Validation of Approximate Likelihood and Emulator Models for Computationally Intensive Simulations

Niccolò Dalmasso  
Department of Statistics & Data Science  
Carnegie Mellon University  
ndalmass@stat.cmu.edu

Ann B. Lee  
Department of Statistics & Data Science  
Carnegie Mellon University  
anlee@stat.cmu.edu

Rafael Izbicki  
Department of Statistics  
Federal University of São Carlos  
rafaelizbicki@gmail.com

Taylor Pospisil  
Department of Statistics & Data Science  
Carnegie Mellon University  
popt23@gmail.com

Chieh-An Lin  
Institute for Astronomy  
University of Edinburgh  
calin@roe.ac.uk

Abstract

Complex phenomena are often modeled with computationally intensive feedforward simulations for which a tractable analytic likelihood does not exist. In these cases, it is sometimes necessary to use an approximate likelihood or faster emulator model for efficient statistical inference. We describe a new two-sample testing framework for quantifying the quality of the fit to simulations at fixed parameter values. This framework can leverage any regression method to handle complex high-dimensional data and attain higher power in settings where well-known distance-based tests would not. We also introduce a statistically rigorous test for assessing global goodness-of-fit across simulation parameters. In cases where the fit is inadequate, our method provides valuable diagnostics by allowing one to identify regions in both feature and parameter space which the model fails to reproduce well. We provide both theoretical results and examples which illustrate the effectiveness of our approach.

1 Introduction

The likelihood function \( L(x; \theta) \) links the unknown components \( \theta \) of the data-generating mechanism with the observable data \( x \) 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, and which can encode, e.g., complex observational effects, selection biases, etc. In situations where there is no tractable expression for the likelihood function, but a stochastic numerical simulator is available, approximate inference of parameters of interest is possible. This is referred to as likelihood-free inference (LFI), among which Approximate Bayesian Computation (ABC) \cite{1,2} is one of the best known methods.

However, in some disciplines such as cosmology and climate science, accurate analyses require highly realistic simulations. Such simulations are often computationally intensive and unrealistic to generate...
“on the fly” as required by ABC. 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. For example, cosmological N-body simulations, which compute gravity between particle pairs, are usually created at a fixed cosmology \cite{4,5} or on a grid of carefully chosen parameter sets \cite{6,7}. Such simulations can take months to run on thousands of CPUs \cite{8,9}. Similarly, modern climate and weather forecasting models, e.g., CESM \cite{10}, incorporate complex representations of the atmosphere, ocean, land, ice, etc., on fine spatial and temporal resolutions. These dynamical simulations are also commonly run as an ensemble with different initial conditions (see \cite{11} and references within).

Given the above scenario, a solution to make LFI feasible is to replace a computationally expensive simulator with a faster approximate emulator model that can speed up probabilistic modeling by several orders of magnitude. Some common models are Gaussian synthetic likelihoods \cite{12,13,14} and Gaussian process emulators \cite{15,16,17}. There have also been recent algorithmic advances on “approximate likelihood” models that directly estimate the likelihood \( L(x; \theta) \) nonparametrically based on simulator outputs; see, e.g., \cite{18} and “emulator networks” \cite{19,20,21,22}. (There are also approaches that directly target the posterior instead, either via density estimation methods \cite{23,24,25,26} or through likelihood ratio estimation \cite{27,28,29}.) ABC and approximate posterior methods benefit from emulators as well but we do not discuss them in this work.)

The development of various LFI methods raises the challenge of validation: determining whether an approximate likelihood or emulator model reproduces to the extent possible the targeted simulations. If the model is inadequate, then the question of diagnostics becomes relevant; such as, pinpointing how and where the emulator differs from the simulator in a potentially high-dimensional feature space across different parameters. Up to now, popular approaches to simulation-based validation \cite{30,31,32} are valuable as consistency checks, but cannot always identify likelihood models that are clearly misspecified (see Section 2 for an example). Also, these tools only provide limited information when an emulator model is underperforming. Similarly, loss functions built into many machine learning algorithms (e.g., Kullback-Leibler divergences for emulator networks) have shortcomings, as they only return a relative measure of performance rather than a goodness of fit to simulated data.

In this paper, we propose general procedures for validating likelihood models. These procedures are inspired by classical hypothesis testing, but generalize to complex high-dimensional data with an LFI setting, and can identify any statistically significant deviation from the simulated distribution. We use a new regression-based two-sample test \cite{33} to first compare the simulator and emulator models locally, i.e., at fixed parameters; these local tests are then turned into a “global” goodness-of-fit test that is statistically consistent (see Theorem 2). Our framework can adopt any machine-learning regression method to handle different structures in high-dimensional data. As Theorem 1 and Figure 1 (right) show, this property translates to high power (for a fixed computational budget) under a variety of practical scenarios.

The validation methods we present go beyond binary reject/accept decisions and allow one to answer the following questions:

(i) if one needs more simulations to improve emulators for reliable inference from observed data (this question is answered by our consistent global procedure; see Figure 2, left)

(ii) where in parameter space one, if needed, should propose the next batch of simulations (answered by our local procedure; see Figure 2, right), and

(iii) how emulated and high-resolution simulated data may be different in feature space (answered by our regression test; see Figure 3, right, and Supplementary Material C), providing valuable information for further improvements.

To date, there are no diagnostics or validation techniques in the emulator and LFI literature that are fully consistent (i.e., that can distinguish any misspecified estimator from the true likelihood), and that in addition can answer the above if/where/how questions.

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 tests over some existing methods. Then in Section 5 we show how our tools can be used to assess and diagnose models for cosmological parameter inference. Proofs of theorems, additional examples and details on how to estimate p-values and confidence regions for \( \theta \) are provided in Supplementary Material.
Throughout the paper, we indicate with $X$ the feature space and with $\Theta$ the parameter settings where the computationally expensive ensemble simulations from the “true” model $\mathcal{L}(x; \theta)$ are available. We denote the approximate likelihood from the emulator model by $\hat{\mathcal{L}}(x; \theta)$. Both likelihood functions are normalized over $X$; that is, $\int_X \mathcal{L}(x; \theta) dx = \int_X \hat{\mathcal{L}}(x; \theta) dx = 1$ 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 $\Theta$. The validation procedure is as follows: For each $\theta_0 \in \Theta$, we first test the null hypothesis $H_0 : \hat{\mathcal{L}}(x; \theta_0) = \mathcal{L}(x; \theta_0)$ for all $x \in X$. 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 $\theta_0$ which has not been used to fit $\hat{\mathcal{L}}(x; \theta)$). A challenging problem is how to perform a two-sample test that is able to handle different types of data $x$, and which in addition informs us on how two samples differ in feature space $X$; in Section 2.1 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 $\hat{\mathcal{L}}(x; \theta) = \mathcal{L}(x; \theta)$ for all $\theta \in \Theta$. Section 2.1.1 provides theoretical guarantees that the global test is indeed consistent for any sampling/weighting scheme $r(\theta)$ over $\Theta$ in Algorithm 2, and any consistent local test in Algorithm 1.

### Algorithm 1 Local Test for Fixed $\theta$

**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 $L(x; \theta_0) = \hat{L}(x; \theta_0)$ for every $x \in X$

1: Sample $S_0 = \{X_{1}^{\theta_0}, \ldots, X_{n_{\text{sim},0}}^{\theta_0}\}$ from $\mathcal{L}(x; \theta_0)$.
2: Sample $S_1 = \{X_{1}^{\hat{\theta}_0}, \ldots, X_{n_{\text{sim},1}}^{\hat{\theta}_0}\}$ from $\hat{\mathcal{L}}(x; \hat{\theta}_0)$.
3: Compute p-value $p_{\theta_0}$ for the comparison between $S_0$ and $S_1$.
4: return $p_{\theta_0}$

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

**Input**: reference distribution $r(\theta)$, $B$, uniform testing procedure (e.g. Kolmogorov-Smirnov, Cramér-von Mises)

**Output**: p-value $p$ for testing if $L(x; \theta) = \hat{L}(x; \theta)$ for every $x \in X$ and $\theta \in \Theta$

1: for $i \in \{1, \ldots, B\}$ do
2: sample $\theta_i \sim r(\theta)$
3: compute $p_{\theta_i}$ using Algorithm 1
4: end for
5: Compute p-value $p$ for testing if $(p_{\theta_i})_{i=1}^{B}$ has a uniform distribution.
6: return $p$

### 2.1 Two-Sample Test via Regression

Traditional approaches to comparing two distributions [40] are often not easily generalizable to high-dimensional and non-Euclidean data. More recent non-parametric extensions (see [41] for a review), e.g., maximum mean discrepancy (MMD) [42], energy distance (ED) [43], divergence [44, 45] and classification tests [46] 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 use a new regression-based approach to two-sample testing [47] that can adapt to any structure in $X$ where there is a suitable regression method; Theorem 1 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 $\hat{\mathcal{L}}(x; \theta_0)$ and $\mathcal{L}(x; \theta_0)$ in feature space $X$. We briefly describe the method below; see Supplementary Material E and [59] for theoretical details, and see Sections 2.2 and 3.2 for examples based on random forests regression.

1Note that a “global” test in feature space $X$ corresponds to a “local” test for fixed parameters $\theta \in \Theta$ (as in Algorithm 1) in our LFI setting.
Let \( P_0 \) be the distribution over \( \mathcal{X} \) induced by \( \mathcal{L}(x; \theta_0) \) and let \( P_1 \) be the distribution over \( \mathcal{X} \) induced by \( \mathcal{L}(x; \theta_0) \). 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 \( f(x|Y = 0) \) and \( f(x|Y = 1) \). The local null hypothesis is then equivalent to the hypothesis \( H_0 : f_0(x) = f_1(x) \) for all \( x \in \mathcal{X}_0 := \{ x \in \mathcal{X} : f(x) > 0 \} \), which in turn is equivalent to

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

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

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 \( m(x) = \mathbb{P}(Y = 1|X = x) \), we can adapt to nontraditional data settings involving mixed data types and various structures. More specifically, let \( \hat{m}(x) \) be an estimate of \( m(x) \) based on the sample \( \{(X_i, Y_i)\}_{i=1}^n \), and let \( \hat{\pi}_1 = \frac{1}{n} \sum_{i=1}^n I(Y_i = 1) \). We define our test statistic as

\[
\hat{T} = \frac{1}{n} \sum_{i=1}^n (\hat{m}(X_i) - \hat{\pi}_1)^2.
\]

(1)

Note that the difference \( |\hat{m}(x) - \hat{\pi}_1| \) for each particular value of \( x \in \mathcal{X} \) 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. Theorem 1 shows that if \( \hat{m} \), 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 given a limited number of simulations, we should choose a regression method that predicts their “class membership” \( Y \) well.

**Algorithm 3** Two-Sample Regression Test via Permutations

**Input:** two i.i.d. samples \( S_0 \) and \( S_1 \) from distributions resp. densities \( f_0 \) and \( f_1 \); number of permutations \( M \); a regression method \( \hat{m} \)

**Output:** \( p \)-value for testing if \( f_0(x) = f_1(x) \) for every \( x \)

1: Define an augmented sample \( \{X_i, Y_i\}_{i=1}^n \), where \( \{X_i\}_{i=1}^n = S_0 \cup S_1 \), and \( Y_i = I(X_i \in S_1) \).
2: Calculate the test statistic \( \hat{T} \) in Equation 1.\n3: Randomly permute \( Y_1, \ldots, Y_n \). Refit \( \hat{m} \) and calculate the test statistic using the permuted data.
4: Repeat the previous step \( M \) times to obtain \( \{\hat{T}^{(1)}, \ldots, \hat{T}^{(M)}\} \).
5: Approximate the permutation \( p \)-value by \( p = \frac{1}{M+1} \left( 1 + \sum_{m=1}^M I(\hat{T}^{(m)} > \hat{T}) \right) \).
6: return \( p \)

**Theorem 1.** Suppose that the regression estimator \( \hat{m}(x) \) is a linear smoother satisfying \( \sup_{m \in \mathcal{M}} \mathbb{E} \int_X (\hat{m}(x) - m(x))^2 dP_X(x) \leq C_0 \delta_n \), where \( C_0 \) is a positive constant, \( \delta_n = o(1) \), \( \delta_n \geq n^{-1} \), and \( \mathcal{M} \) is a class of regressions \( m(x) \) containing constant functions. Let \( t^*_\alpha \) be the upper \( \alpha \) quantile of the permutation distribution of the test statistic \( \hat{T}' \) on validation data. Then for any \( \alpha, \beta \in (0, 1/2) \) and \( n \) sufficiently large, there exists a universal constant \( C_1 \) such that

Type I error: \( \mathbb{P}_0 \left( \hat{T}' \geq t^*_\alpha \right) \leq \alpha \), and Type II error: \( \sup_{m \in \mathcal{M}(C_1 \delta_n)} \mathbb{P}_1 \left( \hat{T}' < t^*_\alpha \right) \leq \beta \)

against the class of alternatives \( \mathcal{M}(C_1 \delta_n) := \left\{ m \in \mathcal{M} : \int_X (m(x) - \pi_1)^2 dP_X(x) \geq C_1 \delta_n \right\} \).

**2.1.1 Theoretical Guarantees for Global Test**

Next we provide sufficient assumptions for the global test to be statistically consistent; i.e., to be able to detect a misspecified distribution (as in Example 1) for large sample sizes. These results apply regardless of the choice of the local test statistic.

**Definition 1.** Define \( \mathbb{D}_{B, n_{sim}} = \{P_0^{n_{sim}}, \ldots, P_{B}^{n_{sim}}\} \), where \( P_0^{n_{sim}}, \ldots, P_{B}^{n_{sim}} \) are the \( p \)-values obtained by Algorithm 3 using \( n_{sim, 1} = n_{sim, 2} = n_{sim} \), and \( \theta_1, \ldots, \theta_B \ i.i.d. \sim r(\theta) \). Let \( S(\mathbb{D}_{B, n_{sim}}) \) be the test
statistic for the global test. Also, denote by \( S(\mathcal{U}_B) \) the test statistic when \( \mathcal{U}_B = \{U_1, \ldots, U_B\} \), with 
\[ U_1, \ldots, U_B \overset{i.i.d.}{\sim} U(0, 1). \]

Assumption 1. Let \( D = \{ \theta : \mu_{\hat{\mathcal{L}}(\cdot: \theta)} \neq \mu_{\mathcal{L}(\cdot)} \} \), where \( \mu_{\mathcal{L}(\cdot)} \) (\( \mu_{\mathcal{L}(\cdot: \theta)} \)) is the measure over \( \mathcal{X} \) induced by \( \mathcal{L}(\cdot: \theta) \) (\( \hat{\mathcal{L}}(\cdot; \theta) \)). Assume that \( \mu_{\hat{\mathcal{L}}}(D) > 0 \), where \( \mu_{\hat{\mathcal{L}}} \) is the measure over \( \Theta \) induced by \( r(\theta) \).

Assumption 2. Assume that if \( \theta_1 \in D \), then the local test is such that \( p_{\hat{\theta}_1}^{\text{sim}} \overset{p}{\longrightarrow} 0 \). Moreover, if \( \theta_1 \notin D \), then the local test is such that \( p_{\hat{\theta}_1}^{\text{sim}} \sim U(0, 1) \).

Assumption 3. For every \( 0 < \alpha < 1 \), the test statistic \( S \) is such that \( F_{S(\mathcal{U}_B)}^{-1}(1 - \alpha) \overset{B \rightarrow \infty}{\longrightarrow} 0 \).

Assumption 4. Under Assumptions 1 and 2 there exists \( \alpha > 0 \) such that the test statistic \( S \) satisfies 
\[ S(\mathcal{D}_{B,n_{\text{sim}}}) \overset{p}{\longrightarrow} B, n_{\text{sim}} \rightarrow \infty \alpha. \]

Assumption 1 states that the set of parameter values where the likelihood function is incorrectly estimated has positive mass under the reference distribution. Assumption 2 states that the test chosen to perform the local comparison is statistically consistent and that its p-value has uniform distribution under the null hypothesis. Assumptions 3 and 4 state that the test statistic for the global comparison in step 5 of Algorithm 2 is statistically consistent, i.e., (i) it approaches zero under the null hypothesis when \( B \) increases, and (ii) it converges to a positive number if the null hypothesis is false.

Theorem 2. Let \( \phi \) be an \( \alpha \)-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 
\[ P(\phi_S(\mathcal{D}_{B,n_{\text{sim}}}) = 1) \overset{B, n_{\text{sim}} \rightarrow \infty}{\longrightarrow} 1 \]

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 Examples

We next use two synthetic examples to illustrate some of the advantages of our global and local tests to state-of-the-art validation techniques; such as, simulation-based calibration tests and distance-based two-sample tests like the MMD and ED tests.

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 2). Diagnostic tools like the Posterior Quantiles (PQ) technique [36] and Simulation-Based Calibration (SBC) [38] are often used to validate approximate likelihood models (see, e.g., [20]), but they are sometimes not able to tell the difference between the true model and a clearly misspecified model as illustrated by the following toy example where \( \theta_i \sim \text{Gamma}(1, 1) \), \( i = 1, \ldots, 500 \), and \( X_1, \ldots, X_{1000} \sim \text{Beta}(\theta_i, \theta_i) \).

The PQ test is based on the fact that, given a sample \( \hat{\theta} \) from the prior distribution, the posterior quantile \( g(\hat{\theta}) = \int f(\theta | x) I(\theta < \hat{\theta}) d\theta \) is uniformly distributed. Similarly, the SBC test relies on the fact that, given any ranking function \( g(\theta) \) and a posterior sample \( \{\theta_1, \ldots, \theta_L\} \), the rank 
\[ r \left( g(\theta_1), \ldots, g(\theta_L), g(\hat{\theta}) \right) = \sum_{i=1}^L I(g(\theta_i) < g(\hat{\theta})) \]
is uniformly distributed. Both PQ and SBC assess goodness-of-fit by checking if a histogram of respective statistics (posterior quantiles and ranks) is close to uniform.

Figure 1 left, shows the distribution of the statistics computed for PQ and SBC (along with confidence regions that describe what one would expect under uniformity) and the distribution of our local p-values (recall that the global test in Algorithm 2 is based on formally testing whether the local p-values are uniformly distributed) for two different cases: In the top row, we consider a case where \( \hat{\mathcal{L}}_{\hat{\theta}}(\theta) = \mathcal{L}_{\hat{x}}(\theta) \). All tests pass the model, as they should. In the bottom row, we consider a case where \( \hat{\mathcal{L}}_{\hat{\theta}}(\theta) \propto 1 \), a poor approximation of the likelihood function (see Supplementary Material A for examples). Our regression test (which is based on uniformity of the local p-values) clearly rejects this model. PQ and SBC, on the other hand, fail to distinguish the difference between the true likelihood.
and the misspecified model. Similar results 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).

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, \( H_0 : \hat{L}(x; \theta_0) = L(x; \theta_0) \) for every \( x \in X \). 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. We illustrate this with a synthetic example where \( X \in \mathbb{R}^D \), with a dimension \( D \) that could be large. We consider the following cases where the approximate likelihood \( \hat{L} \) and the true likelihood \( L \) only differ in the first dimension – that is, we test against a sparse alternative:

Bernoulli Case. True likelihood \( L(x; \theta) = f_b(x_1; \theta) \prod_{d=2}^D N(x_d; \theta, 1) \), where \( f_b(x_1; \theta) \) is a Bernoulli distribution with parameter \( \theta \) and \( N(x_d; \theta, 1) \) is a normal distribution with mean \( \theta \) and variance 1. The approximate likelihood \( \hat{L}(x; \theta) = \prod_{d=1}^D N(x_d; \theta, 1) \). Moreover, \( \theta \sim U(0, 1) \).

Scaling Case. True likelihood \( L(x; \theta) = N(x_1; 0, \theta) \prod_{d=2}^D N(x_d; 0, 1) \). The approximate likelihood \( \hat{L}(x; \theta) = \prod_{d=1}^D N(x_d; 0, 1) \). Moreover, \( \theta \sim U(0, 1) \).

Mixture of Gaussians Case. True likelihood \( L(x; \theta) = f_m(x_1; \theta, 1) \prod_{d=2}^D N(x_d; 0, 1) \) and approximate likelihood \( \hat{L}(x; \theta) = \prod_{d=1}^D N(x_d; 0, 1) \), where \( f_m \) is a mixture of two Gaussians centered at \( -\theta \) and \( \theta \) respectively such that \( x_1 \sim 1/2N(-\theta, 1) + 1/2N(\theta, 1) \). Moreover, \( \theta \sim U(-5, 5) \).

For each \( \theta \in \Theta \), we compute a local p-value by comparing samples of size \( n = 100 \) from \( L \) and \( \hat{L} \), respectively (Algorithm 3). This procedure is repeated 100 times to estimate the power function. We apply the local test for three different test statistics in these settings; namely: (i) the test statistic in Equation 1 using random forests regression, (ii) the maximum mean discrepancy (MMD) test statistic \[ Eq. 5 \] with a Gaussian kernel, and (iii) the energy test statistic \[ Eq. 5 \] using the Euclidean norm. Figure 1, right, shows how the power function varies with \( \theta \) at dimension \( D = 100 \) (top row) and how the power, averaged over \( \theta \), varies with \( D \) (bottom row) for each setting. When \( D = 1 \) (highlighted with circles in the right panel) distance-based tests yield higher power, but their performance quickly degrades with increasing \( D \). On the other hand, our regression-based test is able to achieve higher power in high-dimensional settings by leveraging some advantages of random forests regression; such as, the ability to select features, and the ability to tell discrete versus continuous distributions apart. For instance, in the scaling case (middle row) our regression test has higher power for small values of \( \theta \), 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 a cosmological application with weak lensing peak counts. The parameters of interest are the abundance of the matter distribution

---

**Figure 1**: Left: 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). (The grey ribbon represents the 99% confidence interval for the test of uniformity.) Right: Local test power as a function of \( \theta \) at \( D = 100 \) and across dimension \( D \). Distance-based tests are more powerful at \( D = 1 \) (highlighted with circles in the right panel), but their power is severely affected with increasing dimension. Our regression tests achieve higher power for large \( D \) by leveraging the advantages of random forests regression in high-dimensional settings with sparse structure.
We mimic these two properties by drawing (over 100 trials) MMD distance between the two samples, (ii) the power of a global test based on the local test distance. To better understand why the Gaussian and Poisson models fit poorly we can turn to the local test statistics. This is particularly true for the Gaussian fits, where we observe significant deviations from the data-generating process in the parameter space the model fits are insufficient. More specifically: for the Gaussian fits, the p-values in the left (θ₁ < 0.5) region are very small as are the p-values for the lower (θ₂ < 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 the uninformative model and eventually 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 (nₜₛₐₘ = 200). The data-generating process in our synthetic example induces four quadrants with different behaviors. Figure 2, 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 (θ₁ < 0.5) region are very small as are the p-values for the lower (θ₂ < 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 the uniform distribution of p-values at nₜₛₐₘ = 50 and eventually achieving p-values drawn from the uniform distribution for large values of nₜₛₐₘ. Our global test makes this observation rigorous.

3.2 Peak Count Data Example

For weak lensing peaks, we consider a 2D parameter space θ = (Ωₚ, σₚ), and design a grid of 50 different cosmologies. For each cosmology θ, we use CAMELUS simulator [49] to simulate peak counts with nₜₛₐₘ = 1000 and nₜₛₐₘ = 200. The peak histogram of each simulated map is a vector X ∈ ℝD where D = 13 is the number of bins. To assess models we compute the Kullback-Leibler (KL) divergence loss for the nₜₛₐₘ = 200 test simulations at each θ (see Supplementary Material B). We find that the Gaussian model performs best with a KL loss of 3.91 with the Poisson model closely behind at 4.02 and the KDE model last at 7.92. Based on the KL loss we would choose the Gaussian model. However, these
are only relative comparisons. We now use our test to find out whether the Gaussian fit is a good fit or merely better than the alternatives. As seen in Figure 3(left, top row), the local tests for the Gaussian model reject the null hypothesis of equality at all \( \theta \); thus the global hypothesis is also rejected. The Poisson and KDE models are rejected by the global test as well. As our initial tests indicate that the likelihoods are not well estimated, we increase \( n_{\text{train}} \) to 5000 for each \( \theta \) for illustrative purposes. With this sample size the KDE estimates are good enough to pass the global goodness-of-fit test (p-value = 0.239), whereas the Gaussian model still fails the test (similar to Figure 2 left).

When simulations are expensive, our objective is to achieve the best likelihood estimates at the lowest possible computational cost. Our local tests allow us to identify regions of \( \Theta \) where the current model is failing; hence allowing for better strategies as to where to place the next batch of additional simulations. Furthermore, our tests can provide insights into how the two distributions \( \hat{L}(x; \theta_0) \) and \( L(x; \theta_0) \) differ in feature space \( X \); more specifically, by evaluating how the estimate of the regression function \( \hat{m}(x) \) in Equation 1 varies with \( x \) for a fixed \( \theta \) (\( \hat{m}(x) \text{ far from } \hat{\pi}_1 \) is an indication that the model is not well estimated in that location of the feature space). We illustrate such an analysis for our fitted Gaussian model with \( n_{\text{train}} = 1000 \). According to the random forests regression used to build our test statistic, the most influential variables correspond to bins with low counts. In Figure 3 right, we explore the fit on one of these bins (variable \( X_9 \)) by a partial dependence plot (which shows the marginal effect of this variable on \( \hat{m}(x) \) [52]). Another example of detailed diagnostics in a multivariate feature space is provided in Supplementary Material C 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.
Figure 3: Left: Local goodness-of-fit $p$-values for peak-count data with $n_{\text{train}} = 1000$ training sample points. The Gaussian model is rejected at all $\theta$; Poisson and KDE perform better but the global test rejects for $n_{\text{train}} = 1000$ while it accepts KDE for $n_{\text{train}} = 5000$. Right: Partial dependence plot for variable $X_9$ for peak count data obtained using our local regression test on the Gaussian model. The estimated regression $\hat{m}(x)$ (solid curve) is often far from $\hat{\pi}_1$ (dashed horizontal line), which (according to our test statistic in Equation 1) indicates that the Gaussian model is not well fit. Moreover the difference $|\hat{m}(x) - \hat{\pi}_1|$ changes at integer values, showing that the regression test is distinguishing between the discrete true distribution and the continuous Gaussian distribution.

References

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Masanori Sato, Masahiro Takada, Takashi Hamana, and Takahiko Matsubara. Simulations of wide-field weak-lensing surveys II: covariance matrix of real-space correlation functions. *The Astrophysical Journal*, 734(2):76, 2011.

[9] Joachim Harnois-Déraps and Ludovic van Waerbeke. Simulations of weak gravitational lensing – II. including finite support effects in cosmic shear covariance matrices. *Monthly Notices of the Royal Astronomical Society*, 450(3):2857–2873, 2015.

[10] 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.

[11] 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.

[12] Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102, 2010.
[13] 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.

[14] 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.

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

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

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

[18] 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.

[19] Yanan Fan, David J. Nott, and Scott A. Sisson. Approximate bayesian computation via regression density estimation. *Stat*, 2(1):34–48, 2013.

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

[21] Jan-Matthis Lueckmann, Giacomo Bassetto, Theofanis Karaletsos, and Jakob H Macke. Likelihood-free inference with emulator networks. *arXiv preprint arXiv:1805.09294*, 2018.

[22] Justin Alsing, Benjamin Wandelt, and Stephen Feeney. Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology. *Monthly Notices of the Royal Astronomical Society*, 477(3):2874–2885, 03 2018.

[23] Justin Alsing, Tom Charnock, Stephen Feeney, and Benjamin Wandelt. Fast likelihood-free cosmology with neural density estimators and active learning. *arXiv preprint arXiv:1903.00007*, 02 2019.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Xin Tong. A plug-in approach to neyman-pearson classification. *Journal of Machine Learning Research*, 14:3011–3040, 2013.

[30] Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02166*, 2015.
[31] 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.

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

[33] Johann Brehmer, Kyle Cranmer, Gilles Louppe, and Juan Pavez. Constraining effective field theories with machine learning. Physical Review Letters, 121, 09 2018.

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

[35] Michael Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. Statistics and Computing, 07 2014.

[36] 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.

[37] 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.

[38] 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.

[39] Ilmun Kim, Ann B. Lee, and Jing Lei. Global and local two-sample tests via regression. arXiv preprint arXiv:1812.08927, 2018.

[40] Olivier Thas. Comparing distributions. Springer, 2010.

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

[46] Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two sample testing. arXiv preprint arXiv:1602.02210, 2016.

[47] LSST-DESC Photometric Redshift Working Group. An assessment of photometric redshift PDF performance in the context of LSST. Under internal review by LSST-DESC.

[48] L. Baringhaus and C. Franz. On a new multivariate two-sample test. J. Multivar. Anal., 88(1):190–206, January 2004.

[49] 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.

[50] M. P. Wand and Chris Jones. Multivariate plug-in bandwidth selection. Computational Statistics, 9(2):97–116, 1994.

[51] David S. Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic posterior transformation for likelihood-free inference. ICML, 2019.
[52] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[53] P E. Freeman, I Kim, and Ann Lee. Local two-sample testing: A new tool for analysing high-dimensional astronomical data. *Monthly Notices of the Royal Astronomical Society*, 471, 07 2017.

[54] 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. Roskan Cooray, D. J. Croton, E. Daddi, T. Dahlen, R. Davé, D. E. 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. Vilforth, 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. 197:35, December 2011.

[55] 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.

[56] P. E. Freeman, R. Izbicki, A. B. Lee, J. A. Newman, C. J. Conselice, A. M. Koekemoer, J. M. Lotz, and M. Mozena. New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society*, 434(1):282–295, 06 2013.

[57] 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.

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

[59] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.
Supplementary Material: Validation of Approximate Likelihood and Emulator Models for Computationally Intensive Simulations

A Example 1 (Consistency of Global Test)

In Example 1, we tested the null hypothesis that \( \hat{L}(\theta) = L_x(\theta) \) for data simulated according to \( \theta \sim \text{Gamma}(1, 1) \) and \( X|\theta \sim \text{Beta}(\theta, \theta) \). Figure 4 (left) shows the true likelihood \( L_x(\theta) \) for some different values of \( \theta \), comparing these functions to the likelihood approximation \( \hat{L}(\theta) \propto 1 \). Such an approximation is valid when \( \theta = 1 \), as \( \text{Beta}(1, 1) \) is indeed just the uniform distribution, whereas the approximation is clearly wrong for the other values of \( \theta \sim \text{Gamma}(1, 1) \).

B Peak Count Data Example

The KL divergence for model comparison is estimated by:

\[
KL(L, \hat{L}) = -\mathbb{E} \left[ \log \left( \frac{\hat{L}(x; \theta)}{L(x; \theta)} \right) \right] \\
= -\mathbb{E} \left[ \log \left( \hat{L}(x; \theta) \right) \right] + K \approx -\frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \hat{L}(x_{ij}; \theta_j) + K
\]

where \( K \) does not depend on \( \hat{L} \); \( \{\theta_j\}_{j=1}^{m} \) with \( m = 50 \) denotes the parameters used by the simulator; \( \{x_{ij}\}_{i=1}^{n_j} \) (with \( n_j = 200 \) for all \( \theta_j \)) denotes the test simulations at \( \theta_j \); and \( \sum_{j=1}^{m} n_j = n \) is the total number of test simulations.

Figure 4 (right) shows the grid of 50 parameters settings \( \theta = (\Omega_m, \sigma_8) \) which we use for the CAMELUS batch simulations.

C Galaxy Morphology Example: Investigating Difference Between Two Populations in Feature Space

Here we show how our regression approach can be used to investigate differences in distribution for two galaxy populations in a seven-dimensional morphology feature space \( X \); for details see [53]. We consider galaxies in the COSMOS, EGS, GOODS-North and UDS fields from CANDELS program [54] [55]. The available data consist of seven morphology summary statistics \( \{M, I, D, G, M_{20}, C, A\} \) [56] from 2736 galaxies, together with their star formation rates (SFR).
We first sort the galaxies according to their star formation rates, and we define two populations – with “high” \((Y = 1)\) versus “low” \((Y = 0)\) SFR – by taking the top and bottom 25\(^{th}\) quantiles, respectively. We then use 65\% of the data to train a random forests regression, and we use the remaining 35\% for testing. For every test point \(x\) (that is, for every galaxy in the test set), we compute the absolute difference \(|\hat{m}(x) - \hat{\pi}_1|\) between the estimated regression function and the proportion of high-SFR galaxies in the training sample. In Figure 5, we visualize the data via a two-dimensional diffusion map \([57]\). The colored points in the figure denote the regions in feature space where the local difference \(|\hat{m}(x) - \hat{\pi}_1|\) is statistically significant according to a permutation test with a false discovery rate correction at \(\alpha = 0.05\) via Benjamini-Hochberg’s method\([6]\). The blue points have \(\hat{m}(x) > \hat{\pi}_1\); these “high-SFR regions” are associated with extended, disturbed galaxy morphologies. The red points have \(\hat{m}(x) < \hat{\pi}_1\); these “low-SFR regions” are associated with concentrated, undisturbed morphologies. These results are consistent with what astronomers would expect, and illustrate the utility of the regression statistic \(|\hat{m}(x) - \hat{\pi}_1|\) in describing differences of two samples in a potentially high-dimensional feature space.

![Figure 5: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR galaxies. The test points are visualized via a two-dimensional diffusion map. (Courtesy of Ilmun Kim [39])](image)

**D Approximate P-Values and Confidence Regions**

Consider testing \(H_0 : \theta \in \Theta_0\). Let \(\lambda(x)\) be the likelihood ratio statistic for testing \(H_0\), i.e.,

\[
\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(x; \theta)}{\sup_{\theta \in \Theta} L(x; \theta)}.
\]

We estimate \(\lambda(x)\) using the estimated likelihood:

\[
\hat{\lambda}(x) = \frac{\sup_{\theta \in \Theta_0} \hat{L}(x; \theta)}{\sup_{\theta \in \Theta} \hat{L}(x; \theta)}.
\]

The estimated p-value is then

\[
\hat{p}(x) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\hat{\lambda}(X) > \hat{\lambda}(x))
\]

If \(\Theta_0 = \{\theta_0\}\), \(\hat{p}(x)\) can be estimated using data that is simulated under \(\theta = \theta_0\). If \(|\Theta_0| > 1\), the distribution of the test statistic can be approximated using the \(\chi^2\) approximation for the likelihood ratio test \([58]\). Confidence intervals may be obtained by inverting the hypothesis tests \([58]\).

\(^2\) Note that these tests are local in feature space and not the same as the “local tests” in parameter space described in Algorithm \([4]\).
E Proofs for the Local Test

Lemma 1. Suppose that we have a regression estimate satisfying

\[
\sup_{m \in \mathcal{M}} E \int_S (\hat{m}(x) - m(x))^2 \, dP_X(x) \leq C_0 \delta_n. \tag{2}
\]

We reject the null hypothesis when \( \hat{T}' \geq t_\alpha \) where \( t_\alpha = 2 \max\{C_0, 1/4\} \alpha^{-1} \delta_n \). Then for any \( \alpha, \beta \in (0, 1/2) \), there exists a universal constant \( C_1 \) such that

- Type I error: \( \mathbb{P}_0 (\hat{T}_{\text{global}}' \geq t_\alpha) \leq \alpha \) and
- Type II error: \( \sup_{m \in \mathcal{M}(C_1 \delta_n)} \mathbb{P}_1 (\hat{T}_{\text{global}}' < t_\alpha) \leq \beta \)

for a sufficiently large \( n \).

*Proof.* We start with analyzing the type I error of the test.

**Type I Error Control**

Under the null hypothesis, Markov’s inequality shows that

\[
\mathbb{P}_0 (\hat{T}_{\text{global}}' \geq t_\alpha) \leq \frac{\mathbb{E}_0 |\hat{T}_{\text{global}}'|}{t_\alpha} \\
\leq \frac{2}{t_\alpha} (\mathbb{E}_0 \left[ \int_S (\hat{m}(x) - \pi_1)^2 \, dP_X(x) \right] + \mathbb{E}_0 \left[ (\hat{\pi}_1 - \pi_1)^2 \right]) \\
\leq \frac{2}{t_\alpha} (C_0 \delta_n + \pi_1 (1 - \pi_1) n^{-1}) \leq \frac{2 \max\{C_0, 1/4\} \delta_n}{t_\alpha} = \alpha.
\]

Hence the result follows. Next, we control the type II error.

**Type II Error Control**

Based on the inequality \( (x - y)^2 \leq 2(x - z)^2 + 2(z - y)^2 \), we lower bound the test statistic as

\[
\hat{T}' = \frac{1}{n} \sum_{i=n+1}^{2n} (\hat{m}(X_i) - \hat{\pi}_1)^2 \\
\geq \frac{1}{2n} \sum_{i=n+1}^{2n} (m(X_i) - \hat{\pi}_1)^2 - \frac{1}{n} \sum_{i=n+1}^{2n} (\hat{m}(X_i) - m(X_i))^2 \\
\geq \frac{1}{4n} \sum_{i=n+1}^{2n} (m(X_i) - \pi_1)^2 - \frac{1}{2} (\pi_1 - \hat{\pi}_1)^2 - \frac{1}{n} \sum_{i=n+1}^{2n} (\hat{m}(X_i) - m(X_i))^2. \tag{3}
\]

Define the events \( \mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3 \) such that

\[
\mathcal{A}_1 = \left\{ (\pi_1 - \hat{\pi}_1)^2 < C_2 \delta_n \right\}, \\
\mathcal{A}_2 = \left\{ \frac{1}{n} \sum_{i=n+1}^{2n} (\hat{m}(X_i) - m(X_i))^2 < C_3 \delta_n \right\}, \\
\mathcal{A}_3 = \left\{ \frac{1}{n} \sum_{i=n+1}^{2n} (m(X_i) - \pi_1)^2 - \mathbb{E} [(m(X) - \pi_1)^2] < \frac{1}{2} \mathbb{E} [(m(X) - \pi_1)^2] \right\}.
\]
Using Markov’s inequality, we have
\[
P (A_1^t) \leq \frac{\pi_1(1 - \pi_1)}{C_2 n \delta_n},
\]
\[
P (A_2^t) \leq \frac{1}{C_3 n} \mathbb{E} \left[ \int_S (\hat{m}(x) - m(x))^2 dP_X(x) \right] \leq \frac{C_0}{C_3},
\]
by the condition in (3). For the third event, denote \( \Delta_n = \mathbb{E} [(m(X) - \pi_1)^2] \) and use Chebyshev’s inequality to have
\[
P (A_3^t) \leq \frac{4}{n \Delta_n^2} \text{Var} ((m(X) - \pi_1)^2)
\]
\[
\leq \frac{4}{n \Delta_n^2} \mathbb{E} [ (m(X) - \pi_1)^4 ]
\]
\[
\leq \frac{4}{n \Delta_n^2} \mathbb{E} [ (m(X) - \pi_1)^2 ] \quad \text{since } |m(X) - \pi_1| \leq 1
\]
\[
\leq \frac{4}{C_1 n \delta_n},
\]
where the last inequality uses the assumption that \( \Delta_n \geq C_1 \delta_n \). Hence, we obtain
\[
P ((A_1 \cap A_2 \cap A_3)) \leq P (A_1^t) + P (A_2^t) + P (A_3^t) < \beta,
\]
by choosing sufficiently large \( C_1, C_2, C_3 > 0 \) with the assumption that \( \delta_n \geq n^{-1} \). Using (3), the type II error of the regression test is bounded by
\[
P_1(\hat{T} \leq t_\alpha)
\]
\[
\leq P_1 \left( \frac{1}{4n} \sum_{i=n+1}^{2n} (m(X_i) - \pi_1)^2 - \frac{1}{2} (\pi_1 - \hat{\pi}_1)^2 \right) - \frac{1}{n} \sum_{i=n+1}^{2n} (\hat{m}(X_i) - m(X_i))^2 < t_\alpha
\]
\[
\leq P_1 \left( \frac{1}{4n} \sum_{i=n+1}^{2n} (m(X_i) - \pi_1)^2 - \frac{1}{2} (\pi_1 - \hat{\pi}_1)^2 - \frac{1}{n} \sum_{i=n+1}^{2n} (\hat{m}(X_i) - m(X_i))^2 < t_\alpha, A_1 \cap A_2 \cap A_3 \right) + P_1 \left( (A_1 \cap A_2 \cap A_3)^c \right)
\]
\[
\leq P_1 (\Delta_n < C_4 \delta_n) + \beta,
\]
where \( C_4 \) can be chosen by \( C_4 = 4C_2 + 8C_3 + 16 \max\{C_0, 1/4\}/\alpha \). Now by choosing \( C_1 > C_4 \) for sufficiently large \( n \), the type II error can be bounded by an arbitrary \( \beta > 0 \). Hence the result follows.

**Proof of Theorem 7.** The exact type I error control of the permutation test is well-known [see e.g. Chapter 15 of 59]. Hence we focus on the type II error control.

Let \( \eta = (\eta_1, \ldots, \eta_n)^T \) be a permutation of \( \{1, \ldots, n\} \). Now conditioned on the data \( X_{2n} = \{(X_1, Y_1), \ldots, (X_{2n}, Y_{2n})\} \), we denote the probability and expectation over permutations by \( P_{\eta[\cdot]} = P_{\eta[\cdot | X_{2n}]} \) and \( \mathbb{E}_{\eta[\cdot]} = \mathbb{E}_{\eta[\cdot | X_{2n}]} \) respectively. Then by Markov’s inequality
\[
P_{\eta} (\hat{T} \geq t_\alpha) = P_{\eta} \left( \frac{1}{n} \sum_{i=n+1}^{2n} (\hat{m}_{\eta}(X_i) - \hat{\pi}_1)^2 \geq t_\alpha \right) \leq \frac{1}{t_\alpha^2} \sum_{i=n+1}^{2n} \mathbb{E}_{\eta} [(\hat{m}_{\eta}(X_i) - \hat{\pi}_1)^2],
\]
where \( \hat{m}_{\eta}(x) = \sum_{i=1}^{n} w_i(x) Y_{\eta_i}. \) Since \( \sum_{i=1}^{n} w_i(x) = 1 \) for any \( x \in S \),
\[
\mathbb{E}_{\eta} [\hat{m}_{\eta}(x)] = \sum_{i=1}^{n} w_i(x) \mathbb{E}_{\eta}[Y_{\eta_i}] = \sum_{i=1}^{n} w_i(x) \hat{\pi}_1 = \hat{\pi}_1.
\]
Further note that
\[ E_\eta \left[ (\hat{m}_\eta(x) - \hat{\pi}_1)^2 \right] = \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} w_{i_1}(x)w_{i_2}(x)E_\eta \left[ (Y_{i_1} - \hat{\pi}_1)(Y_{i_2} - \hat{\pi}_1) \right] \]
\[ \leq \sum_{i=1}^{n} w_i^2(x)E_\eta \left[ (Y_i - \hat{\pi}_1)^2 \right] \]
\[ = \hat{\pi}_1(1 - \hat{\pi}_1) \sum_{i=1}^{n} w_i^2(x) \leq \frac{1}{4} \sum_{i=1}^{n} w_i^2(x), \]
where the first inequality uses \( E_\eta \left[ (Y_{i_1} - \hat{\pi}_1)(Y_{i_2} - \hat{\pi}_1) \right] \leq 0 \) when \( i_1 \neq i_2 \).

Note that the permutation samples are not i.i.d. and thus in order to use the condition in (2) which holds for i.i.d. samples, we will associate the upper bound in (4) with i.i.d. samples. To do so, let \( (Y_1^*, \ldots, Y_n^*) \) be i.i.d. Bernoulli random variables with parameter \( p = 1/2 \) independent of \( \{X_1, \ldots, X_{2n}\} \). Then
\[ E_{Y^*} \left[ (\hat{m}(x) - 1/2)^2 | X_1, \ldots, X_{2n} \right] \]
\[ = E_{Y^*} \left[ (\sum_{i=1}^{n} w_i(x)Y_i^* - 1/2)^2 | X_1, \ldots, X_{2n} \right] \]
\[ = E_{Y^*} \left[ (\sum_{i=1}^{n} w_i(x)(Y_i^* - 1/2))^2 | X_1, \ldots, X_{2n} \right] \]
\[ = \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} w_{i_1}(x)w_{i_2}(x)E_{Y^*}[(Y_{i_1}^* - 1/2)(Y_{i_2}^* - 1/2)] \]
\[ = \frac{1}{4} \sum_{i=1}^{n} w_i^2(x). \]

Therefore, we obtain
\[ E_\eta \left[ (\hat{m}_\eta(x) - \hat{\pi}_1)^2 \right] \leq E_{Y^*} \left[ (\hat{m}(x) - 1/2)^2 | X_1, \ldots, X_{2n} \right] \]
which in turn implies that
\[ \mathbb{P}_\eta \left( \hat{T}' \geq t'^*_\alpha \right) \leq \frac{1}{\alpha n} \sum_{i=n+1}^{2n} E_{Y^*} \left[ (\hat{m}(X_i) - 1/2)^2 | X_1, \ldots, X_{2n} \right]. \]

So the critical value of the permutation distribution is bounded by
\[ t'^*_\alpha \leq \frac{1}{\alpha n} \sum_{i=n+1}^{2n} E_{Y^*} \left[ (\hat{m}(X_i) - 1/2)^2 | X_1, \ldots, X_{2n} \right]. \]

Next, define the event
\[ \mathcal{A} = \left\{ \frac{1}{n} \sum_{i=n+1}^{2n} E_{Y^*} \left[ (\hat{m}(X_i) - 1/2)^2 | X_1, \ldots, X_{2n} \right] \leq C'\delta_n \right\}. \]

Now, because we assume that
\[ \sup_{m \in \mathcal{M}} \mathbb{E} \int_{S} (\hat{m}(x) - m(x))^2 dP_X(x) \leq C_0 \delta_n, \]
by Markov’s inequality it holds that
\[ \mathbb{P} (\mathcal{A}^c) \leq \mathbb{P} \left( \frac{1}{n} \sum_{i=n+1}^{2n} E_{Y^*} \left[ (\hat{m}(X_i) - 1/2)^2 | X_1, \ldots, X_{2n} \right] > C'_\alpha \delta_n \right) \leq \frac{C_0}{C'_2}. \]
As a result, the type II error of the permutation test is bounded by
\[
P_1 \left( \tilde{T}' < t^*_n \right) \leq P_1 \left( \tilde{T}' < t^*_n, A \right) + P_1 \left( A^c \right) \leq P_1 \left( \tilde{T}' < \frac{C'_2 \delta_n}{\alpha} \right) + \frac{C_0}{C'_2},
\]
Now we choose \( C'_2 \) sufficiently large so that
\[
\frac{C_0}{C'_2} < \frac{\beta}{2}.
\]
Next we follow the proof of Lemma [1] to show that
\[
P_1 \left( \tilde{T}' < \frac{C'_2 \delta_n}{\alpha} \right) < \frac{\beta}{2},
\]
which completes the proof.

### F Proofs for the Global Test

**Lemma 2.** Let \( \hat{F}_{D_{B,n_{sim}}} \) be the empirical cumulative distribution of the p-values in \( \mathbb{D}_{B,n_{sim}} \),
\[
KS(\mathbb{D}_{B,n_{sim}}) = \sup_{0 \leq z \leq 1} |\hat{F}_{D_{B,n_{sim}}}(z) - z|,
\]
be the Kolmogorov-Smirnoff test statistic and
\[
CVM(\mathbb{D}_{B,n_{sim}}) = \int_0^1 \left( \hat{F}_{D_{B,n_{sim}}}(z) - z \right)^2 dz
\]
be the Cramér-von Mises test statistic. Both KS and CVM satisfy Assumptions [3] and [7].

**Proof.** Let \( U \sim U(0, 1) \). From the law of large numbers,
\[
KS(U_B) = \sup_{0 \leq z \leq 1} |\hat{F}_{U_B}(z) - z| \xrightarrow{a.s.} \sup_{0 \leq z \leq 1} |\mathbb{P}(U \leq z) - z| = 0,
\]
which proofs the first statement of the theorem. Similarly, for every \( n_{sim} \in \mathbb{N} \),
\[
KS(\mathbb{D}_{B,n_{sim}}) = \sup_{0 \leq z \leq 1} |\hat{F}_{D_{B,n_{sim}}}(z) - z| \xrightarrow{a.s.} \sup_{0 \leq z \leq 1} |\mathbb{P}(p_{\theta_1}^{n_{sim}} \leq z) - z|.
\]
(6)

Now, Under Assumption [2] for every \( \theta_1 \in D \),
\[
\mathbb{P}(p_{\theta_1}^{n_{sim}} \leq z | \theta_1) \xrightarrow{n_{sim} \rightarrow \infty} 1
\]
uniformly over \( z \in (0, 1) \). Thus, under Assumption [1], for every \( 0 < \epsilon_z < 1 - z \), there exists \( n_{sim} \in \mathbb{N} \) such that, for every \( n_{sim}' > n_{sim} \),
\[
\mathbb{P}(p_{\theta_1}^{n_{sim}'} \leq z | \theta_1 \in D) \geq (1 - \epsilon_z) \mathbb{P}(\theta_1 \in D) + z \mathbb{P}(\theta_1 \notin D)
\]
\[
= (1 - \epsilon_z + z - z) \mathbb{P}(\theta_1 \in D) + z \mathbb{P}(\theta_1 \notin D)
\]
\[
= (1 - \epsilon_z - z) \mathbb{P}(\theta_1 \in D) + z
\]
(7)

It follows from Equations [6] and [7] and by taking \( \epsilon_z = (1 - z)/2 \) that
\[
\sup_{0 \leq z \leq 1} |\mathbb{P}(p_{\theta_1}^{n_{sim}'} \leq z) - z| \geq \sup_{0 \leq z \leq 1} (1 - \epsilon_z - z) \mathbb{P}(\theta_1 \in D)
\]
\[
\geq \mathbb{P}(\theta_1 \in D) \sup_{0 \leq z \leq 1} \left( \frac{1 - z}{2} \right) = \frac{\mathbb{P}(\theta_1 \in D)}{2},
\]
and hence
\[
\lim_{n_{sim} \rightarrow \infty} \sup_{0 \leq z \leq 1} |\mathbb{P}(p_{\theta_1}^{n_{sim}} \leq z) - z| \geq \frac{\mathbb{P}(\theta_1 \in D)}{2} > 0,
\]
which concludes the proof for the KS statistic. The proof for the CVM statistic is analogous. \( \square \)
Proof of Theorem 2. Assumption 2 implies that $\phi_S$ is such that

$$\phi_S(D_{B,n}) = 1 \iff S(D_{B,n}) \geq F_{S(U_B)}^{-1}(1 - \alpha).$$

It follows that

$$P(\phi_S(D_{B,n}) = 1) = P\left(S(D_{B,n}) - F_{S(U_B)}^{-1}(1 - \alpha) \geq 0\right)$$

$$\geq P\left(|S(D_{B,n}) - a - F_{S(U_B)}^{-1}(1 - \alpha)| \leq a\right) \xrightarrow{B,n \to \infty} 1,$$

where the last line follows from Assumptions 3 and 4.

Proof of Corollary 1. It follows directly from Theorem 2 and Lemma 2.