Supplementary Material: Validation of Approximate Likelihood and Emulator Models for Computationally Intensive Simulations

A Identifying Differences Between Two Multivariate Distributions

Here we show how our regression approach can be used to identify and visualize locally significant differences between two multivariate distributions P_0 and P_1 defined over a "feature space" $\mathcal{X}$; we denote samples from the respective distributions by $\mathcal{S}_0$ and $\mathcal{S}_1$. In this example, which is adapted from Kim et al. (2016), S_0 and S_1 are both real observed data (galaxy images), but in our LFI setting S_0 would represent a sample from the simulator or target likelihood $\mathcal{L}(\mathbf{x};\theta)$ at a fixed parameter value θ, and S_1 would represent a sample from the emulator or approximate likelihood $\widehat{\mathcal{L}}(\mathbf{x};\theta)$ (for the same parameter value). The goal would then be to identify with statistical confidence the regions in $\mathcal{X}$ which may be under- or over-represented in $\mathcal{S}_1$ (as compared to S_0). Our techniques can also be used to validate and diagnose output from generative adversarial networks (GANs) and other so-called implicit generative models Mohamed and Lakshminarayanan (2016); e.g., this type of analysis could be relevant for recent GAN models of galaxy images (Ravanbakhsh et al., 2017) and weak lensing convergence maps (Mustafa et al., 2019).

Algorithm 4 Local Test in Feature Space

```
Input: i.i.d. training data from two populations $\{\mathbf{X}_i, Y_i\}_{i=1}^n$; testing data $\{\mathbf{X}_j\}_{j=1}^J$; number of permutations $M$; significance level $\alpha$; a regression method $\hat{m}$
Output: p-values $\{p_j\}_{j=1}^M$ for testing significance of difference $|\widehat{m}(\mathbf{X}_j) - \widehat{\pi}_1|$ for every test point
$\widehat{\pi}_1 = 1/n \sum_{i=1}^n Y_i$;
Train regression method $\hat{m}$ on training data $\{\mathbf{X}_i, Y_i\}_{i=1}^n$;
Calculate the test statistics on each of the test points
    $\hat{\nu}(\mathbf{X}_j) = (\hat{m}(\mathbf{X}_j) - \widehat{\pi}_1)^2$;
for $k$ in $1, ..., M$ do
    Randomly permute $Y_1, ..., Y_n$ and train regression method on permuted data $\hat{m}^{(k)}$;
    Calculate the test statistics on the permuted data $\{\hat{\nu}^{(k)}(\mathbf{X}_j) = (\hat{m}^{(k)}(\mathbf{X}_j) - \widehat{\pi}_1)^2\}_{j=1}^J$;
end for

Approximate permutation p-values $p_j$ for every test point $\mathbf{X}_j$:
    $p_j = \frac{1}{M+1} \sum_{k=1}^M \left(1 + \mathbb{I}(\hat{\nu}^{(k)}(\mathbf{X}_j) > \hat{\nu}(\mathbf{X}_j))\right)$
Apply a multiple test procedure to control false discovery rate;
return $\{p_j\}_{j=1}^J$
```

Galaxy Morphology Example

Here we consider galaxies in the COSMOS, EGS, GOODS-North and UDS fields from CANDELS program (Grogin et al., 2011; Koekemoer et al., 2011). The available data consist of seven morphology summary statistics 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" SFR ($Y = 1$) versus "low" SFR ($Y = 0$) — by taking the top and bottom 25th quantiles, respectively. Figure 5 shows a random subset of 12 galaxies from each sample.

To compare the two populations in distribution, we use 65% of the data to train a random forests regression, and the remaining 35% for testing. For every test point $\mathbf{x}$ (that is, for every galaxy images in the test set), we compute the absolute difference $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ between the estimated regression function and the proportion of high-SFR galaxies in the training sample. We then calculate whether the difference $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ is *statistically significant* according to a permutation test with a false discovery rate correction at $\alpha = 0.05$ via Benjamini-Hochberg's method. The details of the local test in feature space are outlined in Algorithm 4.

Figure 6 shows examples of galaxies associated with the highest significant difference $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$; galaxies that are more representative of one sample than the other. In Figure 7 we visualize the test data via a two-dimensional diffusion map (Coifman et al., 2005), where we color the test points that occur in regions of feature space where the local differences in the two distributions are statistically significant. The blue points have $\widehat{m}(\mathbf{x}) > \widehat{\pi}_1$; these "high-SFR regions" are associated with extended, disturbed galaxy morphologies. The red points have $\widehat{m}(\mathbf{x}) < \widehat{\pi}_1$; 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 $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ in describing differences of two samples in a potentially high-dimensional feature space. For further details, see (Kim et al., 2016; Freeman et al., 2017).

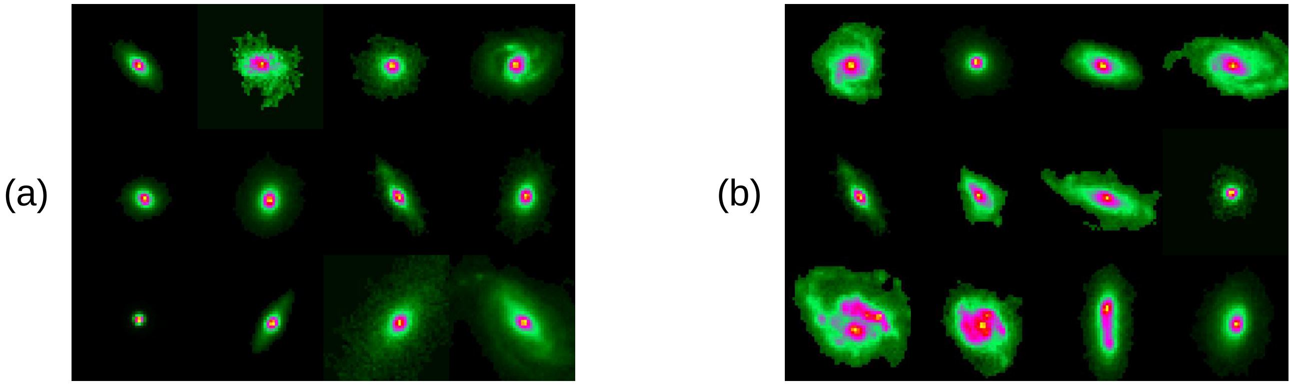

Figure 5: Examples of galaxies from (a) the low-SFR sample $\mathcal{S}_0$ versus (b) the high-SFR sample $\mathcal{S}_1$.

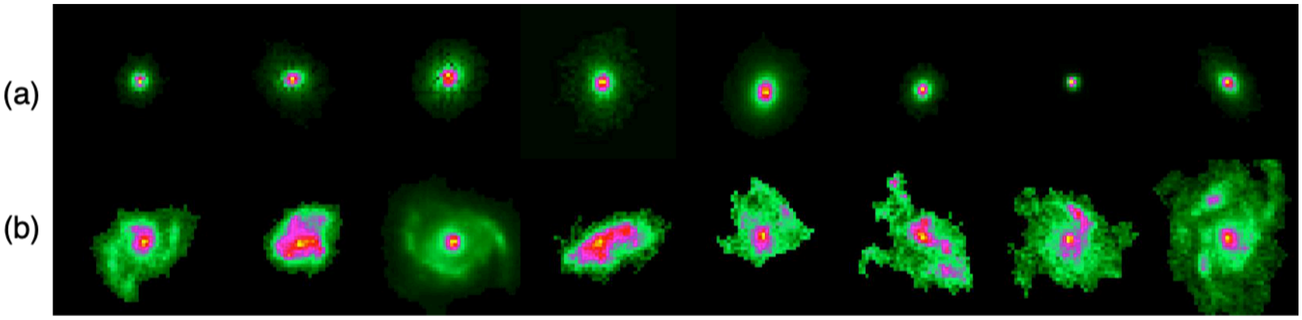

Figure 6: Galaxies in the test set with the highest significant difference $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ according to our local test in feature space, Algorithm 4. (a) Galaxies that are more representative of the low-SFR sample $\mathcal{S}_0$, and (b) galaxies more representative of the high-SFR sample $\mathcal{S}_1$. The first group of galaxies presents undisturbed and concentrated morphologies, while the latter galaxies appear more extended and/or disturbed. This is in line with what is expected by astronomers when comparing actual low-SFR and high-SFR galaxies.

B Proofs for the Global Test

In this section we first provide sufficient assumptions for Theorem 1 to hold. Corollary 1 then follows by the fact that both Kolmogorov-Smirnoff and Cramér-von Mises test statistic are statistically consistent (i.e., satisfy Assumptions 3 and 4).

Definition 1. *Let $S(\mathbb{D}_{B,n_{sim}})$ be the test statistic for the global test. Also, denote by $S(\mathbb{U}_B)$ the test statistic when $\mathbb{U}_B = (U_1, \ldots, U_B)$, with $U_1, \ldots, U_B \overset{i.i.d.}{\sim} U(0,1)$.*

Assumption 1. *Let $D = \left\{\theta : \mu_{\widehat{\mathcal{L}}(\cdot;\theta)} \neq \mu_{\mathcal{L}\cdot(\theta)}\right\}$, where $\mu_{\widehat{\mathcal{L}}(\cdot;\theta)}$ ($\mu_{\mathcal{L}(\cdot;\theta)}$) is the measure over $\mathcal{X}$ induced by $\mathcal{L}(\cdot;\theta)$ ($\widehat{\mathcal{L}}(\cdot;\theta)$). Assume that $\mu_r(D) > 0$, where μ_r is the measure over Θ induced by $r(\theta)$.*

Assumption 2. *Assume that if $\theta_1 \in D$, then the local test is such that $p_{\theta_1}^{n_{sim}} \xrightarrow[n_{sim}\longrightarrow\infty]{\mathbb{P}} 0$. Moreover, if $\theta_1 \notin D$, then the local test is such that $p_{\theta_1}^{n_{sim}} \sim U(0,1)$.*

Assumption 3. *For every $0 < \alpha < 1$, the test statistic S is such that $F^{-1}_{S(\mathbb{U}_B)}(1-\alpha) \xrightarrow{B\longrightarrow\infty} 0$.*

Assumption 4. *Under Assumptions 1 and 2, there exists $a > 0$ such that the test statistic S satisfies*

$$S(\mathbb{D}_{B,n_{sim}}) \xrightarrow[B,n_{sim}\longrightarrow\infty]{\mathbb{P}} a.$$

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 comparisons 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. Under these four assumptions, we can guarantee statistical consistency.

Lemma 1. *Let $\widehat{F}_{\mathbb{D}_{B,n_{sim}}}$ be the empirical cumulative distribution of the p-values in $\mathbb{D}_{B,n_{sim}}$,*

$$KS(\mathbb{D}_{B,n_{sim}}) = \sup_{0\leq z\leq 1} |\widehat{F}_{\mathbb{D}_{B,n_{sim}}}(z) - z|,$$

be the Kolmogorov-Smirnoff test statistic and

$$CVM(\mathbb{D}_{B,n_{sim}}) = \int_0^1 \left(\widehat{F}_{\mathbb{D}_{B,n_{sim}}}(z) - z\right)^2 dz$$

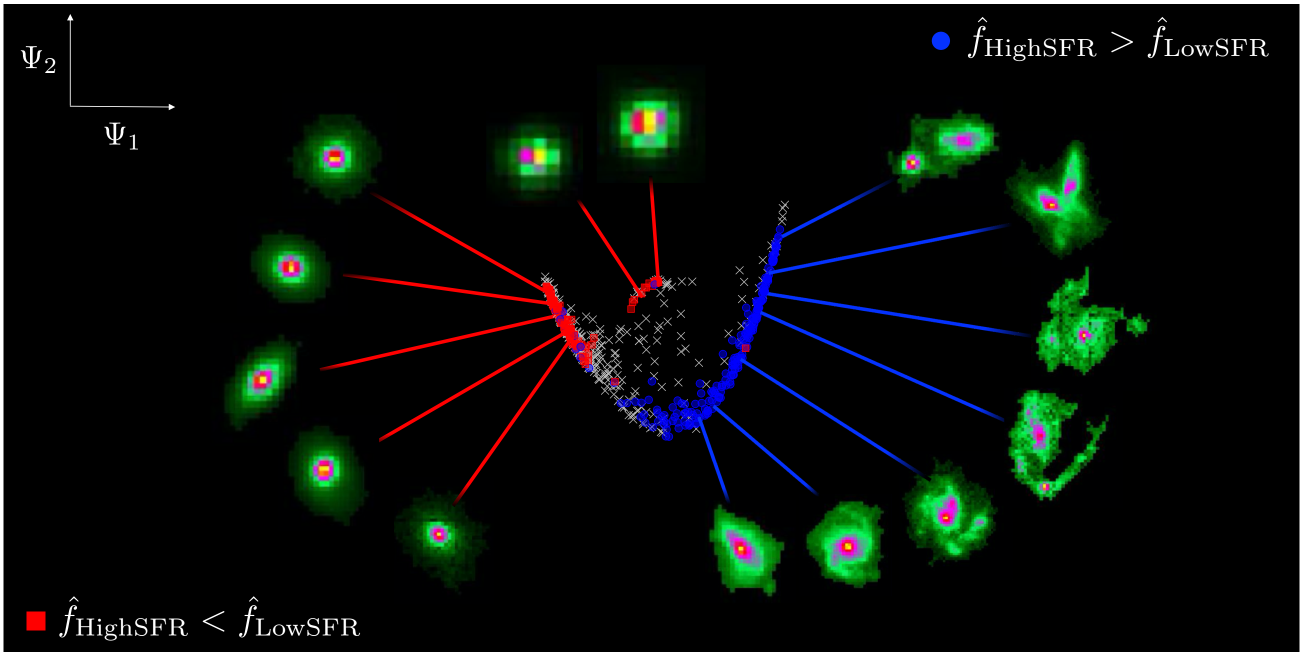

Figure 7: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR galaxies. The test points are visualized via a two-dimensional diffusion map. Figure adapted from Kim et al. (2018).

be the Cramér-von Mises test statistic. Both KS and CVM satisfy Assumptions 3 and 4.

Proof of Lemma 1. Let $U \sim U(0,1)$. From the law of large numbers,

$$\mathrm{KS}(\mathbb{U}_B) = \sup_{0\le z\le 1} |\widehat{F}_{\mathbb{U}_B}(z) - z| \xrightarrow[B\longrightarrow\infty]{\text{a.s.}} \sup_{0\le z\le 1} |\mathbb{P}(U \le z) - z| = 0,$$

which proves the first statement of the theorem. Similarly, for every $n_{\text{sim}} \in \mathbb{N}$,

$$\mathrm{KS}(\mathbb{D}_{B,n_{\text{sim}}}) = \sup_{0\le z\le 1} |\widehat{F}_{\mathbb{D}_{B,n_{\text{sim}}}}(z) - z| \xrightarrow[B\longrightarrow\infty]{\text{a.s.}} \sup_{0\le z\le 1} |\mathbb{P}(p_{\theta_1}^{n_{\text{sim}}} \le z) - z|. \tag{2}$$

Now, Under Assumption 2, for every $\theta_1 \in D$,

$$\mathbb{P}(p_{\theta_1}^{n_{\text{sim}}} \le z|\theta_1) \xrightarrow{n_{\text{sim}}\longrightarrow\infty} 1$$

uniformly over $z \in (0,1)$. Thus, under Assumption 1, for every $0 < \epsilon_z < 1 - z$, there exists $n_{\text{sim}} \in \mathbb{N}$ such that, for every $n'_{\text{sim}} > n_{\text{sim}}$,

$$\begin{aligned}\mathbb{P}(p_{\theta_1}^{n'_{\text{sim}}} \le z) &= \mathbb{P}(p_{\theta_1}^{n'_{\text{sim}}} \le z|\theta_1 \in D)\mathbb{P}(\theta_1 \in D) + \\ &\quad \mathbb{P}(p_{\theta_1}^{n'_{\text{sim}}} \le z|\theta_1 \notin D)\mathbb{P}(\theta_1 \notin D) \\ &\ge (1-\epsilon_z)\mathbb{P}(\theta_1 \in D) + z\mathbb{P}(\theta_1 \notin D) \\ &= (1-\epsilon_z + z - z)\mathbb{P}(\theta_1 \in D) + z\mathbb{P}(\theta_1 \notin D) \\ &= (1-\epsilon_z - z)\mathbb{P}(\theta_1 \in D) + z \end{aligned} \tag{3}$$

It follows from Equations 2 and 3 and by taking $\epsilon_z = (1-z)/2$ that

$$\begin{aligned}\sup_{0\le z\le 1} |\mathbb{P}(p_{\theta_1}^{n'_{\text{sim}}} \le z) - z| &\ge \sup_{0\le z\le 1} (1-\epsilon_z - z)\mathbb{P}(\theta_1 \in D) \\ &\ge \mathbb{P}(\theta_1 \in D) \sup_{0\le z\le 1} \frac{(1-z)}{2} = \frac{\mathbb{P}(\theta_1 \in D)}{2},\end{aligned}$$

and hence

$$\lim_{n'_{\text{sim}}\longrightarrow\infty} \sup_{0\le z\le 1} |\mathbb{P}(p_{\theta_1}^{n'_{\text{sim}}} \le z) - z| \ge \frac{\mathbb{P}(\theta_1 \in D)}{2} > 0,$$

which concludes the proof for the KS statistic. The proof for the CVM statistic is analogous. ☐

Proof of Theorem 1. Assumption 2 implies that ϕ_S is such that

$$\phi_S(\mathbb{D}_{B,n_{\text{sim}}}) = 1 \iff S(\mathbb{D}_{B,n_{\text{sim}}}) \ge F^{-1}_{S(\mathbb{U}_B)}(1-\alpha).$$

It follows that

$$\begin{aligned}&\mathbb{P}\left(\phi_S(\mathbb{D}_{B,n_{\mathrm{sim}}})=1\right)\\&=\mathbb{P}\left(S(\mathbb{D}_{B,n_{\mathrm{sim}}})-F^{-1}_{S(\mathbb{U}_B)}(1-\alpha)\geq 0\right)\\&\geq\mathbb{P}\left(|S(\mathbb{D}_{B,n_{\mathrm{sim}}})-a-F^{-1}_{S(\mathbb{U}_B)}(1-\alpha)|\leq a\right)\\&\xrightarrow{B,n_{\mathrm{sim}}\longrightarrow\infty}1,\end{aligned}$$

where the last line follows from Assumptions 3 and 4. ☐

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

C Proofs for Two-Sample Testing via Regression

Lemma 2. *Suppose that we have a regression estimate satisfying*

$$\sup_{m\in\mathcal{M}}\mathbb{E}\int_S\left(\widehat{m}(x)-m(x)\right)^2dP_X(x)\leq C_0\delta_n.\tag{4}$$

We reject the null hypothesis when $\widehat{\mathcal{T}}'\geq t_\alpha$ *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:* $\mathbb{P}_0\left(\widehat{\mathcal{T}}'\geq t_\alpha\right)\leq\alpha$ *and*
- *Type II error:* $\sup_{m\in\mathcal{M}(C_1\delta_n)}\mathbb{P}_1\left(\widehat{\mathcal{T}}'<t_\alpha\right)\leq\beta$

for a sufficiently large n.

Proof of Lemma 2. We start with analyzing the type I error of the test.

• Type I Error Control

Under the null hypothesis, Markov's inequality shows that

$$\begin{aligned}&\mathbb{P}_0\left(\widehat{\mathcal{T}}'\geq t_\alpha\right)\leq\frac{\mathbb{E}_0[\widehat{\mathcal{T}}']}{t_\alpha}\\&\leq\frac{2}{t_\alpha}(\mathbb{E}_0\left[\int_S\left(\widehat{m}(x)-\pi_1\right)^2dP_X(x)\right]+\mathbb{E}_0\left[\left(\widehat{\pi}_1-\pi_1\right)^2\right])\\&\leq\frac{2}{t_\alpha}\left(C_0\delta_n+\pi_1(1-\pi_1)n^{-1}\right)\\&\leq\frac{2\max\{C_0,1/4\}\delta_n}{t_\alpha}=\alpha.\end{aligned}$$

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

$$\begin{aligned}\widehat{\mathcal{T}}'&=\frac{1}{n}\sum_{i=n+1}^{2n}\left(\widehat{m}(X_i)-\widehat{\pi}_1\right)^2\\&\geq\frac{1}{2n}\sum_{i=n+1}^{2n}\left(m(X_i)-\widehat{\pi}_1\right)^2\\&\quad-\frac{1}{n}\sum_{i=n+1}^{2n}\left(\widehat{m}(X_i)-m(X_i)\right)^2\qquad(5)\\&\geq\frac{1}{4n}\sum_{i=n+1}^{2n}\left(m(X_i)-\pi_1\right)^2-\frac{1}{2}(\pi_1-\widehat{\pi}_1)^2\\&\quad-\frac{1}{n}\sum_{i=n+1}^{2n}\left(\widehat{m}(X_i)-m(X_i)\right)^2.\qquad(6)\end{aligned}$$

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

$$\mathcal{A}_1=\left\{(\pi_1-\widehat{\pi}_1)^2<C_2\delta_n\right\},$$

$$\mathcal{A}_2=\left\{\frac{1}{n}\sum_{i=n+1}^{2n}\left(\widehat{m}(X_i)-m(X_i)\right)^2<C_3\delta_n\right\},$$

$$\begin{aligned}\mathcal{A}_3=\Bigg\{&\Bigg|\frac{1}{n}\sum_{i=n+1}^{2n}\left(m(X_i)-\pi_1\right)^2-\mathbb{E}\left[(m(X)-\pi_1)^2\right]\Bigg|\\&<\frac{1}{2}\mathbb{E}\left[(m(X)-\pi_1)^2\right]\Bigg\}.\end{aligned}$$

Using Markov's inequality, we have

$$\mathbb{P}\left(\mathcal{A}_1^c\right)\leq\frac{\pi_1(1-\pi_1)}{C_2n\delta_n},$$

$$\mathbb{P}\left(\mathcal{A}_2^c\right)\leq\frac{1}{C_3\delta_n}\mathbb{E}\left[\int_S(\widehat{m}(x)-m(x))^2dP_X(x)\right]\leq\frac{C_0}{C_3},$$

by the condition in (4). For the third event, denote $\Delta_n=\mathbb{E}\left[(m(X)-\pi_1)^2\right]$ and use Chebyshev's inequality to have

$$\begin{aligned}\mathbb{P}\left(\mathcal{A}_3^c\right)&\leq\frac{4}{n\Delta_n^2}\mathrm{Var}\left((m(X)-\pi_1)^2\right)\\&\leq\frac{4}{n\Delta_n^2}\mathbb{E}\left[(m(X)-\pi_1)^4\right]\\&\leq\frac{4}{n\Delta_n^2}\mathbb{E}\left[(m(X)-\pi_1)^2\right]\quad\text{since }|m(X)-\pi_1|\leq 1\\&\leq\frac{4}{C_1n\delta_n},\end{aligned}$$

where the last inequality uses the assumption that $\Delta_n\geq C_1\delta_n$. Hence, we obtain

$$\mathbb{P}\left((\mathcal{A}_1\cap\mathcal{A}_2\cap\mathcal{A}_3)^c\right)\leq\mathbb{P}\left(\mathcal{A}_1^c\right)+\mathbb{P}\left(\mathcal{A}_2^c\right)+\mathbb{P}\left(\mathcal{A}_3^c\right)<\beta,$$

by choosing sufficiently large $C_1, C_2, C_3 > 0$ with the assumption that $\delta_n \geq n^{-1}$. Using (6), the type II error of the regression test is bounded by

$$\begin{aligned}
&\mathbb{P}_1(\widehat{\mathcal{T}}' < t_\alpha) \\
\leq\ & \mathbb{P}_1\Big(\frac{1}{4n}\sum_{i=n+1}^{2n}(m(X_i)-\pi_1)^2 - \frac{1}{2}(\pi_1-\widehat{\pi}_1)^2 \\
&\qquad - \frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}(X_i)-m(X_i))^2 < t_\alpha\Big) \\
\leq\ & \mathbb{P}_1\Big(\frac{1}{4n}\sum_{i=n+1}^{2n}(m(X_i)-\pi_1)^2 - \frac{1}{2}(\pi_1-\widehat{\pi}_1)^2 \\
&\qquad - \frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}(X_i)-m(X_i))^2 < t_\alpha, \mathcal{A}_1\cap\mathcal{A}_2\cap\mathcal{A}_3\Big) \\
&\qquad + \mathbb{P}_1\left((\mathcal{A}_1\cap\mathcal{A}_2\cap\mathcal{A}_3)^c\right) \\
\leq\ & \mathbb{P}_1\left(\Delta_n < C_4\delta_n\right) + \beta,
\end{aligned}$$

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 2. The exact type I error control of the permutation test is well-known (see e.g. Chapter 15 of Lehmann and Romano, 2006). Hence we focus on the type II error control.

Let $\eta = (\eta_1, \ldots, \eta_n)^\top$ be a permutation of $\{1, \ldots, n\}$. Now conditioned on the data $\mathcal{X}_{2n} = \{(X_1, Y_1), \ldots, (X_{2n}, Y_{2n})\}$, we denote the probability and expectation over permutations by $\mathbb{P}_\eta[\cdot] = \mathbb{P}_\eta[\cdot|\mathcal{X}_{2n}]$ and $\mathbb{E}_\eta[\cdot] = \mathbb{E}_\eta[\cdot|\mathcal{X}_{2n}]$ respectively. Then by Markov's inequality

$$\begin{aligned}
\mathbb{P}_\eta\left(\widehat{\mathcal{T}}' \geq t_\alpha^*\right) &= \mathbb{P}_\eta\left(\frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}_\eta(X_i)-\widehat{\pi}_1)^2 \geq t_\alpha^*\right) \\
&\leq \frac{1}{t_\alpha^* n}\sum_{i=n+1}^{2n}\mathbb{E}_\eta\left[(\widehat{m}_\eta(X_i)-\widehat{\pi}_1)^2\right],
\end{aligned}$$

where $\widehat{m}_\eta(x) = \sum_{i=1}^n w_i(x)Y_{\eta_i}$. Since $\sum_{i=1}^n w_i(x) = 1$ for any $x \in S$,

$$\mathbb{E}_\eta\left[\widehat{m}_\eta(x)\right] = \sum_{i=1}^n w_i(x)\mathbb{E}_\eta[Y_{\eta_i}] = \sum_{i=1}^n w_i(x)\widehat{\pi}_1 = \widehat{\pi}_1.$$

Further note that

$$\begin{aligned}
&\mathbb{E}_\eta\left[(\widehat{m}_\eta(x)-\widehat{\pi}_1)^2\right] = \\
&\quad \sum_{i_1=1}^n\sum_{i_2=1}^n w_{i_1}(x)w_{i_2}(x)\mathbb{E}_\eta\left[(Y_{\eta_{i_1}}-\widehat{\pi}_1)(Y_{\eta_{i_2}}-\widehat{\pi}_1)\right] \qquad (7) \\
&\quad \leq \sum_{i=1}^n w_i^2(x)\mathbb{E}_\eta\left[(Y_{\eta_i}-\widehat{\pi}_1)^2\right] \\
&\quad = \widehat{\pi}_1(1-\widehat{\pi}_1)\sum_{i=1}^n w_i^2(x) \leq \frac{1}{4}\sum_{i=1}^n w_i^2(x),
\end{aligned}$$

where the first inequality uses $\mathbb{E}_\eta\left[(Y_{\eta_{i_1}}-\widehat{\pi}_1)(Y_{\eta_{i_2}}-\widehat{\pi}_1)\right] \leq 0$ 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 (4) which holds for *i.i.d.* samples, we will associate the upper bound in (8) 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

$$\begin{aligned}
&\mathbb{E}_{Y^*}\left[(\widehat{m}(x)-1/2)^2|X_1,\ldots,X_{2n}\right] \\
=\ & \mathbb{E}_{Y^*}\Big[\big(\sum_{i=1}^n w_i(x)Y_i^* - 1/2\big)^2\big|X_1,\ldots,X_{2n}\Big] \\
=\ & \mathbb{E}_{Y^*}\Big[\big(\sum_{i=1}^n w_i(x)(Y_i^* - 1/2)\big)^2\big|X_1,\ldots,X_{2n}\Big] \\
=\ & \sum_{i_1=1}^n\sum_{i_2=1}^n w_{i_1}(x)w_{i_2}(x)\mathbb{E}_{Y^*}[(Y_{i_1}^*-1/2)(Y_{i_2}^*-1/2)] \\
=\ & \frac{1}{4}\sum_{i=1}^n w_i^2(x).
\end{aligned}$$

Therefore, we obtain

$$\mathbb{E}_\eta\left[(\widehat{m}_\eta(x)-\widehat{\pi}_1)^2\right] \leq \mathbb{E}_{Y^*}\left[(\widehat{m}(x)-1/2)^2|X_1,\ldots,X_{2n}\right]$$

which in turn implies that

$$\begin{aligned}
&\mathbb{P}_\eta\left(\widehat{\mathcal{T}}' \geq t_\alpha^*\right) \leq \\
&\leq \frac{1}{t_\alpha^* n}\sum_{i=n+1}^{2n}\mathbb{E}_{Y^*}\left[(\widehat{m}(X_i)-1/2)^2|X_1,\ldots,X_{2n}\right].
\end{aligned}$$

So the critical value of the permutation distribution is bounded by

$$t_\alpha^* \leq \frac{1}{\alpha n}\sum_{i=n+1}^{2n}\mathbb{E}_{Y^*}\left[(\widehat{m}(X_i)-1/2)^2|X_1,\ldots,X_{2n}\right].$$

Next, define the event

$$\mathcal{A} = \left\{ \frac{1}{n} \sum_{i=n+1}^{2n} \mathbb{E}_{Y^*} \left[(\widehat{m}(X_i) - 1/2)^2 | X_1, \ldots, X_{2n} \right] \le C_2' \delta_n \right\}. \tag{8}$$

Now, because we assume that

$$\sup_{m \in \mathcal{M}} \mathbb{E} \int_S \left(\widehat{m}(x) - m(x)\right)^2 dP_X(x) \le C_0 \delta_n, \tag{9}$$

by Markov's inequality it holds that

$$\begin{aligned} &\mathbb{P}\left(\mathcal{A}^c\right) \\ &\le \mathbb{P}\Big(\frac{1}{n} \sum_{i=n+1}^{2n} \mathbb{E}_{Y^*} \left[(\widehat{m}(X_i) - 1/2)^2 | X_1, \ldots, X_{2n} \right] \\ &\qquad > C_2' \delta_n \Big) \le \frac{C_0}{C_2'}. \end{aligned}$$

As a result, the type II error of the permutation test is bounded by

$$\begin{aligned} \mathbb{P}_1\left(\widehat{\mathcal{T}}' < t^*_\alpha\right) &\le \ \mathbb{P}_1\left(\widehat{\mathcal{T}}' < t^*_\alpha, \mathcal{A}\right) + \mathbb{P}_1\left(\mathcal{A}^c\right) \\ &\le \ \ \mathbb{P}_1\left(\widehat{\mathcal{T}}' < \frac{C_2'}{\alpha}\delta_n\right) + \frac{C_0}{C_2'}. \end{aligned}$$

Now we choose C_2' sufficiently large so that

$$\frac{C_0}{C_2'} < \frac{\beta}{2}.$$

Next we follow the proof of Lemma 2 to show that

$$\mathbb{P}_1\left(\widehat{\mathcal{T}}' < \frac{C_2'}{\alpha}\delta_n\right) < \frac{\beta}{2},$$

which completes the proof. □

D Goodness-of-Fit Regression Test via Monte Carlo Sampling

If the total number of test simulations from $\mathcal{L}(\mathbf{x};\theta_0)$ is small, 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, where we draw several independent Monte Carlo (MC) samples of size n_e from $\widehat{\mathcal{L}}(\mathbf{x};\theta_0)$ to produce a set of values $\{\widehat{T}^{(m)}\}_{m=1}^M$ that are used as a null distribution to test the hypothesis $\mathcal{L}(\mathbf{x};\theta_0) = \widehat{\mathcal{L}}(\mathbf{x};\theta_0)$. (See Algorithm 5 for details; here $f(\mathbf{x})$ denotes the likelihood $\mathcal{L}(\mathbf{x};\theta_0)$ of the simulator at $\theta = \theta_0$, and $f_e(\mathbf{x})$ denotes the approximate likelihood $\widehat{\mathcal{L}}(\mathbf{x};\theta_0)$ of the emulator at the same parameter value.) If the emulations are cheap, we can choose $n_e \gg n_{\text{sim}}$ as well as a large number M. To cite Friedman (Friedman, 2004, Section IV), the goodness-of-fit approach has "the potential for increased power [compared to two-sample testing] 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 repeated MC sampling scheme. To simplify the proof, we again use sample splitting for fitting the regression versus computing the test statistic.

Corollary 2. *Suppose that the regression estimator* $\widehat{m}(\cdot)$ *satisfies*

$$\sup_{m \in \mathcal{M}} \mathbb{E} \int_{\mathcal{X}} (\widehat{m}(\mathbf{x}) - m(\mathbf{x}))^2 dP_X(\mathbf{x}) \le C_0 \delta_n, \tag{10}$$

where C_0 *is a positive constant,* $\delta_n = o(1)$*,* $\delta_n \ge n^{-1}$ *and* $\mathcal{M}$ *is a class of regression* $m(\mathbf{x})$ *containing constant functions. Given* M *such that* $\alpha > (M+1)^{-1}$*, let us define the test via Monte Carlo sampling by*

$$\phi_{\text{MC}} = I \left\{ \frac{1}{M+1} \left(1 + \sum_{i=1}^M I(\widehat{\mathcal{T}}^{(i)}_{\text{split}} > \widehat{\mathcal{T}}_{\text{split}}) \right) \le \alpha \right\}.$$

Then for fixed $\alpha \in (0,1)$ *and* $\beta \in (1-\alpha)$ *and sufficiently large* n_{sim} *and* n_e*, there exists a constant* C_1 *such that*

$$\textit{Type I error: } \mathbb{P}_0(\phi_{\text{MC}} = 1) \le \alpha,$$

$$\textit{Type II error: } \sup_{m \in \mathcal{M}(C_1 \delta_n)} \mathbb{P}_1(\phi_{\text{MC}} = 0) \le \beta,$$

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

Remark. Here in contrast to the permutation approach, we do not assume that the regression is a linear smoother.

D.1 Proof of Corollary 2

We first prove the type I error control and then turn to the type II error control.

- **Type I error.**

With slight abuse of notation, let us write

$$\phi_{\text{MC}}(\mathcal{T}) = I \left\{ \frac{1}{M+1} \left(1 + \sum_{i=1}^M I(\widehat{\mathcal{T}}^{(i)}_{\text{split}} > \mathcal{T}) \right) \le \alpha \right\},$$

so that $\phi_{\text{MC}}(\widehat{\mathcal{T}}_{\text{split}}) = \phi_{\text{MC}}$. By construction, it can be checked that

$$\frac{1}{M} \sum_{i=1}^M \phi_{\text{MC}}(\widehat{\mathcal{T}}^{(i)}_{\text{split}}) \le \alpha.$$

Furthermore we know that $\widehat{\mathcal{T}}_{\text{split}}$ is equal in distribution to $\widehat{\mathcal{T}}^{(i)}_{\text{split}}$ for any $i = 1, \ldots, M$ under the null hypothesis. Thus

$$\frac{1}{M}\sum_{i=1}^{M} \mathbb{E}_0[\phi_{\text{MC}}(\widehat{\mathcal{T}}^{(i)}_{\text{split}})] = \mathbb{E}_0[\phi_{\text{MC}}] \leq \alpha,$$

which verifies the type I error control.

- **Type II error.**

For this part of the proof, we closely follow the proof of Theorem 2.2 in Kim et al. (2018). We let denote the empirical distribution of Monte Carlo samples $\widehat{\mathcal{T}}^{(1)}, \ldots, \widehat{\mathcal{T}}^{(M)}$ by

$$F_M(t) = \frac{1}{M}\sum_{i=1}^{M} I(\widehat{\mathcal{T}}^{(i)} \leq t) \quad \text{for all } t \in \mathbb{R}.$$

Then, by letting $\alpha_M = \alpha(M+1)/M - 1/M$, we can see that $\phi_{\text{MC}} = 1$ if and only if $F_M(t) \geq 1 - \alpha_M$. In other words, we reject the null hypothesis if and only if

$$\widehat{\mathcal{T}}_{\text{split}} \geq c_{1-\alpha_M},$$

where $c_{1-\alpha_M}$ is the upper $1-\alpha_M$ quantile of F_M. One can obtain an upper bound for this quantile by applying Markov's inequality as

$$c_{1-\alpha_M} \leq \frac{1}{\alpha_M}\left(\frac{1}{M}\sum_{i=1}^{M} \widehat{\mathcal{T}}^{(i)}_{\text{split}}\right).$$

Having this observation in mind and putting $\Delta_n = \mathbb{E}[(m(\mathbf{X}) - \pi_1)^2]$, let us define the events $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ such that

$$\mathcal{A}_1 = \left\{\frac{1}{M}\sum_{i=1}^{M} \widehat{\mathcal{T}}^{(i)}_{\text{split}} \leq 3\beta^{-1}C_0\delta_n\right\},$$

$$\mathcal{A}_2 = \left\{\frac{1}{n}\sum_{i=n+1}^{2n} (\widehat{m}(\mathbf{X}_i) - m(\mathbf{X}_i))^2 \leq 3\beta^{-1}C_0\delta_n\right\} \quad \text{and}$$

$$\mathcal{A}_3 = \left\{\left|\frac{1}{n}\sum_{i=n+1}^{2n} (m(\mathbf{X}_i) - \pi_1)^2 - \Delta_n\right| \leq \Delta_n/2\right\}.$$

Then applying Markov's inequality together with condition (10) yields $\mathbb{P}(\mathcal{A}_1^c) \leq \beta/3$ and $\mathbb{P}(\mathcal{A}_2^c) \leq \beta/3$. Moreover, as shown in Kim et al. (2018), we have $\mathbb{P}(\mathcal{A}_3^c) \leq 4/(C_1 n\delta_n)$. Combining these via the union bound, we see that the type II error is bounded by

$$\begin{aligned}
&\mathbb{P}\left(\widehat{\mathcal{T}}_{\text{split}} < c_{1-\alpha_M}\right) \\
&= \mathbb{P}\left(\widehat{\mathcal{T}}_{\text{split}} < c_{1-\alpha_M},\ \mathcal{A}_1\right) + \mathbb{P}\left(\widehat{\mathcal{T}}_{\text{split}} < c_{1-\alpha_M},\ \mathcal{A}_1^c\right) \\
&\leq \mathbb{P}\left(\widehat{\mathcal{T}}_{\text{split}} < 3\alpha_M^{-1}\beta^{-1}C_0\delta_n\right) + \mathbb{P}(\mathcal{A}_1^c) \\
&\leq \mathbb{P}\left(\widehat{\mathcal{T}}_{\text{split}} < 3\alpha_M^{-1}\beta^{-1}C_0\delta_n\right) + \frac{\beta}{3}.
\end{aligned}$$

For the last line, based on the inequality $(x-y)^2 \leq 2(x-z)^2 + 2(z-y)^2$, we further see that

$$\begin{aligned}
&\mathbb{P}\left(\widehat{\mathcal{T}}_{\text{split}} < 3\alpha_M^{-1}\beta^{-1}C_0\delta_n\right) \\
\leq\ &\mathbb{P}\Bigg(\Bigg(\frac{1}{2n}\sum_{i=n+1}^{2n} (m(\mathbf{X}_i) - \pi_1)^2 \\
&-\frac{1}{n}\sum_{i=n+1}^{n} (\widehat{m}(\mathbf{X}_i) - m(\mathbf{X}_i))^2\Bigg) < 3\alpha_M^{-1}\beta^{-1}C_0\delta_n, \\
&\mathcal{A}_2 \cap \mathcal{A}_3\Bigg) + \mathbb{P}\left(\mathcal{A}_2^c \cup \mathcal{A}_3^c\right) \\
\leq\ &\mathbb{P}\left(\Delta_n < 6(1+\alpha_M^{-1})\beta^{-1}C_0\delta_n\right) + \frac{\beta}{3} + \frac{4}{C_1 n\delta_n}.
\end{aligned}$$

Then by taking C_1 sufficiently large, the proof is complete.

E Example 1 (Consistency of Global Test)

In Example 1, we tested the null hypothesis that $\widehat{\mathcal{L}}(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta)$ for data simulated according to $\theta \sim Gamma(1,1)$, and $\mathbf{x} = x_1, ..., x_{1000}|\theta \sim Beta(\theta,\theta)$. Figure 9 (left) shows the true likelihood $\mathcal{L}(\mathbf{x};\theta)$ for some different values of θ but a fixed $\mathbf{x}$ (for simplicity), comparing these functions to the likelihood approximation $\widehat{\mathcal{L}}(\mathbf{x};\theta) \propto 1$. Such an approximation is valid when $\theta = 1$, as $Beta(1,1)$ is indeed just the uniform distribution, whereas the approximation is clearly wrong for the other values of $\theta \sim Gamma(1,1)$.

F Example 2 (Power of Two-Sample Test via Regression)

The practical implications of Theorem 2 are that for a two-sample test via regression one should base the test on the regression method with the smallest mean integrated squared error (MISE) so as to achieve a more powerful test. Table 2 illustrates this for the three settings in Example 2 (Section 2.3): random

Algorithm 5 Goodness-of-Fit Regression Test via Monte Carlo Sampling

```
Input: i.i.d. sample $\mathcal{S}$ of size $n_{\text{sim}}$ from distribution with density $f$; emulator model with density $f_e$; size of Monte Carlo sample $n_e$; number of additional Monte Carlo samples $M$; a regression method $\widehat{m}$
Output: $p$-value for testing if $f(\mathbf{x}) = f_e(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$
Let $n = n_{\text{sim}} + n_e$.
Sample $\mathcal{S}_e = \{\mathbf{X}_1^*, \dots, \mathbf{X}_{n_e}^*\}$ from $f_e$.
Define an augmented sample $\{\mathbf{X}_i, Y_i\}_{i=1}^n$, where $\{\mathbf{X}_i\}_{i=1}^n = \mathcal{S} \cup \mathcal{S}_e$, and $Y_i = I(\mathbf{X}_i \in \mathcal{S}_e)$.
Calculate the test statistic $\widehat{\mathcal{T}}$ in Equation 1.
for $m \in \{1, \dots, M\}$ do
    Sample $\mathcal{S}^{(m)} = \{\mathbf{X}_1^{(m)}, \dots, \mathbf{X}_{n_{\text{sim}}}^{(m)}\}$ from $f$, under the null hypothesis $H_0 : f = f_e$.
    Sample $\mathcal{S}_e^{(m)} = \{\mathbf{X}_1^{*(m)}, \dots, \mathbf{X}_{n_e}^{*(m)}\}$ from $f_e$.
    Define a new augmented sample $\{\mathbf{X}_i, Y_i\}_{i=1}^n$, where $\{\mathbf{X}_i\}_{i=1}^n = \mathcal{S}^{(m)} \cup \mathcal{S}_e^{(m)}$, and $Y_i = I(\mathbf{X}_i \in \mathcal{S}_e^{(m)})$.
    Refit $\widehat{m}$ and calculate the test statistic on the new augmented sample to obtain $\widehat{\mathcal{T}}^{(m)}$ from the null distribution $f = f_e$.
end for
Compute the Monte Carlo $p$-value by $p = \frac{1}{M+1}\left(1 + \sum_{m=1}^{M} I(\widehat{\mathcal{T}}^{(m)} > \widehat{\mathcal{T}})\right)$.
return $p$
```

forest achieves a smaller MISE than nearest neighbor (NN) regression across all settings and, as Figure 2 shows, it also consistently attains a higher power.

Setting / Regression Method	*Random Forest*	*NN*
(a) Bernoulli	0.19	0.73
(b) Scaling	0.35	2.31
(c) Mixture of Gaussians	0.27	1.64

Table 2: Integrated mean squared error (MISE) for regression methods used for two-sample testing in Figure 2. Random forest has the smallest MISE in regression; it also yields the test with highest power, as implied by Theorem 2.

As pointed out in the related work section, classifier two-sample testing methods have also been used for two-sample testing by dichotomizing the regression function and using the classification accuracy as a test statistic. Such dichotomization might result in a loss of power with respect to the respective regression test in certain settings (for more examples, see Kim et al. (2018)). In Figure 8 we consider the same settings as in Example 2, but now also computing the power of the classification accuracy test from Lopez-Paz and Oquab (2017) for both random forest and nearest neighbor classification. The regression test achieves comparable results across the different settings, providing slight improvements in some cases, e.g., with respect to the local power at $D = 100$ (left column). Note that our global procedure can incorporate classification accuracy tests as well, but would then not be able to identify locally significant differences in feature space as in Section A.

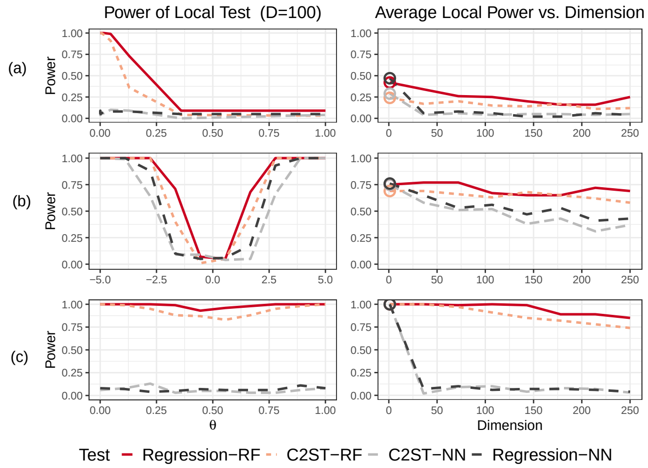

Figure 8: Test power at $D = 100$ (left column) and as a function of dimension D (right column) in the same Example 2 settings, i.e., for (a) Bernoulli, (b) Scaling and (c) Mixture of Gaussians. We include the results for our regression test with random forests (RF) and nearest neighbors (NN), as well as the corresponding results using the classification accuracy test of Lopez-Paz and Oquab (2017) with RF and NN (labeled as C2ST-RF and C2ST-NN, respectively).

G Approximate P-Values and Confidence Regions

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

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\mathbf{x}; \theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\mathbf{x}; \theta)}.$$

We estimate $\lambda(\mathbf{x})$ using the estimated likelihood:

$$\widehat{\lambda}(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_0} \widehat{\mathcal{L}}(\mathbf{x}; \theta)}{\sup_{\theta \in \Theta} \widehat{\mathcal{L}}(\mathbf{x}; \theta)}.$$

The estimated p-value is then

$$\widehat{p}(\mathbf{x}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\widehat{\lambda}(\mathbf{X}) > \widehat{\lambda}(\mathbf{x}))$$

If $\Theta_0 = \{\theta_0\}$, $\widehat{p}(\mathbf{x})$ can be estimated using data that are simulated under $\theta = \theta_0$. If $|\Theta_0| > 1$, the distribution of the test statistic can be approximated using the χ^2 approximation for the likelihood ratio test (Casella and Berger, 2002). Confidence intervals may be obtained by inverting the hypothesis tests (Casella and Berger, 2002).

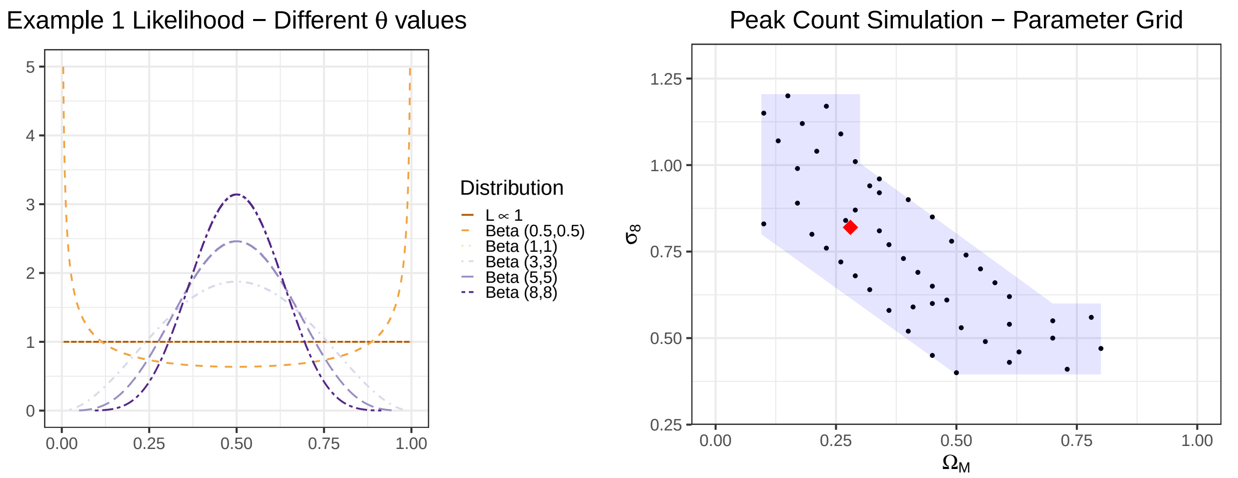

Figure 9: *Left*: The true likelihood for different values of the parameter θ, compared to the approximation $\widehat{\mathcal{L}}(\mathbf{x};\theta) \propto 1$. The approximation is clearly wrong when $\theta \neq 1$. *Right*: Location of the 50 parameter settings for the peak count data simulations using `CAMELUS`, where the blue region indicates the parameter range values and the red diamond indicates the fiducial point θ_0.

H Peak Count Data Example

The KL divergence for model comparison is estimated by:

$$\begin{aligned}
KL(\mathcal{L}, \widehat{\mathcal{L}}) &= -\mathbb{E}\left[\log\left(\frac{\widehat{\mathcal{L}}(\mathbf{x};\theta)}{\mathcal{L}(\mathbf{x};\theta)}\right)\right] \\
&= -\mathbb{E}\left[\log\left(\widehat{\mathcal{L}}(\mathbf{x};\theta)\right)\right] + K \\
&\approx -\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j}\log\left(\widehat{\mathcal{L}}(\mathbf{x}_{ij};\theta_j)\right) + K
\end{aligned}$$

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

Figure 9, right, shows the grid of 50 parameters settings $\theta = (\Omega_m, \sigma_8)$ which we use for the `CAMELUS` batch simulations. The blue shaded region represents the parameter regions from which the parameters are sampled around the fiducial cosmology θ_0 (indicated by a red diamond).

For the conditional MAF, at both $n_{\text{train}} = 200$ and $n_{\text{train}} = 500$ we used 10% of the training data as validation. During training we assessed validation loss and we stopped the training early if the validation loss was not improving for 30 epochs. We explored architectures with $\{5, 10, 15, 20\}$ autoregressive layers and $2^{\{4,..,10\}}$ hidden units, with the best performing having 10 autoregressive layers and either 512 or 1024 hidden units.