregated into a "global" goodness-of-fit test that is statistically consistent (see Theorem 1). Our framework can adopt any machine-learning regression method to handle different structures in high-dimensional data. As Theorem 2 and Figure 2 show, this property translates to *high power* (for a fixed computational budget) under a variety of practical scenarios.

Related Work. Hypothesis testing has recently been used as a goodness-of-fit of generative adversarial networks (GANs, Goodfellow et al. 2014); e.g., Jitkrittum et al. (2016, 2017, 2018) use two-sample tests to detect features and feature space regions which discriminate between real and generated data. Although implicit generative models are not our main focus, our local test has similar diagnostic capabilities (see Suppl. Mat. A and Freeman et al. 2017). There are also close connections between classification accuracy tests (Kim et al., 2016; Lopez-Paz and Oquab, 2017) and our regression test. The main difference lies in the test statistic: classification accuracy tests are based on "global" error rates. Hence classifier tests can tell whether two distributions are different (i.e. they are two-sample tests) but these tests do not *per se* identify locally significant differences between two distributions with statistical confidence; for that one needs to consider the regression or class-conditional probabilities $\mathbb{E}(Y|\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{x})$ (where Y here is the indicator function that $\mathbf{x}$ was generated by the emulator as opposed to the forward-simulator), which is the basis of our regression test statistic (Equation 1).

Novelty. To date, there are no other validation technique in the LFI literature that can answer the following questions in a statistically rigorous way:

(i) **if** one needs to improve emulators for reliable inference from observed data, i.e., whether the difference between the "truth" and the approximation learned with the existing train data is statistically significant; this question is answered by our global procedure (see Figure 3, left and

[1] In this work we will use the terms *emulator* and *approximate likelihood* interchangeably to denote generative models that directly model the relationship between observable data $\mathbf{x}$ and parameters θ.

[2] ABC and methods that target the posterior instead of the likelihood, e.g., some Gaussian process emulators (Rasmussen and Williams, 2005; Wilkinson, 2014; Meeds and Welling, 2014) and approximate posterior models (Papamakarios and Murray, 2016; Gutmann et al., 2016; Lueckmann et al., 2017; Le et al., 2017; Izbicki et al., 2018; Järvenpää et al., 2018; Greenberg et al., 2019), benefit from validation techniques as well but we do not discuss them in this work.

Figure 4, left);

(ii) **where** in parameter space one, if needed, should propose the next batch of simulations; this question is answered by our local procedure (see Figure 3, right) and provides insights as to which simulations to run given a fixed budget; and

(iii) **how** the distributions of emulated and high-resolution simulated data may differ in a potentially high-dimensional feature space; this question is answered by our regression test (see Figure 4, right, and Suppl. Mat. A) and offers valuable information as to what types of observations are under- or over-represented by the emulator and whether such differences are statistically significant. Such insights can guide decisions as to whether it is necessary to improve the emulator model or generate more simulations.

Moreover, we provide theoretical guarantees that ensure that (a) the global test is able to tell if the estimated likelihood is wrong (i.e., no clearly misspecified models can pass the test; Theorem 1), and (b) the local test has high power as long as we have a good estimate of the regression function (Theorem 2).

Organization. The organization of the paper is as follows: In Section 2 we describe our validation method, and provide theoretical guarantees as well as synthetic examples that compare the performance of our goodness-of-fit test over some popular simulation-based calibration and distance-based tests. Then in Section 3, we show how our tools can be used to assess and diagnose models for cosmological parameter inference. Proofs of theorems and details on the high-dimensional sample comparison in feature space are provided in Supplementary Material.

Notation. We indicate with $\mathcal{X}$ the feature space and with Θ the parameter settings where the simulations from the "true" likelihood $\mathcal{L}(\mathbf{x};\theta)$ are available. We denote the approximate likelihood from the emulator model by $\widehat{\mathcal{L}}(\mathbf{x};\theta)$. Both likelihood functions are normalized over $\mathcal{X}$; that is, $\int_{\mathcal{X}} \mathcal{L}(\mathbf{x};\theta)d\mathbf{x} = \int_{\mathcal{X}} \widehat{\mathcal{L}}(\mathbf{x};\theta)d\mathbf{x} = 1$ for every $\theta \in \Theta$.

2 Model Validation by Goodness-of-Fit Test

Our validation approach compares samples from the simulator with samples from the emulator, and can detect local discrepancies for a given parameter setting $\theta_0 \in \Theta$ as well as global discrepancies across parameter settings in Θ. The validation procedure is as follows: For each $\theta_0 \in \Theta$, we first test the null hypothesis $H_0 : \widehat{\mathcal{L}}(\mathbf{x};\theta_0) = \mathcal{L}(\mathbf{x};\theta_0)$ for all $\mathbf{x} \in \mathcal{X}$. This *local test* (Algorithm 1) compares output from the approximate likelihood/emulator model with a "test sample" from the simulator/true likelihood (the latter sample can be a held-out subset of a pre-generated ensemble at θ_0 which has not been used to fit $\widehat{\mathcal{L}}(\mathbf{x};\theta)$). A challenging problem is how to perform a two-sample test that is able to handle different types of data $\mathbf{x}$, and which in addition informs us on how two samples differ in feature space $\mathcal{X}$; in Section 2.2 and Algorithm 3 we propose a new *regression test* that addresses both these questions. After the two-sample comparisons, we combine local assessments into a *global test* (Algorithm 2) for checking if $\widehat{\mathcal{L}}(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta)$ for all $\theta \in \Theta$. The essence of the global test is to pool p-values which, under the null hypothesis, are uniform. Unlike many previous works on pooling p-values for multiple testing (e.g., Lorenz et al. (2016)), the p-values in Algorithm 2 are independent by construction.

The next section provides theoretical guarantees that the global test for our LFI setting is indeed consistent. These results apply for any sampling/weighting scheme $r(\theta)$ over Θ in Algorithm 2, and for any consistent local test in Algorithm 1.

Algorithm 1 Local Test for Fixed θ

```
Input: parameter value $\theta_0$, two-sample testing procedure, number of draws from the true model, $n_{\text{sim},0}$ and from the estimated model, $n_{\text{sim},1}$
Output: $p$-value $p_{\theta_0}$ for testing if $L(\mathbf{x};\theta_0) = \widehat{L}(\mathbf{x};\theta_0)$ for every $\mathbf{x} \in \mathcal{X}$
Sample $\mathcal{S}_0 = \{\mathbf{X}_1^{\theta_0}, \dots, \mathbf{X}_{n_{\text{sim},0}}^{\theta_0}\}$ from $\mathcal{L}(\mathbf{x};\theta_0)$.
Sample $\mathcal{S}_1 = \{\mathbf{X}_1^{*}, \dots, \mathbf{X}_{n_{\text{sim},1}}^{*}\}$ from $\widehat{\mathcal{L}}(\mathbf{x};\theta_0)$.
Compute $p$-value $p_{\theta_0}$ for the comparison between $\mathcal{S}_0$ and $\mathcal{S}_1$.
return $p_{\theta_0}$
```

Algorithm 2 Global Test Across $\theta \in \Theta$

```
Input: reference distribution $r(\theta)$, $B$, uniform testing procedure (e.g. Kolmogorov-Smirnoff, Cramér-von Mises)
Output: $p$-value $p$ for testing if $L(\mathbf{x};\theta) = \widehat{L}(\mathbf{x};\theta)$ for every $\mathbf{x} \in \mathcal{X}$ and $\theta \in \Theta$
for $i \in \{1, \dots, B\}$ do
    sample $\theta_i \sim r(\theta)$
    compute $p_{\theta_i}$ using Algorithm 1
end for
Compute $p$-value $p$ for testing if $\{p_{\theta_i}\}_{i=1}^{B}$ has a uniform distribution.
return $p$
```

2.1 Theoretical Guarantees for Global Test

Theorem 1 shows the global test is statistically consistent; i.e., it is able to detect a misspecified distribution (as in Example 1) for large sample sizes. We provide sufficient assumptions for Theorem 1 in Supp. Mat. B. We assume (i) the set of parameter values where the likelihood function is incorrectly estimated has positive mass under the reference distribution, (ii) statistical consistency of the local comparison test, and (iii) that the test statistic in step 5 of Algorithm 2 converges to zero when the null hypothesis is true, and to a positive number when it is false.

Theorem 1. *Let $\mathbb{D}_{B,n_{sim}} = \{p_{\theta_1}^{n_{sim}}, \dots, p_{\theta_B}^{n_{sim}}\}$, where $p_{\theta_1}^{n_{sim}}, \dots, p_{\theta_B}^{n_{sim}}$ are the p-values obtained by Algorithm 1 with $n_{sim,1} = n_{sim,2} = n_{sim}$ and $\theta_1, \dots, \theta_B \overset{i.i.d.}{\sim} r(\theta)$. Let ϕ be an α-level testing procedure based on the global test statistic S. If the likelihood estimate and the local and global test statistics are such that Assumptions 1–4 hold, then*

$$\mathbb{P}\left(\phi_S(\mathbb{D}_{B,n_{sim}}) = 1\right) \xrightarrow{B,n_{sim} \longrightarrow \infty} 1$$

Corollary 1. *Under Assumptions 1 and 2, the global tests for comparing likelihood models based on Kolmogorov-Smirnoff and Cramér-von Mises statistics are statistically consistent.*

2.2 Two-Sample Test via Regression

Traditional approaches to comparing two distributions (Thas, 2010) are often not easily generalizable to high-dimensional and non-Euclidean data. More recent non-parametric extensions (see Hu and Bai (2016) for a review), e.g., maximum mean discrepancy (MMD, Gretton et al. (2012)), energy distance (ED, Székely and Rizzo (2004)), divergence (Sugiyama et al., 2011; Kanamori et al., 2012), mean embedding (Chwialkowski et al., 2015; Jitkrittum et al., 2016) and classification accuracy tests (Kim et al., 2016; Lopez-Paz and Oquab, 2017) have shown to have power in high dimensions against some alternatives, specifically location and scale alternatives. These methods, however, only provide a binary answer of the form "reject" or "fail to reject" the null hypothesis. Here we propose a new regression-based approach to two-sample testing that can adapt to any structure in $\mathcal{X}$ where there is a suitable regression method; Theorem 2 relates the power of the test to the Mean Integrated Squared Error (MISE) of the regression. Moreover, the regression test can detect and describe local differences (beyond the usual location and scale alternatives) in $\widehat{\mathcal{L}}(\mathbf{x};\theta_0)$ and $\mathcal{L}(\mathbf{x};\theta_0)$ in feature space $\mathcal{X}$. We briefly describe the method below; see Suppl. Mat. E and Kim et al. (2018) for theoretical details, and see Sections 2.3 and 3.2 for examples based on random forest regression.

Let P_0 be the distribution over $\mathcal{X}$ induced by $\mathcal{L}(\mathbf{x};\theta_0)$ and let P_1 be the distribution over $\mathcal{X}$ induced by $\widehat{\mathcal{L}}(\mathbf{x};\theta_0)$. Assume that P_0 and P_1 have density functions f_0 and f_1 relative a common dominating measure. By introducing a random variable $Y \in \{0,1\}$ that indicates which distribution an observation belongs to, we can view f_0 and f_1 as conditional densities $f(\mathbf{x}|Y=0)$ and $f(\mathbf{x}|Y=1)$. The local null hypothesis is then equivalent to the hypothesis $H_0 : f_0(\mathbf{x}) = f_1(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_0 := \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) > 0\}$, which in turn is equivalent to

$$H_0 : \ \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) = \mathbb{P}(Y=1), \quad \text{for all } \mathbf{x} \in \mathcal{X}_0.$$

We test H_0 against the alternative $H_1 : \ \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) \neq \mathbb{P}(Y=1), \quad \text{for some } \mathbf{x} \in \mathcal{X}_0$.

By the above reformulation, we have converted the problem of two-sample testing to a *regression* problem. Depending on the choice of method for estimating the regression function $m(\mathbf{x}) = \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})$, we can adapt to nontraditional data settings involving mixed data types and various structures. More specifically, let $\widehat{m}(\mathbf{x})$ be an estimate of $m(\mathbf{x})$ based on the sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, and let $\widehat{\pi}_1 = \frac{1}{n}\sum_{i=1}^n I(Y_i = 1)$. We define our test statistic as

$$\widehat{\mathcal{T}} = \frac{1}{n}\sum_{i=1}^{n} \left(\widehat{m}(\mathbf{X}_i) - \widehat{\pi}_1\right)^2. \qquad (1)$$

Note that the difference $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ *for each particular value of* $\mathbf{x} \in \mathcal{X}$ also provides information on how well the emulator fits the simulator locally *in feature space*; high values indicate a poor fit. To keep our framework as general as possible, we use a permutation procedure (Algorithm 3) to compute p-values[3]. Theorem 2 shows that if $\widehat{m}$, the chosen regression estimator, has a small MISE, the power of the test is large over a wide region of the alternative hypothesis. What this means in practice is that we should choose a regression method that predicts the "class membership" Y well.

Theorem 2. *Suppose that the regression estimator $\widehat{m}(\mathbf{x})$ is a linear smoother satisfying $\sup_{m \in \mathcal{M}} \mathbb{E}\int_{\mathcal{X}} \left(\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right)^2 dP_X(\mathbf{x}) \leq C_0\delta_n$, where C_0 is a positive constant, $\delta_n = o(1)$, $\delta_n \geq n^{-1}$, and $\mathcal{M}$ is a class of regressions $m(\mathbf{x})$ containing constant functions. Let t_α^* be the upper α quantile of the permutation distribution of the test statistic $\widehat{\mathcal{T}}'$ on validation data*

[3] If the total number of test simulations from $\mathcal{L}(\mathbf{x};\theta_0)$ is small, say $n_{sim} \sim 50$, but the cost of drawing samples from the emulator model $\widehat{\mathcal{L}}(\mathbf{x};\theta_0)$ is negligible, then we can instead of a two-sample permutation test perform a goodness-of-fit test via repeated Monte Carlo sampling from the emulator (see Algorithm 5 in Supp. Mat. D for details). To cite Friedman (Friedman, 2004, Section IV), such an approach has "the potential for increased power at the expense of having to generate many Monte Carlo samples, instead of just one". Corollary 2 states that our main result (Theorem 2) still holds for the alternative goodness-of-fit test.

Algorithm 3 Two-Sample Regression Test via Permutations

```
Input: two i.i.d. samples $\mathcal{S}_0$ and $\mathcal{S}_1$ from distributions with resp. densities $f_0$ and $f_1$; number of permutations $M$; a regression method $\hat{m}$
Output: $p$-value for testing if $f_0(\mathbf{x}) = f_1(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$
1: Define an augmented sample $\{\mathbf{X}_i, Y_i\}_{i=1}^n$, where $\{\mathbf{X}_i\}_{i=1}^n = \mathcal{S}_0 \cup \mathcal{S}_1$, and $Y_i = I(\mathbf{X}_i \in \mathcal{S}_1)$.
2: Calculate the test statistic $\widehat{\mathcal{T}}$ in Equation 1.
3: Randomly permute $\{Y_1, \ldots, Y_n\}$. Refit $\widehat{m}$ and calculate the test statistic on the permuted data.
4: Repeat the previous step $M$ times to obtain $\{\widehat{\mathcal{T}}^{(1)}, \ldots, \widehat{\mathcal{T}}^{(M)}\}$.
5: Approximate the permutation $p$-value by $p = \frac{1}{M+1}\left(1 + \sum_{m=1}^{M} I(\widehat{\mathcal{T}}^{(m)} > \widehat{\mathcal{T}})\right)$.
6: return $p$
```

from sample splitting.[4] *Then for any $\alpha, \beta \in (0, 1/2)$ and n sufficiently large, there exists a universal constant C_1 such that*

$$\textit{Type I error: } \mathbb{P}_0\left(\widehat{\mathcal{T}}' \geq t^*_\alpha\right) \leq \alpha,$$

$$\textit{Type II error: } \sup_{m \in \mathcal{M}(C_1\delta_n)} \mathbb{P}_1\left(\widehat{\mathcal{T}}' < t^*_\alpha\right) \leq \beta$$

against the class of alternatives $\mathcal{M}(C_1\delta_n) := \Big\{m \in \mathcal{M} : \int_{\mathcal{X}} (m(\mathbf{x}) - \pi_1)^2\, dP_X(\mathbf{x}) \geq C_1\delta_n\Big\}$.

2.3 Examples

We next use two synthetic examples to illustrate the advantages of our global and local tests to state-of-the-art validation techniques, in terms of consistency and higher power respectively.

Example 1 (Consistency of Global Test). One key property of our global goodness-of-fit test is that it can detect any misspecified approximation of the likelihood function (Theorem 1). Diagnostic tools like the Posterior Quantiles technique (PQ, Cook et al. (2006)) and Simulation-Based Calibration (SBC, Talts et al. (2018)) are often used to validate approximate likelihood models (see e.g., Papamakarios et al. (2018)) by checking whether a histogram of respective statistics is close to uniform. However, these tests are sometimes not able to discern between the true model and a clearly misspecified model as illustrated by the following toy example, where $\theta_1 \sim \text{Gamma}(1, 1)$, $i = 1, \ldots, 500$, and $X_1, \ldots, X_{1000}|\theta_i \sim \text{Beta}(\theta_i, \theta_i)$. Figure 1 shows the distribution of the PQ and SBC statistics (left and middle columns, respectively) and the distribution of our local p-values (right column) for two different scenarios: In the top row, we consider a case where $\widehat{\mathcal{L}}(\mathbf{x}; \theta) = \mathcal{L}(\mathbf{x}; \theta)$. All tests pass the model, as they should. In the bottom row, we consider a case where $\widehat{\mathcal{L}}(\mathbf{x}; \theta) \propto 1$, a poor approximation of the likelihood function (see Suppl. Mat. A for examples). Our global regression test, which is based on uniformity of the local p-values, clearly rejects this model. PQ and SBC, on the other hand, cannot distinguish between the true likelihood and the misspecified model as these by construction have the same marginal distribution over θ in this toy example. Similar results (Schmidt et al., 2020) have been found for diagnostic tests of conditional density estimates when using quantities related to PQ and SBC (such as, PIT scores and QQ plots).

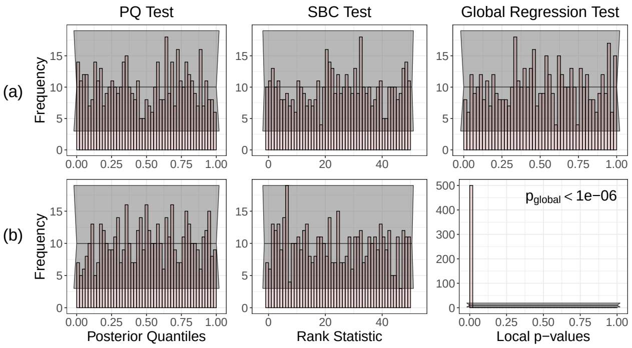

Figure 1: Distribution of posterior quantiles, rank statistics and p-values for PQ, SBC and our global regression test, respectively, for (a) the true model in Example 1, and (b) a clearly misspecified model. Only the global regression test correctly rejects the latter (bottom right plot). (The grey ribbons for the PQ and SBC tests represent the 99% confidence interval for the test of uniformity used in those two tests. Our global test, on the other hand, is based on formally testing whether the local p-values from Algorithm 1 are uniformly distributed.)

Example 2 (Power of Local Test). The power of our goodness-of-fit test will much depend on how we compare samples at fixed $\theta_0 \in \Theta$; that is, on how we test the local null hypothesis, $H_0 : \widehat{\mathcal{L}}(\mathbf{x}; \theta_0) = \mathcal{L}(\mathbf{x}; \theta_0)$ for every $\mathbf{x} \in \mathcal{X}$. An advantage of the regression approach (Algorithm 3) is that we can use any regression technique that efficiently explores the structure of the data at hand; the practical implications of Theorem 2 is that one should choose the regression method with the smallest MISE (a quantity that can be estimated from data) to attain a higher test power (an unknown quantity). We illustrate these ideas with a synthetic example where $\mathbf{x} \in \mathbb{R}^D$, where D could be large. We consider three toy settings where the approximate likelihood and the true likelihood only differ in the first dimension — that is, we test against a sparse alterna-

[4] The proof assumes sample splitting where (for simplicity) half of the data is used to estimate the regression function and the other half is used to estimate the test statistic; i.e., $\widehat{\mathcal{T}}' = 2n^{-1}\sum_{i=n/2}^{n}(\widehat{m}(\mathbf{X}_i) - \widehat{\pi}_1)^2$, where $\widehat{m}$ is estimated using $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{n/2}, Y_{n/2})$.

Example 2 Settings	*True Likelihood* $\mathcal{L}(\mathbf{x};\theta)$	*Approx. Likelihood* $\widehat{\mathcal{L}}(\mathbf{x};\theta)$	*Param. Space* Θ
(a) Bernoulli	$\text{Bern}(x_1;\theta)\prod_{d=2}^{D}\mathcal{N}(x_d;\theta,1)$	$\prod_{d=1}^{D}\mathcal{N}(x_d;\theta,1)$	$(0,1)$
(b) Scaling	$\mathcal{N}(x_1;0,\theta)\prod_{d=2}^{D}\mathcal{N}(x_d;0,1)$	$\prod_{d=1}^{D}\mathcal{N}(x_d;0,1)$	$(0,1)$
(c) Mixture of Gaussians	$f_m(x_1;\theta,1)\prod_{d=2}^{D}\mathcal{N}(x_d;0,1)$, where $f_m(\theta,1)=1/2\mathcal{N}(-\theta,1)+1/2\mathcal{N}(\theta,1)$	$\prod_{d=1}^{D}\mathcal{N}(x_d;0,1)$	$(-5,5)$

Table 1: The three toy settings in Example 2. In each setting, the true and approximate likelihood differ only in the first dimension, x_1. ($\mathcal{N}(x;\mu,\sigma^2)$ is a 1D Gaussian with mean μ and variance σ^2; Bern$(x;\theta)$ is a Bernoulli with parameter θ.)

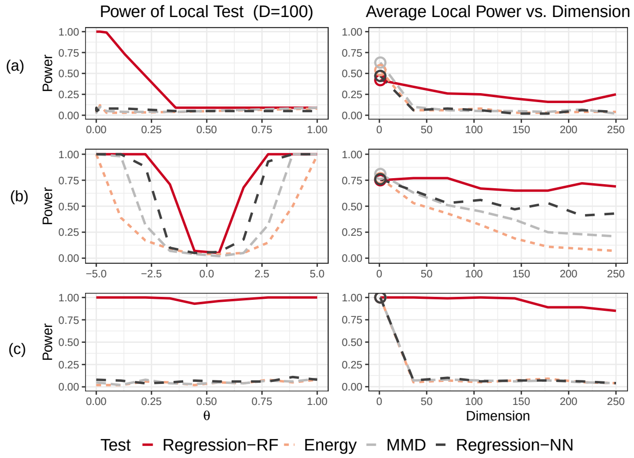

Figure 2: Local test power shown in the left column as a function of θ at D=100, and shown in the right column as a function of the dimension D (averaged over θ) for the (a) Bernoulli, (b) Scaling, and (c) Mixture of Gaussians case. Note that distance-based tests are more powerful at $D=1$ (highlighted with circles in the right column), but their power is severely affected with increasing dimension. Our RF regression test achieves higher power for large D by leveraging the advantages of random forest regression in high-dimensional settings with sparse structure.

tive; see Table 1 for details.

For each $\theta\in\Theta$, we compute a local p-value by comparing samples of size $n=100$ from $\mathcal{L}(\mathbf{x};\theta)$ and $\widehat{\mathcal{L}}(\mathbf{x};\theta)$, respectively (Algorithm 1). This procedure is repeated 100 times to estimate the power function. We apply the local test for three different test statistics; namely: (i) the test statistic in Equation 1 using random forest (RF) or nearest neighbor regression (NN), (ii) the MMD test statistic (Gretton et al., 2012, Eq. 5) with a Gaussian kernel, and (iii) the energy test statistic (Székely and Rizzo, 2004, Eq. 5) using the Euclidean norm. Figure 2 shows how the power function varies with θ at dimension $D=100$ (left column) and how the power, averaged over θ, varies with D (right column) for each setting. When $D=1$ (highlighted with circles in the right column) distance-based tests based on RF yield higher power, but their performance quickly degrades with increasing D. On the other hand, our RF regression test is able to achieve higher power in high-dimensional settings by leveraging random forest regression ability to select features and to tell discrete versus continuous distributions apart (as shown by the red curves). For instance, in the Bernoulli case (top row, a) our regression test has higher power for small values of θ, which is when the distribution of the first coordinate is almost degenerate at 0.

3 Applications

In this section we focus on *validating* approximate likelihood models for cosmological parameter inference with weak lensing peak counts. Weak lensing (WL) is a gravitational deflection effect of light along the line of sight. We can use this effect to estimate parameters of the ΛCDM cosmological model, the most well-supported model within Big Bang cosmology. In particular we can estimate the dark matter density Ω_m and its clumpiness σ_8 through *peak counts*: the number of local maxima in the WL convergence map (a 2D image) binned by the value of the peak Dietrich and Hartlap (2010).

In Section 3.1, we showcase our approach on a synthetic example with known likelihood and properties similar to those of peak counts. To estimate the likelihood we use two parametric models, a Gaussian and a Poisson model, as well as a non-parametric kernel density estimator (KDE) with the bandwidth estimated coordinate-wise according to (Wand and Jones, 1994) and discretized to reflect the integer-valued data. (The Gaussian model with a fixed covariance and varying mean is the current state-of-the-art in cosmological parameter inference Kacprzak et al. (2016).) In Section 3.2, we provide results and insights with peak counts data obtained from the `CAMELUS` simulator (Lin and Kilbinger, 2015), comparing the two parametric models with a conditional masked autoregressive flow (MAF, Papamakarios et al. (2017)), again discretized to reflect the integer-valued data.

3.1 Synthetic Example

Peak count data possess two important properties: (1) data are discrete and (2) counts in different bins are correlated to each other. The first property implies that

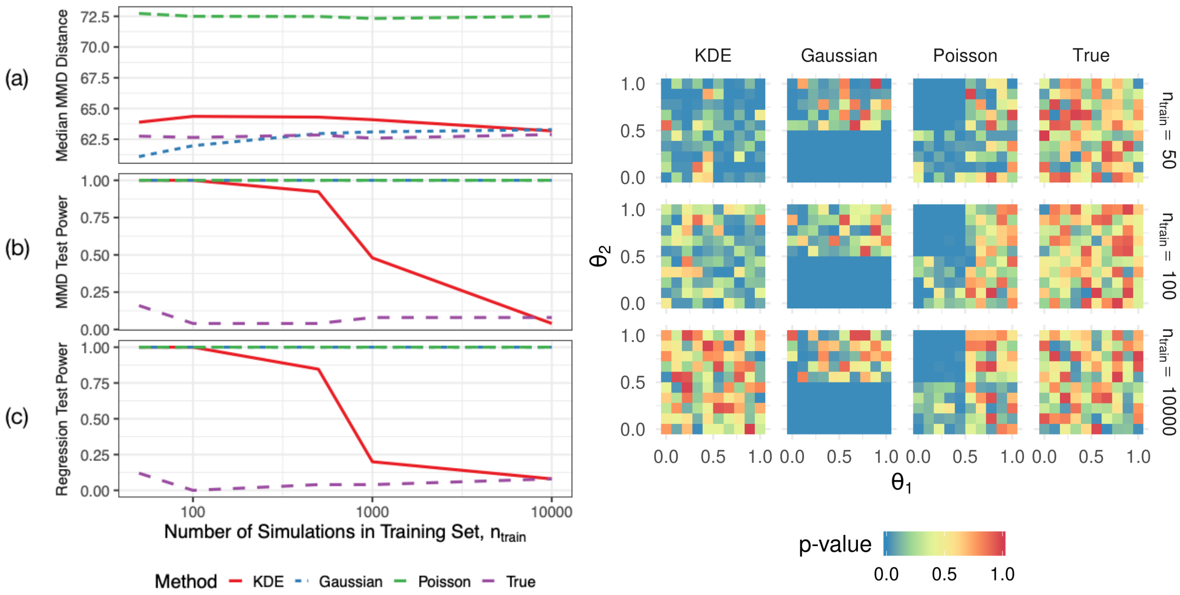

Figure 3: *Left panel*: Median MMD distance and power of global goodness-of-fit tests (100 trials with $\alpha = 0.05$ and $n_{\text{sim}} = 200$) for the synthetic example. (a) The median MMD over all trials is not informative as it does not vary with n_{train}. However, global tests based on (b) the MMD distance and (c) the regression are able to capture that KDE improves as n_{train} increases (the power decreases with n_{train}), while the parametric models do not. *Right panel*: Local test p-values for regression test by model and number of training simulations. We can identify regions where models fit poorly: e.g., the Gaussian model fits poorly for bottom half of θ-space as low counts cannot be adequately approximated as Gaussian.

at high bin counts the data are approximately normally distributed, but for bins with low counts this approximation breaks down. The latter property introduces difficulties in modeling number counts as independent Poisson variables. We mimic these two properties by drawing $X_1, X_2 \overset{\text{indep}}{\sim} \text{Poisson}(\lambda)$, where λ depends on the parameter $\theta = [\theta_1, \theta_2] \in \mathbb{R}^2$. When $\theta_1 < 0.5$, we set $\lambda = 1$, otherwise $\lambda = 10^4$ which makes the normal approximation appropriate due to the Central Limit Theorem. When $\theta_2 < 0.5$ we add the requirement that $X_1 \leq X_2$ (which breaks independence). Our first experiment is to use our global test to assess likelihood models that are fit with different number of simulations ($n_{train} = [50, 100, 10000]$) while holding the size of the test samples fixed ($n_{\text{sim}} = 200$). For the likelihood models mentioned above – KDE, Gaussian and Poisson – we implement Algorithm 2 with a uniform reference distribution over a grid of 100 θ-values evenly spaced in $[0, 1] \times [0, 1]$. For illustrative purposes, we conduct 100 trials resampling the entire dataset to estimate the power of the test. The fit of the likelihood models are assessed using three criteria: (i) the median (over 100 trials) MMD distance between the two samples, (ii) the power of a global test based on the same MMD distances, and (iii) the power of a global regression test with random forest. It is common practice to compare emulator models (see, e.g., Papamakarios et al. (2018); Greenberg et al. (2019)) by computing distances such as MMD, which we here refer to as raw test statistics. Figure 3, left, shows that the "Median MMD Distance" (top, a) is not particularly informative in this example. On the other hand, the "MMD Test Power" (center, b) and the "Regression Test Power" (bottom, c) tell us that both the Poisson and Gaussian models are misspecified; these models are rejected regardless of n_{train}, whereas the KDE model slowly improves with the number of simulations until ultimately achieving a power similar to the true distribution. These results illustrate that the local and global p-values can be more informative than the test statistics themselves.

To better understand why the Gaussian and Poisson models fit poorly we can turn to the local information we calculated for each θ via Algorithm 1 ($n_{sim} = 200$). The data-generating process in our synthetic example induces four quadrants with different behaviors. Figure 3, right, showcases the utility of the local test: it pinpoints *where* in the parameter space the model fits are insufficient. More specifically: for the Poisson fits, the p-values in the left ($\theta_1 < 0.5$) region are very small as are the p-values for the lower ($\theta_2 < 0.5$) region for the Gaussian fits. This is due to the independence and Gaussian assumptions, respectively, breaking down in these two regions. In addition, the KDE model improves as the number of simulations used to train the models increases, starting at poor fits with low p-values at $n_{\text{train}} = 50$ and eventually achieving p-values drawn from the uniform distribution for large values of n_{train}. Our global test makes this observation rigorous.

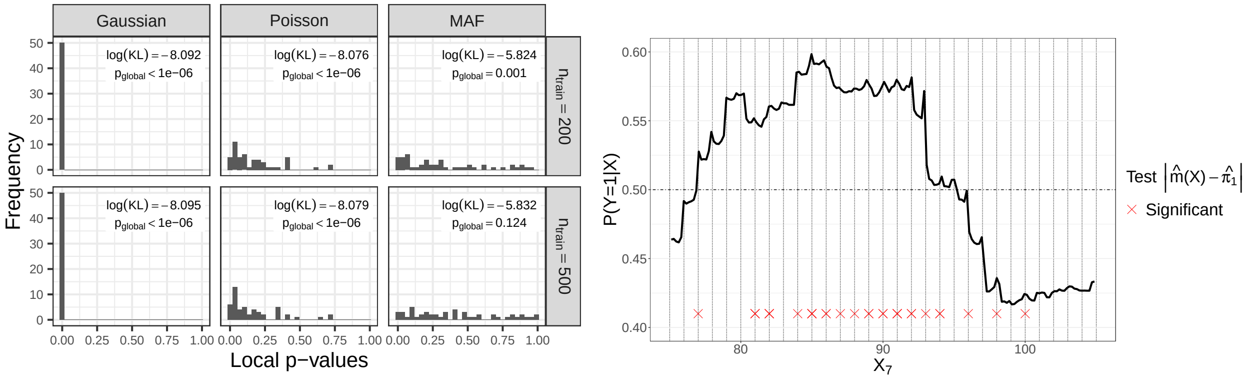

Figure 4: *Left panel*: Local goodness-of-fit for peak-count data with $n_{\text{train}} = 200$ (top row) and $n_{\text{train}} = 500$ (bottom row). Although the Gaussian model is achieving the lowest KL divergence, the estimates are rejected at almost all θ; Poisson and MAF perform better (more uniform-looking distributions of local p-values) but only MAF passes the global test at $n_{\text{train}} = 500$. *Right panel*: Partial dependence plot for variable x_7 (low count bin) for Gaussian model at $n_{\text{train}} = 200$. The red crosses on the x-axis represent the locations where the difference $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ is statistically significant according to a joint analysis in 7 dimensions; these locations coincide with integer values of x_7 and indicate that the regression test is distinguishing between the discrete true distribution of counts and the fitted continuous Gaussian distribution.

3.2 Peak Count Data Example

For WL peaks, we consider a 2D parameter space over $\theta = (\Omega_m, \sigma_8)$ and design a grid of 50 different cosmologies θ around a fiducial (probable) cosmology θ_0 (see Suppl. Mat. D). For each θ-value, we simulate a batch of WL maps ($n_{\text{train}} = 200$, $n_{\text{sim}} = 200$). The peak count data (i.e. histogram of peak intensities in each map) is a vector $\mathbf{x} \in \mathbb{N}^D$ where $D = 7$ is the number of bins. We compare three approximate likelihood models: Gaussian, Poisson and conditional MAF. To assess models, we first compute the Kullback-Leibler (KL) divergence loss for the $n_{sim} = 200$ test simulations at each θ. According to the KL loss, the Gaussian model performs best; however, these are only relative comparisons. We now use a RF regression test to find out whether the Gaussian model actually fits the simulated data well. As indicated in Figure 4 (left, top row), the local tests for the Gaussian model reject the null hypothesis $\widehat{\mathcal{L}}(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta)$ at every θ; thus the global hypothesis is also rejected. The Poisson and MAF models are rejected by the global test as well but have a more uniform-looking distribution of local p-values. Now if we increase the the number of train simulations to $n_{\text{train}} = 500$ (while holding $n_{\text{sim}} = 200$ fixed), the fitted MAF model passes the global test whereas the Gaussian and Poisson models still do not as indicated by the bottom row (these qualitative results stay the same for $n_{\text{train}} = 5000$).

Finally, our local regression tests can provide insights into *how* the two distributions $\widehat{\mathcal{L}}(\mathbf{x};\theta)$ and $\mathcal{L}(\mathbf{x};\theta)$ differ in feature space $\mathcal{X}$; more specifically, by evaluating how the estimate of the regression function $\widehat{m}(\mathbf{x})$ in Equation 1 varies with $\mathbf{x}$ for a fixed θ (a significant difference $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ is an indication that the model is not well estimated at that location in feature space). We illustrate such an analysis for our fitted Gaussian model for $n_{\text{train}} = 200$ and $\theta = \theta_0$. According to the RF regression used to construct our test statistic, the most influential variables correspond to bins with low counts. In Figure 4, right, we visualize the fit on such a bin (variable x_7) by a partial dependence plot (which shows the marginal effect of this variable on $\widehat{m}(\mathbf{x})$ (Friedman, 2001)). On the x-axis, we mark the locations where the difference $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ is statistically significant according to a joint analysis in 7 dimensions (see Algorithm 4 in Suppl. Mat. A for details). These locations coincide with integer values of x_7, showing that the regression test is distinguishing between the discrete true distribution for bin counts and the fitted continuous Gaussian distribution (these results also explain why the Poisson model may fare better). In Suppl. Mat. A we provide a detailed analysis of how one can identify and visualize areas of significant differences in multivariate distributions for galaxy morphology images.

4 Final Remarks

We have developed validation methods of approximate emulator models that are able to identify a misspecified model and give insights on how to improve such a model; more specifically, inform the user as to what regions of the parameter space new simulations (if needed) should be added as well as how emulated and simulated data may differ in a high-dimensional feature space. Future work involves using these results to design more efficient strategies for guided simulations that can balance statistical performance with computational costs.

Acknowledgements

The authors are grateful to the anonymous reviewers for their insightful comments that helped improve the manuscript. ABL, ND and TP were supported in part by the National Science Foundation under DMS-1520786. RI was supported in part by FAPESP (2017/03363-8 and 2019/11321-9) and CNPq (306943/2017-4).

References

T. Abbott, F. B. Abdalla, S. Allam, A. Amara, Annis, et al. Cosmology from cosmic shear with Dark Energy Survey Science Verification data. *Physical Review D*, 94(2):022001, July 2016.

AH Baker, DM Hammerling, MN Levy, H Xu, JM Dennis, BE Eaton, J Edwards, C Hannay, SA Mickelson, RB Neale, et al. A new ensemble-based consistency test for the community earth system model. *Geoscientific Model Development Discussions*, 8(9), 2015.

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

Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Mining gold from implicit models to improve likelihood-free inference. *CoRR*, abs/1805.12244, 2018.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Kacper Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 1981–1989, Cambridge, MA, USA, 2015. MIT Press. URL `http://dl.acm.org/citation.cfm?id=2969442.2969461`.

R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0500334102.

Samantha R Cook, Andrew Gelman, and Donald 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.

Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

J. P. Dietrich and J. Hartlap. Cosmology with the shear-peak statistics. *Monthly Notices of the Royal Astronomical Society*, 402(2):1049–1058, 2010.

Traiko Dinev and Michael U. Gutmann. Dynamic Likelihood-free Inference via Ratio Estimation (DIRE). *arXiv e-prints*, art. arXiv:1810.09899, Oct 2018.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *CoRR*, abs/1410.8516, 2014.

P. E. Freeman, I. Kim, and A. B. Lee. Local two-sample testing: a new tool for analysing high-dimensional astronomical data. *Monthly Notices of the Royal Astronomical Society*, 471(3):3273–3282, 07 2017. ISSN 0035-8711. doi: 10.1093/mnras/stx1807.

Jerome Friedman. On multivariate goodness-of-fit and two-sample testing. Technical report, Stanford Linear Accelerator Center, Menlo Park, CA (US), 2004.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

David S. Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic posterior transformation for likelihood-free inference. *ICML*, 2019.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, March 2012. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2188385.2188410`.

N. A. Grogin, D. D. Kocevski, S. M. Faber, H. C. Ferguson, A. M. Koekemoer, A. G. Riess, V. Acquaviva, D. M. Alexander, O. Almaini, M. L. N. Ashby, M. Barden, E. F. Bell, F. Bournaud, T. M. Brown, K. I. Caputi, S. Casertano, P. Cassata, M. Castellano, P. Challis, R.-R. Chary, E. Cheung, M. Cirasuolo, C. J. Conselice, A. Roshan Cooray, D. J. Croton, E. Daddi, T. Dahlen, R. Davé, D. F. de Mello, A. Dekel, M. Dickinson, T. Dolch, J. L. Donley, J. S. Dunlop, A. A. Dutton, D. Elbaz, G. G. Fazio, A. V. Filippenko, S. L. Finkelstein, A. Fontana, J. P. Gardner, P. M. Garnavich, E. Gawiser, M. Giavalisco, A. Grazian, Y. Guo, N. P. Hathi, B. Häussler, P. F. Hopkins, J.-S. Huang, K.-H. Huang, S. W. Jha, J. S. Kartaltepe, R. P. Kirshner, D. C. Koo,

K. Lai, K.-S. Lee, W. Li, J. M. Lotz, R. A. Lucas, P. Madau, P. J. McCarthy, E. J. McGrath, D. H. McIntosh, R. J. McLure, B. Mobasher, L. A. Moustakas, M. Mozena, K. Nandra, J. A. Newman, S.-M. Niemi, K. G. Noeske, C. J. Papovich, L. Pentericci, A. Pope, J. R. Primack, A. Rajan, S. Ravindranath, N. A. Reddy, A. Renzini, H.-W. Rix, A. R. Robaina, S. A. Rodney, D. J. Rosario, P. Rosati, S. Salimbeni, C. Scarlata, B. Siana, L. Simard, J. Smidt, R. S. Somerville, H. Spinrad, A. N. Straughn, L.-G. Strolger, O. Telford, H. I. Teplitz, J. R. Trump, A. van der Wel, C. Villforth, R. H. Wechsler, B. J. Weiner, T. Wiklind, V. Wild, G. Wilson, S. Wuyts, H.-J. Yan, and M. S. Yun. Candels: The cosmic assembly near-infrared deep extragalactic legacy survey. *The Astrophysical Journal, Supplement*, 197:35, December 2011. doi: 10.1088/0067-0049/197/2/35.

Arushi Gupta, José Manuel Zorrilla Matilla, Daniel Hsu, and Zoltán Haiman. Non-Gaussian information from weak lensing data via deep learning. *Physical Review D*, 97(10):103515, 2018.

Michael U. Gutmann, Jukka Cor, and er. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125):1–47, 2016. URL `http://jmlr.org/papers/v17/15-017.html`.

H. Hildebrandt, M. Viola, C. Heymans, S. Joudaki, K. Kuijken, et al. KiDS-450: cosmological parameter constraints from tomographic weak gravitational lensing. *Monthly Notices of the Royal Astronomical Society*, 465:1454–1498, February 2017.

J. Hu and Z. Bai. A review of 20 years of naive tests of significance for high-dimensional mean vectors and covariance matrices. *Science China Mathematics*, 59 (12):2281–2300, 2016.

James W Hurrell, Marika M Holland, Peter R Gent, and Steven Ghan. The community earth system model: a framework for collaborative research. *Bulletin of the American Meteorological Society*, 94(9):1339–1360, 2013.

R. Izbicki, A. Lee, and C. Schafer. High-dimensional density ratio estimation with extensions to approximate likelihood computation. In *Artificial Intelligence and Statistics*, pages 420–429, 2014.

Rafael Izbicki, Ann B. Lee, and Taylor Pospisil. ABC–CDE: Toward approximate bayesian computation with complex high-dimensional data and limited simulations. *Journal of Computational and Graphical Statistics*, pages 1–20, nov 2018. doi: 10.1080/10618600.2018.1546594.

Marko Järvenpää, Michael U. Gutmann, Aki Vehtari, and Pekka Marttinen. Gaussian process modelling in approximate bayesian computation to estimate horizontal gene transfer in bacteria. *The Annals of Applied Statistics*, 12(4):2228–2251, 12 2018.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 181–189. Curran Associates, Inc., 2016.

Wittawat Jitkrittum, Wenkai Xu, Zoltan Szabo, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. 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 262–271. Curran Associates, Inc., 2017.

Wittawat Jitkrittum, Heishiro Kanagawa, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, and Arthur Gretton. Informative features for model comparison. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 808–819. Curran Associates, Inc., 2018.

Tomasz Kacprzak, D Kirk, O Friedrich, A Amara, A Refregier, et al. Cosmology constraints from shear peak statistics in Dark Energy Survey Science verification data. *Monthly Notices of the Royal Astronomical Society*, 463(4):3653–3673, 2016.

T. Kanamori, T. Suzuki, and M. Sugiyama. f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2): 708–720, Feb 2012. ISSN 0018-9448.

Jennifer Kay, C. Deser, A Phillips, A Mai, Cecile Hannay, G. Strand, J. Arblaster, Susan Bates, G. Danabasoglu, James Edwards, M. Holland, Paul Kushner, Jean-Franois Lamarque, D. Lawrence, Keith Lindsay, A Middleton, Ernesto Munoz, R. Neale, Keith Oleson, and Mariana Vertenstein. The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.*, 08 2015. doi: 10.1175/BAMS-D-13-00255.1.

Ilmun Kim, Ann B. Lee, Peter Freeman, and Jeffrey Newman. Comparing distributions of galaxy morphologies. Technical report, Carnegie Mellon University, Department of Statistics & Data Science, 2016.

Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two sample testing. *arXiv e-prints*, art. arXiv:1602.02210, Feb 2016.

Ilmun Kim, Ann B. Lee, and Jing Lei. Global and Local Two-Sample Tests via Regression. *arXiv e-prints*, art. arXiv:1812.08927, Dec 2018.

Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4743–4751, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9. URL `http://dl.acm.org/citation.cfm?id=3157382.3157627`.

Anton Koekemoer, Sabien Faber, Henry Ferguson, Norman A. Grogin, Dale D. Kocevski, David C. Koo, Kamson Lai, Jennifer M. Lotz, Ray Lucas, Elizabeth J. McGrath, Sara Ogaz, Abhijith Rajan, Adam G. Riess, Steve A. Rodney, Louis Strolger, Stefano Casertano, Marco Castellano, Tomas Dahlen, Mark Dickinson, and and Min S. Yun. Candels: The cosmic assembly near-infrared deep extragalactic legacy survey. *The Astrophysical Journal Supplement Series*, 197:36, 12 2011. doi: 10.1088/0067-0049/197/2/36.

Tuan Anh Le, Atilim Gunes Baydin, and Frank Wood. Inference Compilation and Universal Probabilistic Programming. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1338–1348, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL `http://proceedings.mlr.press/v54/le17a.html`.

Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Chieh-An Lin and Martin Kilbinger. A new model to predict weak-lensing peak counts-I. comparison with N-body simulations. *Astronomy & Astrophysics*, 576: A24, 2015.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL `https://openreview.net/forum?id=SJkXfE5xx`.

Ruth Lorenz, AJ Pitman, and Scott A Sisson. Does amazonian deforestation cause global effects; can we be sure? *Journal of Geophysical Research: Atmospheres*, 121(10):5567–5584, 2016.

Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob 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.

Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6): 1167–1180, 2012.

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

Shakir Mohamed and Balaji Lakshminarayanan. Learning in Implicit Generative Models. *arXiv e-prints*, art. arXiv:1610.03483, Oct 2016.

Mustafa Mustafa, Deborah Bard, Wahid Bhimji, Zarija Luki, Rami Al-Rfou, and Jan Kratochvil. Cosmogan: creating high-fidelity weak lensing convergence maps using generative adversarial networks. *Computational Astrophysics and Cosmology*, 6, 12 2019. doi: 10.1186/s40668-019-0029-9.

Victor M. H. Ong, David J. Nott, Minh-Ngoc Tran, Scott A. Sisson, and Christopher C. Drovandi. Variational bayes with synthetic likelihood. *Statistics and Computing*, 28:971–988, 2018.

George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, pages 1028–1036, 2016.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 2335–2344, USA, 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4. URL `http://dl.acm.org/citation.cfm?id=3294771.3294994`.

George Papamakarios, David C Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. *arXiv preprint arXiv:1805.07226*, 2018.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *arXiv e-prints*, art. arXiv:1912.02762, December 2019.

Dennis Prangle, Michael GB Blum, G Popovic, and SA Sisson. Diagnostic tools for approximate Bayesian computation using the coverage property. *Australian & New Zealand Journal of Statistics*, 56(4):309–329, 2014.

Leah F Price, Christopher C Drovandi, Anthony Lee, and David J Nott. Bayesian synthetic likelihood.

Journal of Computational and Graphical Statistics, 27(1):1–11, 2018.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

Siamak Ravanbakhsh, Francois Lanusse, Rachel Mandelbaum, Jeff Schneider, and Barnabás Póczos. Enabling dark energy science with deep generative models of galaxy images. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 1488–1494. AAAI Press, 2017. URL http://dl.acm.org/citation.cfm?id=3298239.3298456.

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 The LSST Dark Energy Science Collaboration. Evaluation of probabilistic photometric redshift estimation approaches for LSST. *arXiv e-prints*, art. arXiv:2001.03621, Jan 2020.

Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC, 2018.

Markus Stoye, Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Likelihood-free inference with an improved cross-entropy estimator. *CoRR*, abs/1808.00973, 2018.

Masashi Sugiyama, Taiji Suzuki, Yuta Itoh, Takafumi Kanamori, and Manabu Kimura. Least-squares two-sample test. *Neural networks : the official journal of the International Neural Network Society*, 24:735–51, 04 2011. doi: 10.1016/j.neunet.2011.04.003.

Gábor J. Szákely and Maria L. Rizzo. Testing for equal distributions in high dimensions. *InterStat*, 2004.

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.

Olivier Thas. *Comparing distributions*. Springer, 2010.

Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U Gutmann. Likelihood-free inference by ratio estimation. *arXiv preprint arXiv:1611.10242*, 2016.

Xin Tong. A plug-in approach to neyman-pearson classification. *Journal of Machine Learning Research*, 14: 3011–3040, 2013. URL http://jmlr.org/papers/v14/tong13a.html.

Benigno Uria, Iain Murray, and Hugo Larochelle. A deep and tractable density estimator. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 467–475, Bejing, China, 22–24 Jun 2014. PMLR.

Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. *Journal of Machine Learning Research*, 17(205):1–37, 2016. URL http://jmlr.org/papers/v17/16-272.html.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv e-prints*, art. arXiv:1609.03499, Sep 2016a.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel Recurrent Neural Networks. *arXiv e-prints*, art. arXiv:1601.06759, Jan 2016b.

Aäron van den van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4797–4805, USA, 2016c. Curran Associates Inc. ISBN 978-1-5108-3881-9. URL http://dl.acm.org/citation.cfm?id=3157382.3157633.

M. P. Wand and Chris Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9 (2):97–116, 1994. URL http://oro.open.ac.uk/28244/.

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

Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102, 2010.