Validation of Emulators and Approximate Likelihood Models

Ann B. Lee
Department of Statistics & Data Science
Carnegie Mellon University

Joint work with Taylor Pospisil (CMU), Rafael Izbicki (UCSCar), and Niccolo Dalmasso (CMU)

STAMPS: Statistical Methods for the Physical Sciences

STAMPS, New focus group at CMU (2018-)

Potential application areas:

- Oceanography
- Particle physics
- Remote sensing
- Atmospheric science
- Environmental science
- Astrophysics
- Cosmology
- ...

Statistical commonalities:

- Spatio-temporal data
- Ill-posed inverse problems
- Uncertainty quantification
- Computationally intensive simulations
- Massive datasets
- ...

"We want to extract as much information as possible from our data—but hopefully not more than that."

Louis Lyons

My Research Interests

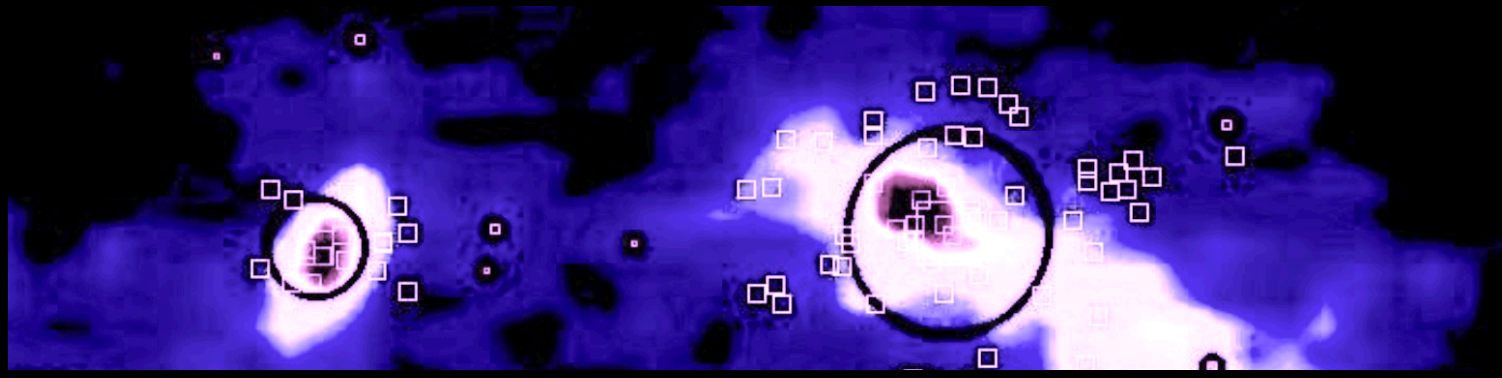
- Leverage ML/prediction tools to solve classical statistical problems (e.g. uncertainty quantification, and hypothesis testing) in non-traditional settings with different types of complex data $\mathbf{x}$
- In the context of LFI, I've been working on
  - direct estimation of approximate likelihoods $f(\mathbf{x}|\theta)$ [Izbicki/Lee/Schafer, 2014] and posteriors $f(\theta|\mathbf{x})$ [Izbicki/Lee/Pospisil, 2017] via "ABC-CDE" (Delphi?)
  - validation of emulators and approximate likelihoods $f(\theta|\mathbf{x})$ [Pospisil/Lee/Izbicki/Dalmasso, 2018]

My Research Interests

- Leverage ML/prediction tools to solve classical statistical problems (e.g. uncertainty quantification, and hypothesis testing) in non-traditional settings with different types of complex data **x**
- In the context of LFI, I've been working on
  - direct estimation of approximate likelihoods $f(\mathbf{x}|\theta)$ [Izbicki/Lee/Schafer, 2014] and posteriors $f(\theta|\mathbf{x})$ [Izbicki/Lee/Pospisil, 2017] via "ABC-CDE" (Delphi?)
  - validation of emulators and approximate likelihoods $f(\theta|\mathbf{x})$ [Pospisil/Lee/Izbicki/Dalmasso, 2018]

Setting: Extremely Slow Simulations

- Setting with extremely computationally intensive simulations (simulated in batches), where one needs to fit an emulator to the slow simulations. Batch setting common in physics, e.g.:
  - In high-precision cosmology analysis, standard to run a batch of ~1000 N-body or hydrodynamic simulations at a fixed cosmology (can take months with thousands of CPU's)



Scinet LIghtCone Simulations [Credit: Joachim Harnois-Derap]

Validation of Emulators: Objectives

- When fitting an emulator or approximate likelihood model $L(x;\theta)=p(x|\theta)$ to slow simulations, we want to answer:
  - IF one needs to run more computationally intensive simulations to better fit an emulator to the simulations
  - WHERE in parameter space one, if needed, should propose the next batch of simulations
  - HOW the emulated and high-resolution simulated data are different in observable space X (to provide valuable information toward improving the emulator)

Validation of Emulators: Open Problems

- Typically, validation is done via ML loss functions or distance measures, or goodness-of-fit tests based on posterior quantiles or rank statistics
- To date, there are no diagnostics or validation techniques in the emulator/LFI literature that are fully consistent (that is, that can distinguish any bad estimator from the ``true'' reference likelihood), and that in addition can answer the if/where/how questions above.

$$\begin{aligned} \text{Test} \quad & H_0 : \widehat{\mathcal{L}}(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta) \text{ for every } \mathbf{x} \in \mathcal{X} \text{ and } \theta \in \Theta \\ \text{versus} \quad & H_1 : \widehat{\mathcal{L}}(\mathbf{x};\theta) \neq \mathcal{L}(\mathbf{x};\theta) \text{ for some } \mathbf{x} \in \mathcal{X} \text{ and } \theta \in \Theta \end{aligned}$$

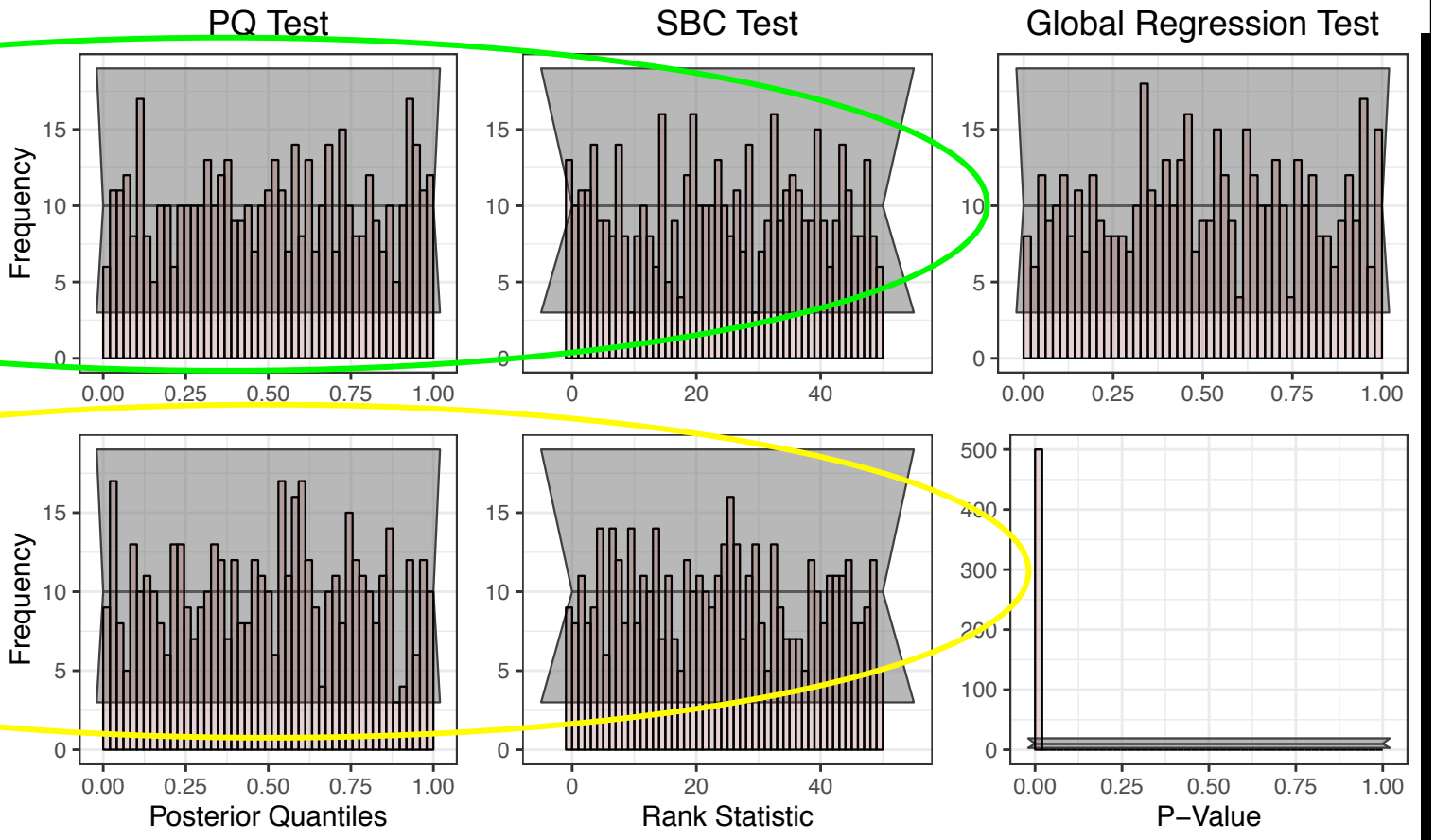
Common Diagnostics (PQ, SBC) Can Sometimes Not Distinguish Between a Good Model (top row) and a Clearly Misspecified Model (bottom row)

For each $\theta_i \sim Gamma(1,1)$, we sample $X_1, \ldots, X_n | \theta_i \sim Beta(\theta_i, \theta_i)$.

Compare Posterior Quantile (PQ) test [Cook et al 2006], Simulation-Based Calibration (SBC) test [Talts et al 2018], and our proposed global test.

$\widehat{\mathcal{L}}(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta)$ (a)

$\widehat{\mathcal{L}}(\mathbf{x};\theta) \propto 1$ (b)



Our Proposed Approach: Perform a Local Test at Each Parameter and then Test if the Local p-values has a Uniform Distribution [Pospisil et al, 2019]

Algorithm 1 Local Test

Input: parameter value θ_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_{θ_0} for testing if $L(\mathbf{x};\theta_0)=\widehat{L}(\mathbf{x};\theta_0)$ for every $\mathbf{x}$

1: Sample $\mathcal{S}_0 = \{\mathbf{X}_1^{\theta_0}, \ldots, \mathbf{X}_{n_{\text{sim},0}}^{\theta_0}\}$ from $\mathcal{L}(\mathbf{x};\theta_0)$.
2: Sample $\mathcal{S}_1 = \{\mathbf{X}_1^{*}, \ldots, \mathbf{X}_{n_{\text{sim},1}}^{*}\}$ from $\widehat{\mathcal{L}}(\mathbf{x};\theta_0)$.
3: Compute p-value p_{θ_0} for the comparison between $\mathcal{S}_0$ and $\mathcal{S}_1$.
4: **return** p_{θ_0}

Algorithm 3 Global Test

Input: reference distribution $r(\theta)$, B, uniform testing procedure

Output: p-value p for testing if $L(\mathbf{x};\theta) = \widehat{L}(\mathbf{x};\theta)$ for every $\mathbf{x}$ and θ

1: **for** $i \in \{1, \ldots, B\}$ **do**
2: sample $\theta_i \sim r(\theta)$
3: compute p_{θ_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

Our Proposed Approach: Perform a Local Test at Each Parameter and then Test if the Local p-values has a Uniform Distribution [Pospisil et al, 2019]

Algorithm 1 Local Test

Input: parameter value θ_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_{θ_0} for testing if $L(\mathbf{x};\theta_0)=\widehat{L}(\mathbf{x};\theta_0)$ for every $\mathbf{x}$

1: Sample $\mathcal{S}_0 = \{\mathbf{X}_1^{\theta_0}, \ldots, \mathbf{X}_{n_{\text{sim},0}}^{\theta_0}\}$ from $\mathcal{L}(\mathbf{x};\theta_0)$.
2: Sample $\mathcal{S}_1 = \{\mathbf{X}_1^{*}, \ldots, \mathbf{X}_{n_{\text{sim},1}}^{*}\}$ from $\widehat{\mathcal{L}}(\mathbf{x};\theta_0)$.
3: Compute p-value p_{θ_0} for the comparison between $\mathcal{S}_0$ and $\mathcal{S}_1$.
4: **return** p_{θ_0}

- Global test is consistent if the local test is consistent
- Challenge: To find a two-sample test that can handle complex, high-dimensional data x and that's interpretable

Algorithm 3 Global Test

Input: reference distribution $r(\theta)$, B, uniform testing procedure

Output: p-value p for testing if $L(\mathbf{x};\theta) = \widehat{L}(\mathbf{x};\theta)$ for every $\mathbf{x}$ and θ

1: **for** $i \in \{1, \ldots, B\}$ **do**
2: sample $\theta_i \sim r(\theta)$
3: compute p_{θ_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

A Two-Sample Test via Regression
[Kim, Lee and Lei, 2019]

Suppose we have two samples:

$$\mathbf{X}_1^0, \ldots, \mathbf{X}_m^0 \sim P_0 \quad \text{and} \quad \mathbf{X}_1^1, \ldots, \mathbf{X}_n^1 \sim P_1$$

A two sample-test would ask whether P_0 and P_1 are the same; i.e., it would test the null hypothesis

$$H_0 : f(\mathbf{x}|Y = 0) = f(\mathbf{x}|Y = 1) \quad \text{for all } \mathbf{x} \in \mathcal{X}$$

A Two-Sample Test via Regression

[Kim, Lee and Lei, 2019]

Suppose we have two samples:

$$\mathbf{X}_1^0, \ldots, \mathbf{X}_m^0 \sim P_0 \quad \text{and} \quad \mathbf{X}_1^1, \ldots, \mathbf{X}_n^1 \sim P_1$$

A two sample-test would ask whether P_0 and P_1 are the same; i.e., it would test the null hypothesis

$$H_0 : f(\mathbf{x}|Y = 0) = f(\mathbf{x}|Y = 1) \quad \text{for all } \mathbf{x} \in \mathcal{X}$$

By Bayes rule, this is equivalent to testing

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

Convert a Regression Into a Two-Sample Test

Our null and alternative hypotheses are

$$H_0: \ \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) \ = \ \mathbb{P}(Y=1) \ \text{ for all } \mathbf{x} \in \mathcal{X}$$
$$H_1: \ \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) \ \neq \ \mathbb{P}(Y=1) \ \text{ for some } \mathbf{x} \in \mathcal{X}$$

Define the regression function $m(\mathbf{x}) = \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})$.
Let $\widehat{m}(\mathbf{x})$ be an estimate of $m(\mathbf{x})$ based on the sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$.
Let $\widehat{\pi}_1 = \frac{1}{n}\sum_{i=1}^n I(Y_i = 1)$.

We define our test statistic as

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

Why Two-Sample Test via Regression?

$$
\begin{aligned}
H_0 &: \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) = \mathbb{P}(Y=1) \text{ for all } \mathbf{x} \in \mathcal{X} \\
H_1 &: \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) \neq \mathbb{P}(Y=1) \text{ for some } \mathbf{x} \in \mathcal{X}
\end{aligned}
$$

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

- Can adapt to any structure in X for which there is a suitable regression technique
- The power of the test is directly related to the the MISE of the chosen regression estimator [Kim et al, 2019]
-

If the chosen regression estimator has a small MISE, the power of the test is large over a wide region of the alternative hypothesis

Theorem 1. *Suppose that the regression estimator $\widehat{m}(\mathbf{x})$ is a linear smoother satisfying*

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

where C_0 is a positive constant, $\delta_n = o(1)$, $\delta_n \geq n^{-1}$, and $\mathcal{M}$ is a class of regressions $m(\mathbf{x})$ containing constant functions. Let t^*_α be the upper α quantile of the permutation distribution of the test statistic $\widehat{T}'$ on validation data.[1] *Then for any $\alpha, \beta \in (0, 1/2)$, there exists a universal constant C_1 such that*

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

against the class of alternatives $\mathcal{M}(C_1\delta_n)$ defined by

$$\left\{ m \in \mathcal{M} : \int_{\mathcal{X}} (m(\mathbf{x}) - \pi_1)^2 \, dP_X(\mathbf{x}) \geq C_1 \delta_n \right\},$$

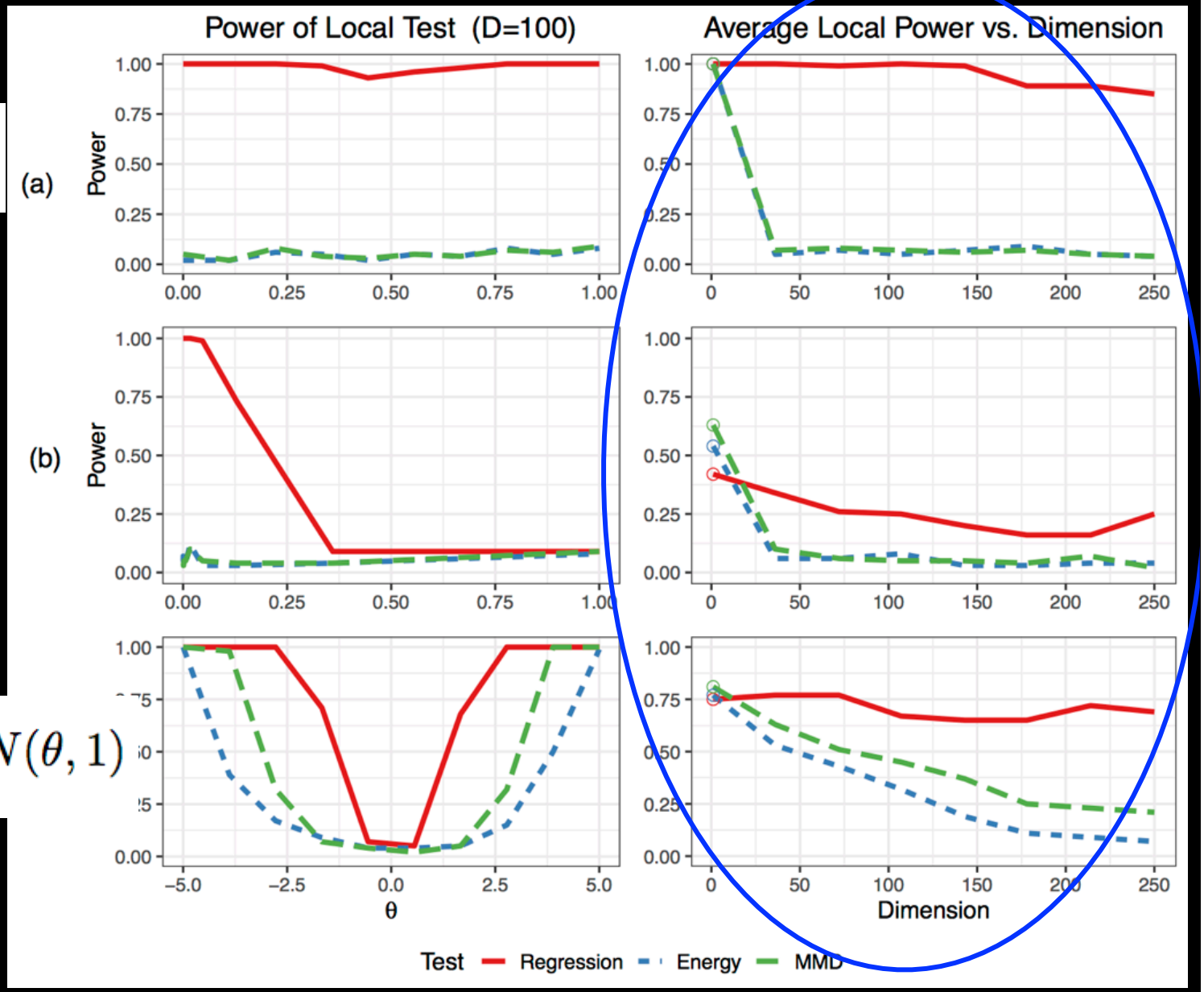
for n sufficiently large.

Power of Local Test (Sparse Structure in High Dimensions) for RF regression vs MMD

$$X_1 \sim Bernoulli(\theta)$$

$$X_1 \sim N(0, \theta)$$

$$X_1 \sim \tfrac{1}{2}N(-\theta, 1) + \tfrac{1}{2}N(\theta, 1)$$



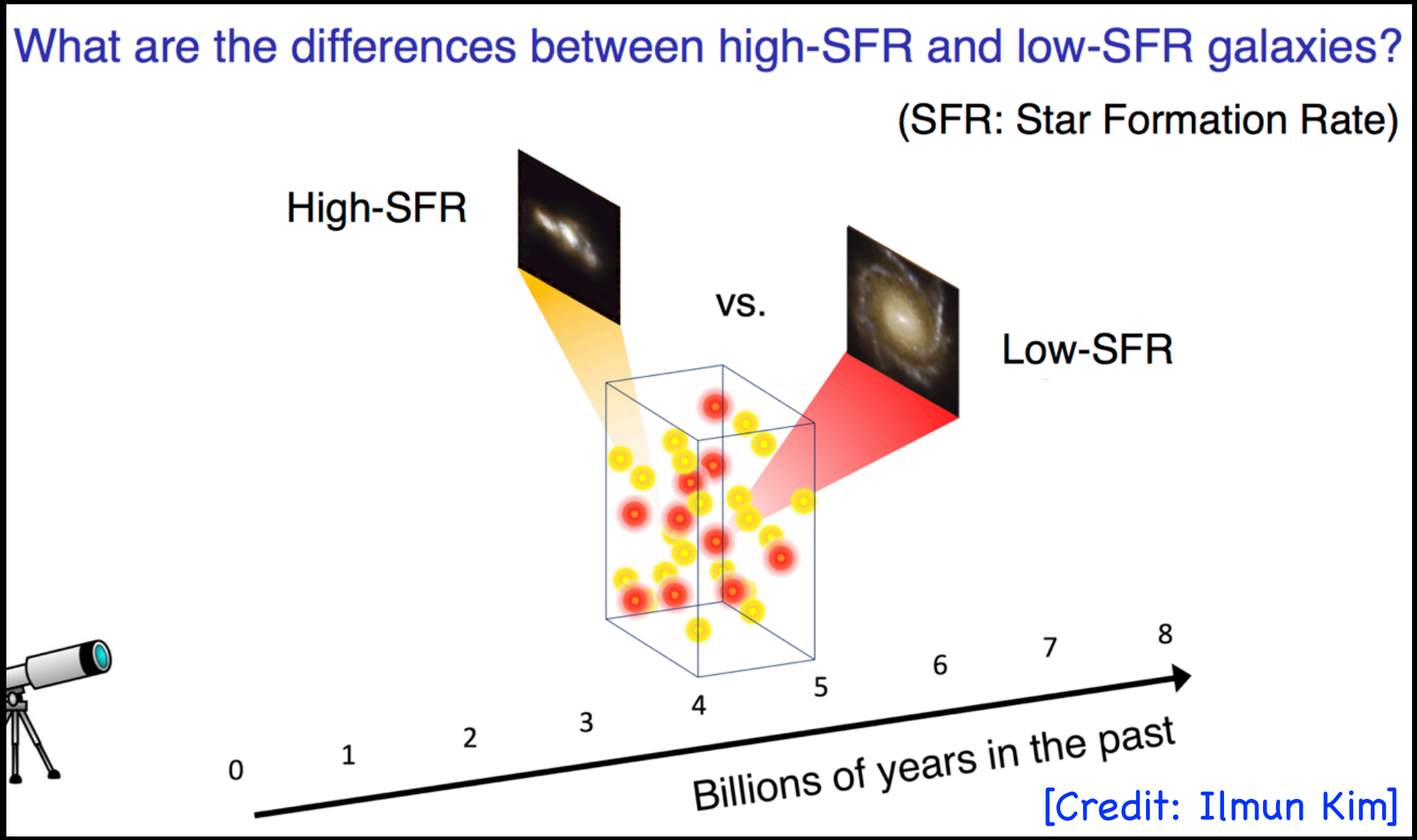
Why Two-Sample Test via Regression?

$$
\begin{aligned}
H_0 &: \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) = \mathbb{P}(Y=1) \text{ for all } \mathbf{x} \in \mathcal{X} \\
H_1 &: \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) \neq \mathbb{P}(Y=1) \text{ for some } \mathbf{x} \in \mathcal{X}
\end{aligned}
$$

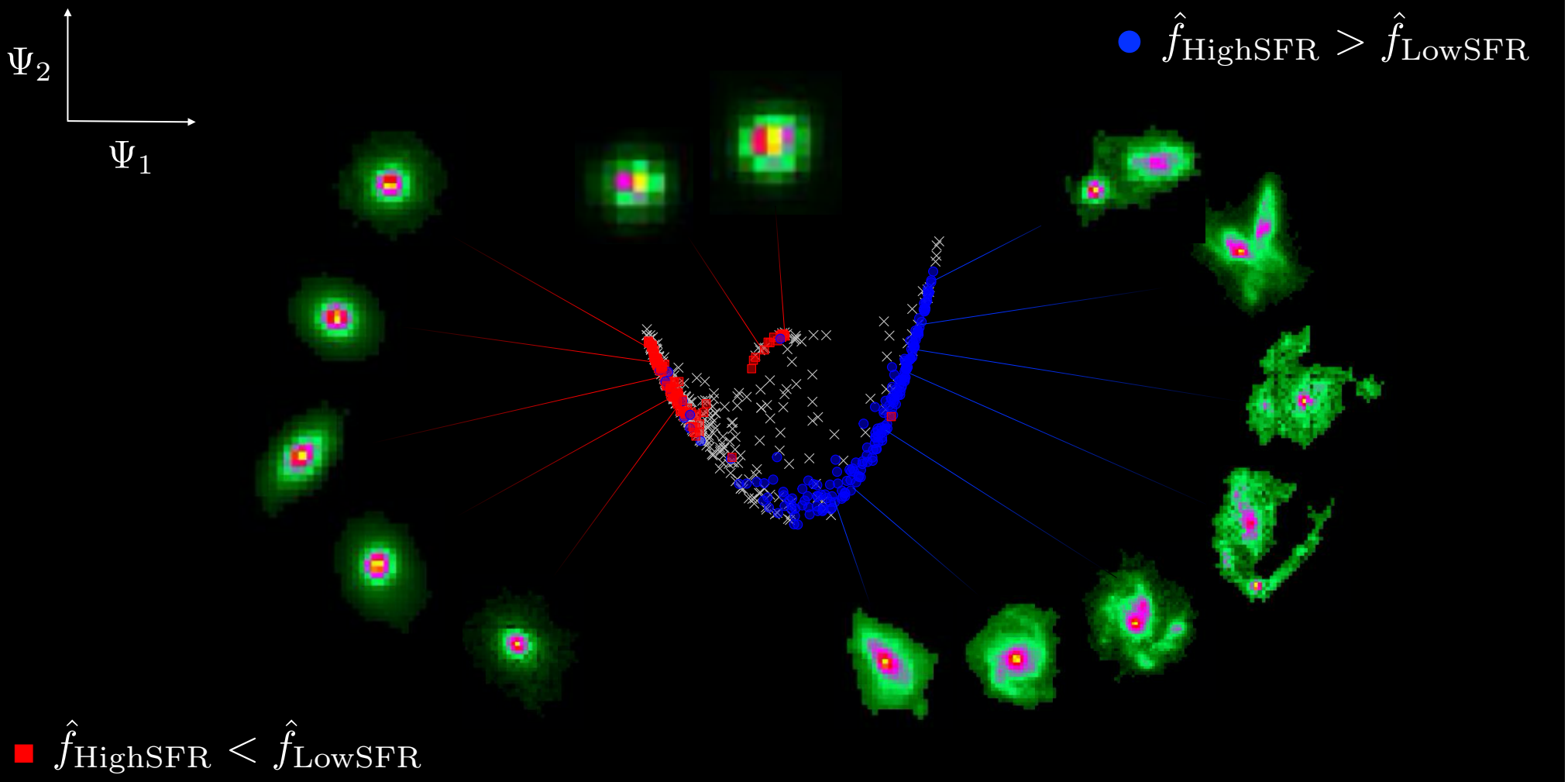
$$
\widehat{T} = \frac{1}{n}\sum_{i=1}^{n} (\widehat{m}(\mathbf{X}_i) - \widehat{\pi}_1)^2 .
$$

- Can adapt to any structure in X for which there is a suitable regression technique
- The power of the test is directly related to the the MISE of the chosen regression estimator [Kim et al, 2019]
- The regression test tells you not only if but also how the two samples are different in space of observables

Ex: Analyze How Two Samples are Different in Feature Space [Freeman/Kim/Lee, MNRAS`2017]



Ex: Regression Test to Analyze How Two Distributions (High-SFR vs Low-SFR Galaxies) Differ in High-Dimensional Space



Let's Now Return to the Problem of Validating Emulators

$$\text{Test} \quad H_0 : \widehat{\mathcal{L}}(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta) \text{ for every } \mathbf{x} \in \mathcal{X} \text{ and } \theta \in \Theta$$
$$\text{versus} \quad H_1 : \widehat{\mathcal{L}}(\mathbf{x};\theta) \neq \mathcal{L}(\mathbf{x};\theta) \text{ for some } \mathbf{x} \in \mathcal{X} \text{ and } \theta \in \Theta$$

Algorithm 1 Local Test

Input: parameter value θ_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_{θ_0} for testing if $L(\mathbf{x};\theta_0)$=$\widehat{L}(\mathbf{x};\theta_0)$ for every $\mathbf{x}$

1: Sample $\mathcal{S}_0 = \{\mathbf{X}_1^{\theta_0}, \ldots, \mathbf{X}_{n_{\text{sim},0}}^{\theta_0}\}$ from $\mathcal{L}(\mathbf{x};\theta_0)$.
2: Sample $\mathcal{S}_1 = \{\mathbf{X}_1^{*}, \ldots, \mathbf{X}_{n_{\text{sim},1}}^{*}\}$ from $\widehat{\mathcal{L}}(\mathbf{x};\theta_0)$.
3: Compute p-value p_{θ_0} for the comparison between $\mathcal{S}_0$ and $\mathcal{S}_1$.
4: **return** p_{θ_0}

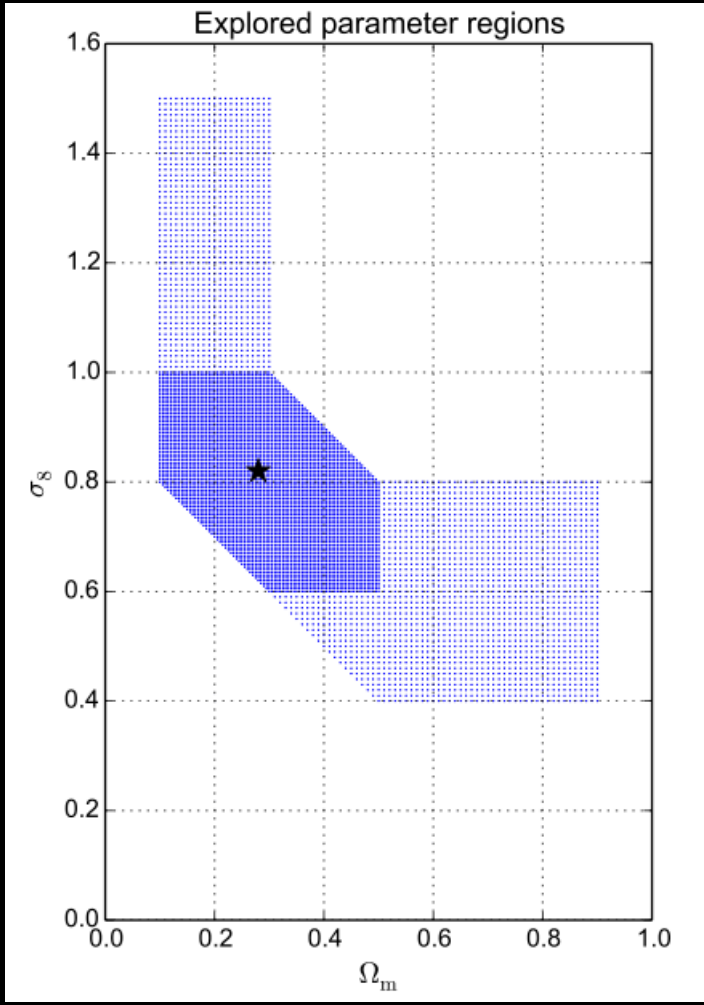
Algorithm 3 Global Test

Input: reference distribution $r(\theta)$, B, uniform testing procedure

Output: p-value p for testing if $L(\mathbf{x};\theta) = \widehat{L}(\mathbf{x};\theta)$ for every $\mathbf{x}$ and θ

1: **for** $i \in \{1, \ldots, B\}$ **do**
2: sample $\theta_i \sim r(\theta)$
3: compute p_{θ_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

Application to Weak Lensing Peak Count Data



- Use CAMELUS [Lin & Kilbinger 2015] to simulate weak lensing convergence maps => binned peak counts $x \in R^{13}$. A batch of ~1000 realizations at each grid point or setting of cosmological parameters $\theta=(\Omega_M, \sigma_8)$.
- Fit three different likelihood models: Gaussian, Poisson, KDE

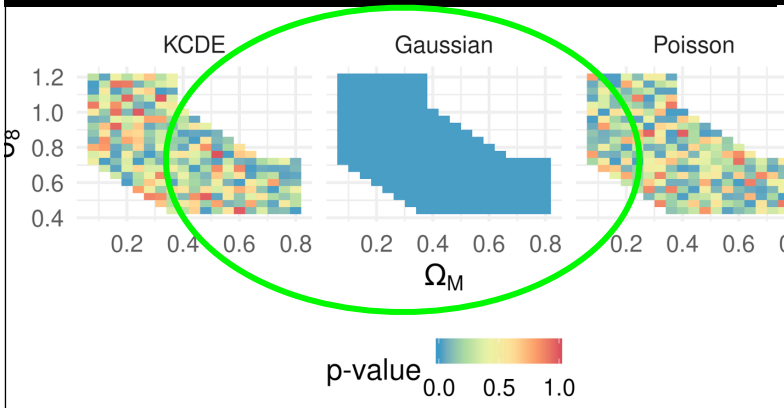
$$\text{Test} \quad H_0 : \widehat{\mathcal{L}}(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta) \text{ for every } \mathbf{x} \in \mathcal{X} \text{ and } \theta \in \Theta$$
$$\text{versus} \quad H_1 : \widehat{\mathcal{L}}(\mathbf{x};\theta) \neq \mathcal{L}(\mathbf{x};\theta) \text{ for some } \mathbf{x} \in \mathcal{X} \text{ and } \theta \in \Theta$$

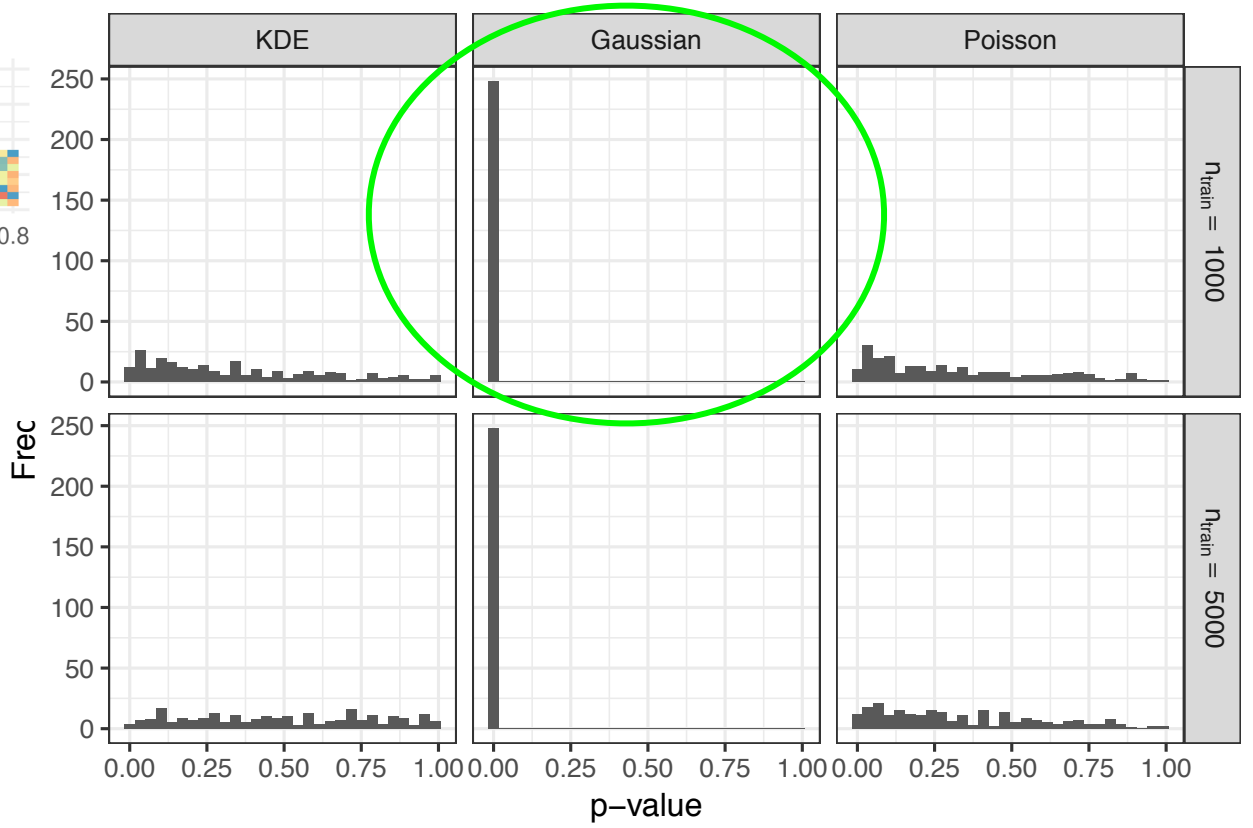
Do We Need More Simulations to Fit the Data Well or Are the Current Fits Good Enough?

- Start with $n_{train}=1000$ train and $n_{sim}=200$ test simulations at each parameter setting (i.e. grid point)
- Suppose we compute the KL-divergence between the simulated and approximate likelihoods on test data. We then get that the Gaussian model performs best with a loss of 34.96, Poisson model closely behind at 35.53, and KDE last at 114.09.
- Based on the KL loss we would choose the Gaussian model -- but these are only relative comparisons...

Do We Need More Simulations to Fit the Data Well or Are the Current Fits Good Enough?

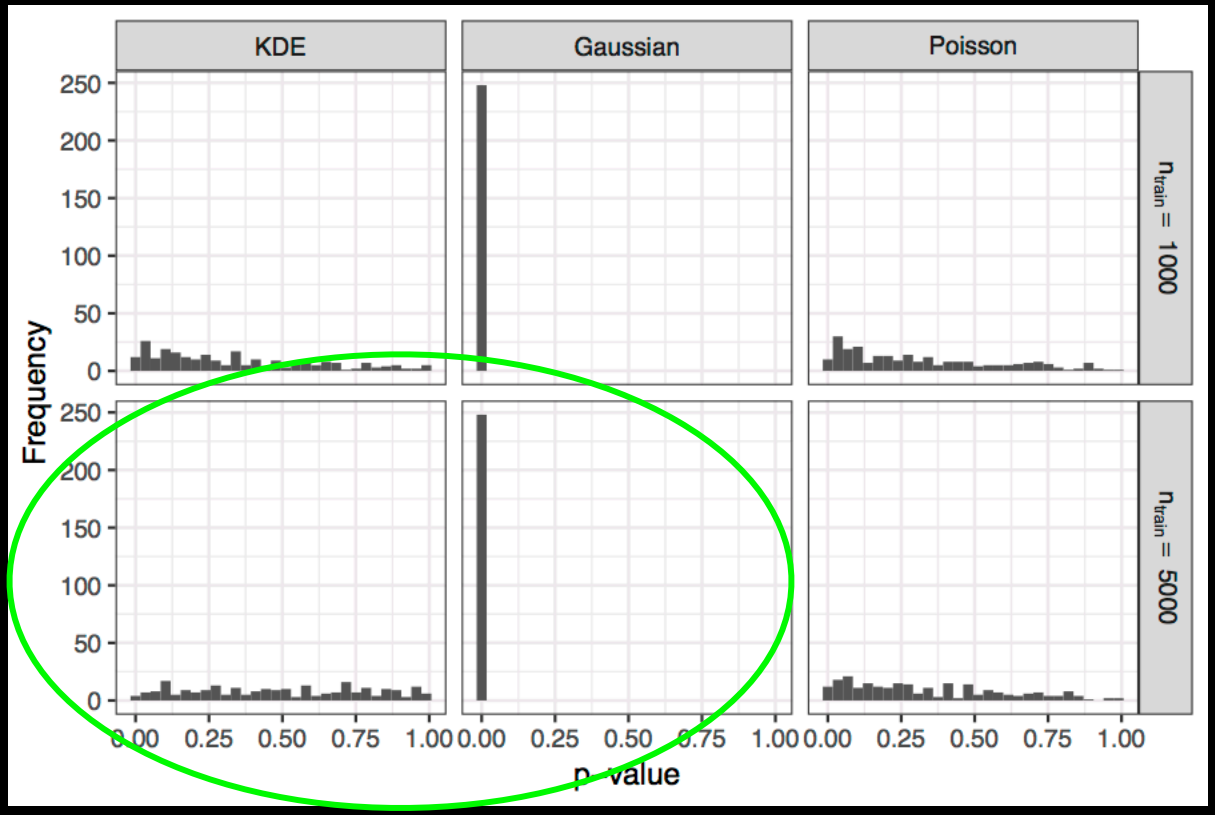
- Based on the KL loss we would choose the Gaussian likelihood model -- but our local test p-values reveal that the Gaussian model is rejected at all θ





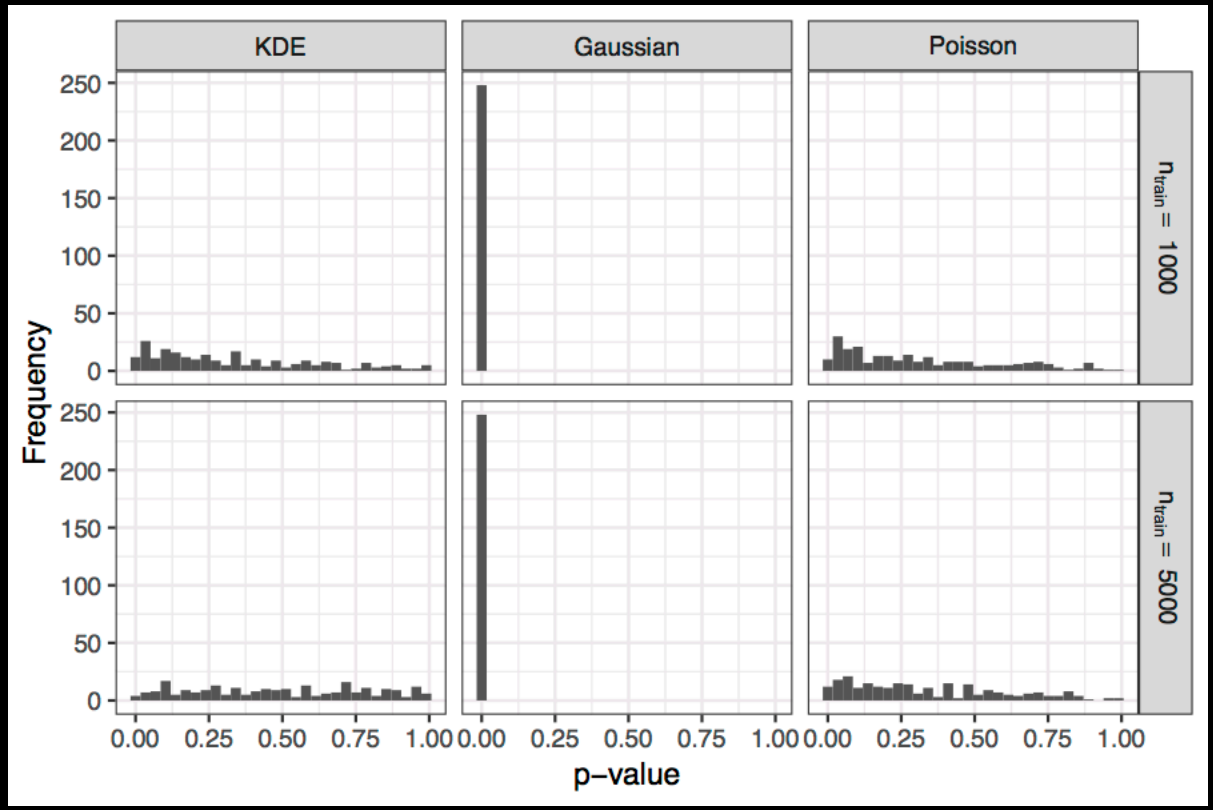
Do We Need More Simulations to Fit the Data Well or Are the Current Fits Good Enough?

- If we increase the number of train simulations, KDE (bottom left) passes our global validation test, but the Gaussian model (bottom center) still fails the test



Do We Need More Simulations to Fit the Data Well or Are the Current Fits Good Enough?

- If we increase the number of train simulations, KDE (bottom left) passes our global validation test, but the Gaussian model (bottom center) still fails the test



- Even if it's not feasible to generate new simulations, our regression local test still provides valuable information on how to improve the emulator (diagnostics)...

But Let Us First Look at a Synthetic Example where L(x;θ) is Known…

- Construct an example with two salient features:
  - The data (X_1, X_2) themselves are discrete counts: at high bin counts approximately normally distributed, but the Gaussian model breaks down at low bin counts

$$X_1, X_2 \sim Poisson(\lambda),$$
$$\text{where } \lambda = \begin{cases} 1, & \text{if } \theta_1 < 0.5 \\ 10000, & \text{otherwise} \end{cases}$$

  - Counts in different bins are (negatively) correlated ⇒ the independent Poisson model breaks down

$$\text{When } \theta_2 < 0.5, \text{ add requirement that } X_1 \leq X_2$$

Global Goodness-of-Fit

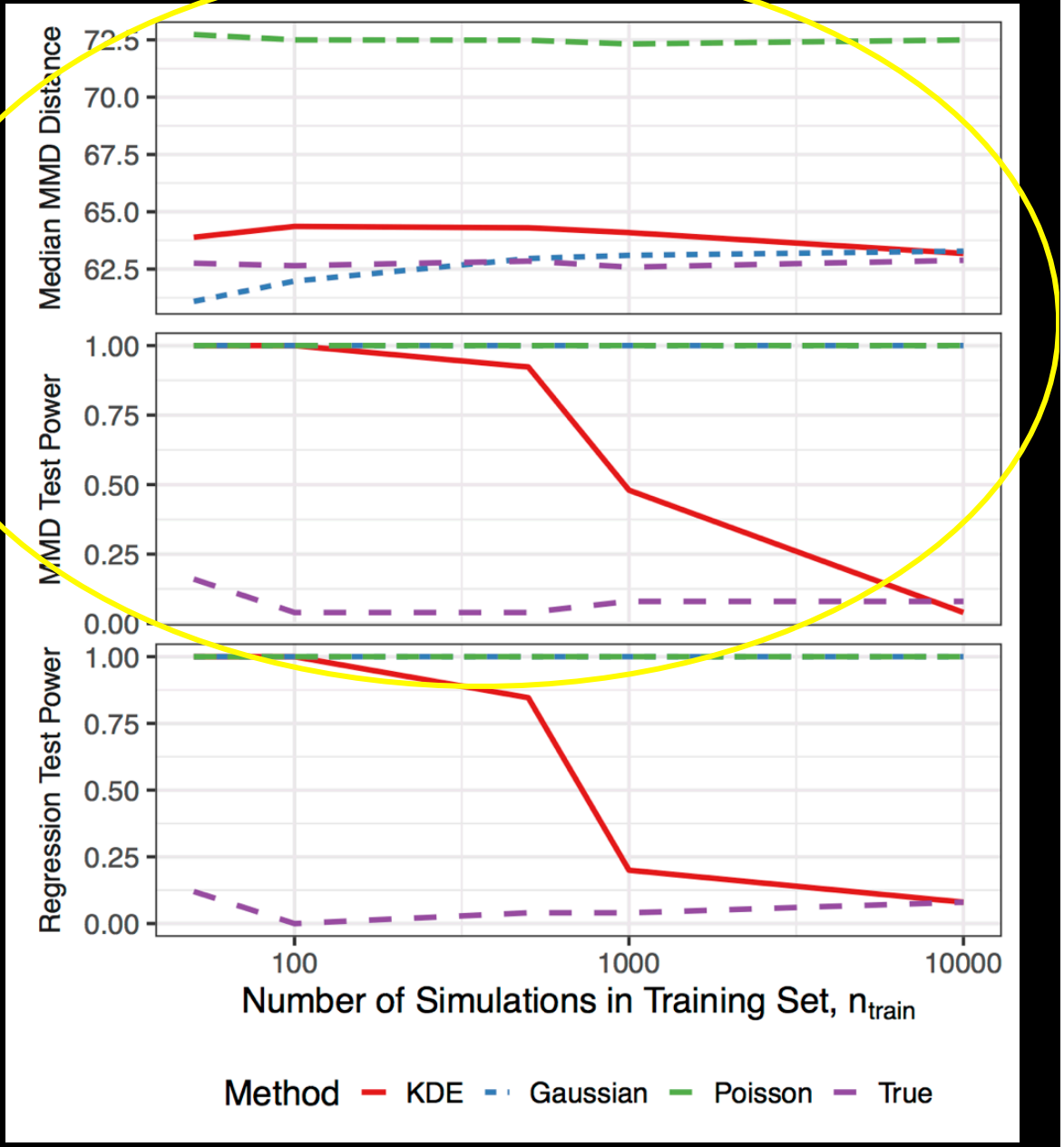
- Fit KDE, Gaussian and Poisson models with varying number of train simulations, n_{train}=[50, 100, 10000], but fixed number of test simulations, n_{sim}=200
- Global test on a grid of 100 values evenly spaced in [0,1] [0,1]. Repeat 100 times to compute power of test.
- To emphasize that it is the global test procedure itself that matters in this case use three different criteria:

1. median MMD distance between two samples
2. power of global test based on the same distance
3. power of global test based on RF regression

Why Not Just Use the Distance Metric?

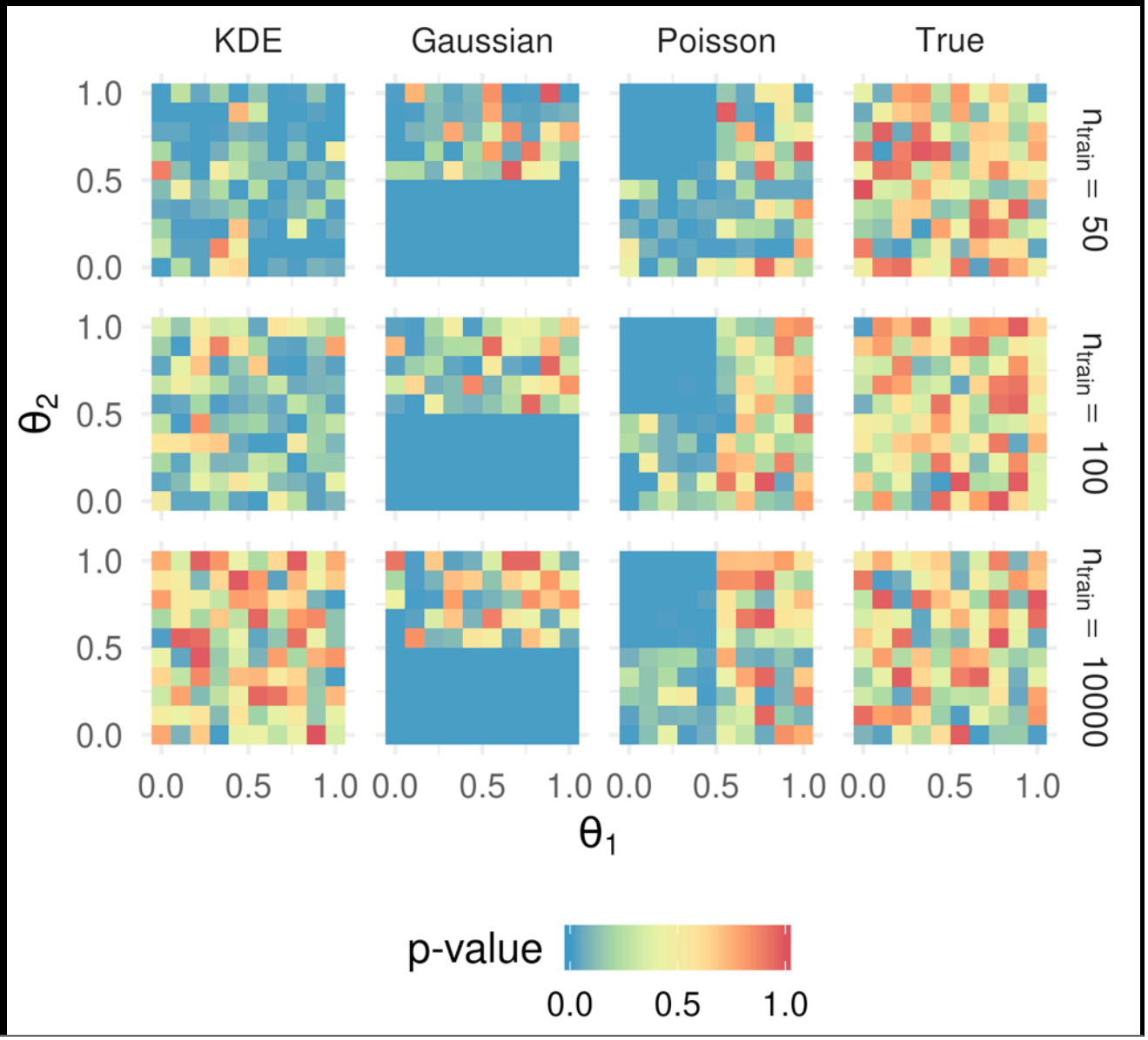
Median MMD distance (top) vs power of global test based on the same distance (center).

While the global test procedure can capture that KDE improves with the number of train simulations, the median distance is not informative.



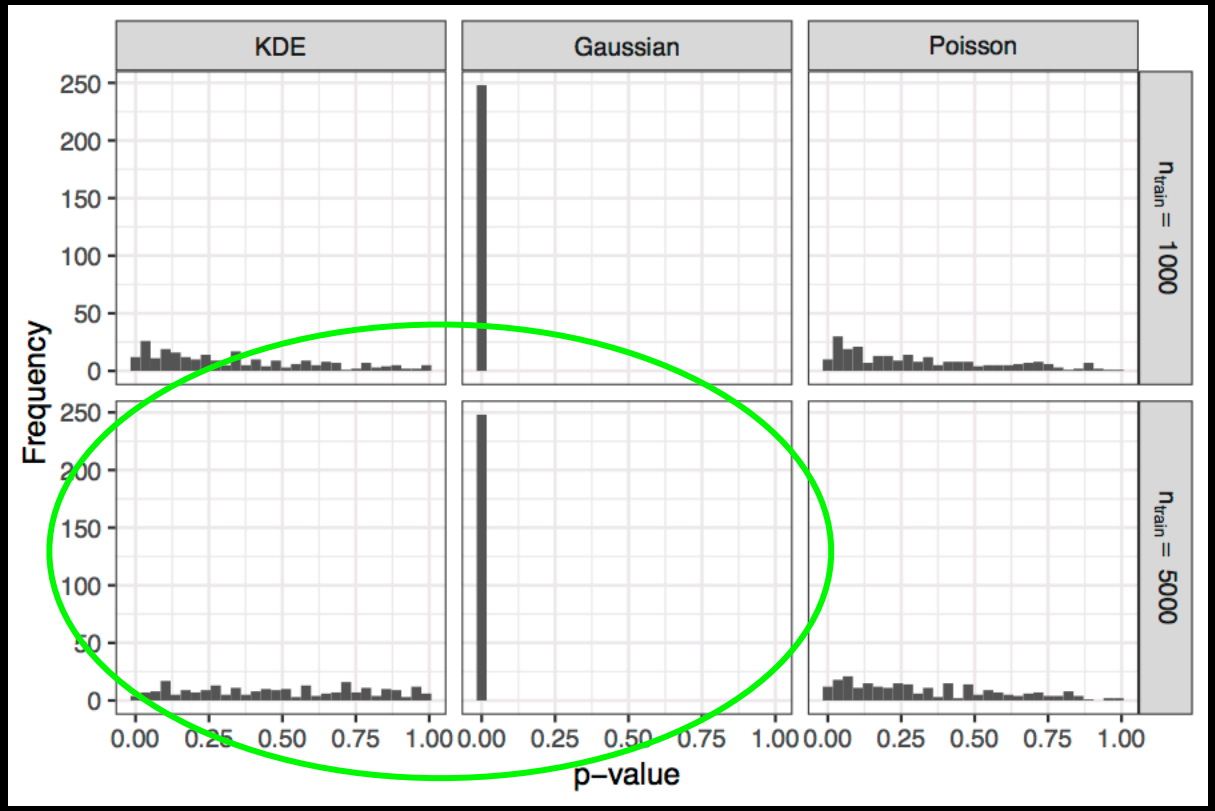
Local Regression Test (WHERE)

- The local regression p-values identifies regions where models fit poorly
- E.g. Gaussian model fits poorly for bottom half of parameter space where bins have low counts



Now Back to CAMELUS Peak Count Data

- If we increase the number of train simulations, KDE (bottom left) passes our global validation test, but the Gaussian model (bottom center) still fails the test



- Even if it's not feasible to generate new simulations, our regression local test still provides valuable information on how to improve the emulator (diagnostics)

Recall: Framework for Local Regression Test

Our null and alternative hypotheses are

$$H_0: \ \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) \ = \ \mathbb{P}(Y=1) \ \text{ for all } \mathbf{x} \in \mathcal{X}$$
$$H_1: \ \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) \ \neq \ \mathbb{P}(Y=1) \ \text{ for some } \mathbf{x} \in \mathcal{X}$$

Define the regression function $m(\mathbf{x}) = \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})$.
Let $\widehat{m}(\mathbf{x})$ be an estimate of $m(\mathbf{x})$ based on the sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$.
Let $\widehat{\pi}_1 = \frac{1}{n}\sum_{i=1}^n I(Y_i = 1)$.

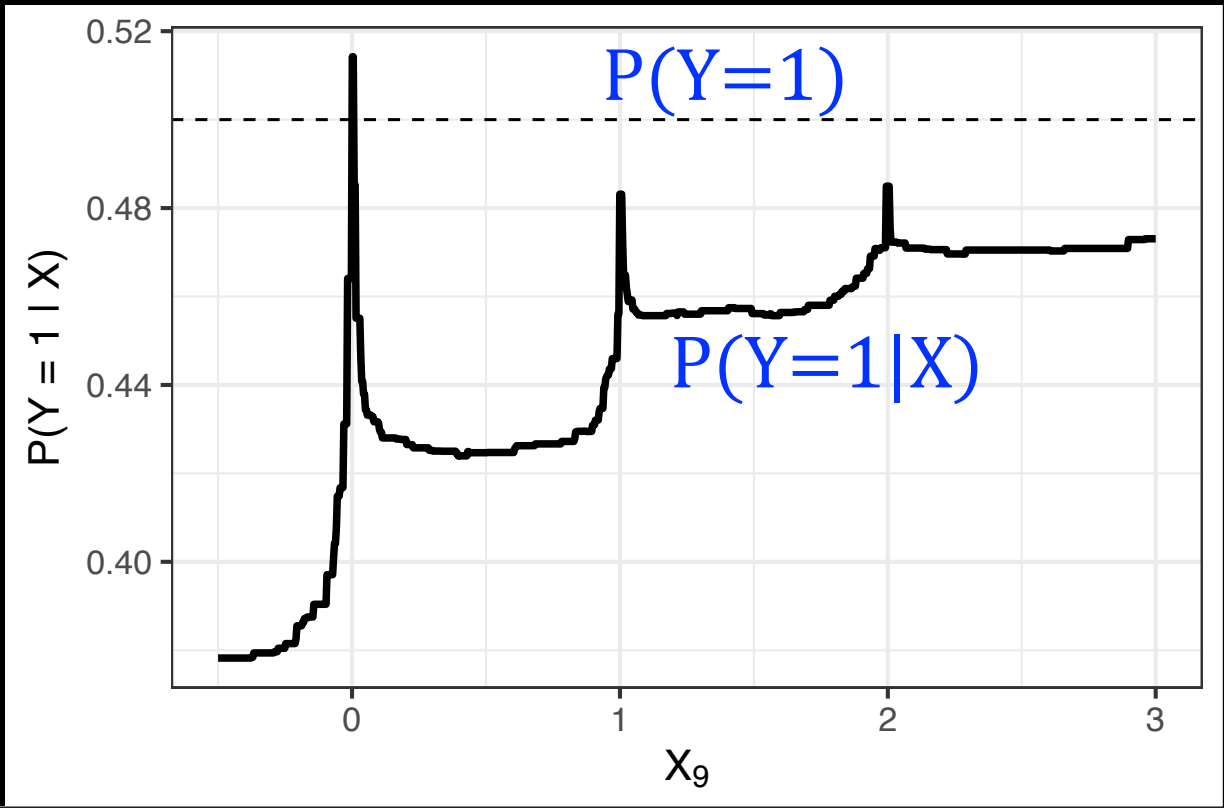
We define our test statistic as

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

Emulator Diagnostics: Our Regression Test Tells Us HOW the Two Samples are Different in X

- According to our random forest regression, bins with low counts (e.g. bin X_9) contribute the most to the rejection of the Gaussian model.

- Partial dependence plot for variable X_9. The regression test is distinguishing between the discrete true distribution and the approximate Gaussian continuous model.



SUMMARY: Validation of Emulators Fit to Slow Ensemble Simulations

- IF one needs to run more computationally intensive simulations to better fit an emulator to the simulations, or if the fit is close enough

- WHERE in parameter space one, if needed, should propose the next batch of simulations

- HOW emulated and high-resolution simulated data are different in high-dimensional observable space

SUMMARY: Validation of Emulators Fit to Slow Ensemble Simulations

- IF one needs to run more computationally intensive simulations to better fit an emulator to the simulations, or if the fit is close enough (answered by our fully consistent global procedure)
- WHERE in parameter space one, if needed, should propose the next batch of simulations
- HOW emulated and high-resolution simulated data are different in high-dimensional observable space

SUMMARY: Validation of Emulators Fit to Slow Ensemble Simulations

- IF one needs to run more computationally intensive simulations to better fit an emulator to the simulations, or if the fit is close enough (answered by our fully consistent global procedure)
- WHERE in parameter space one, if needed, should propose the next batch of simulations (answered by our local procedure)
- HOW emulated and high-resolution simulated data are different in high-dimensional observable space

SUMMARY: Validation of Emulators Fit to Slow Ensemble Simulations

- IF one needs to run more computationally intensive simulations to better fit an emulator to the simulations, or if the fit is close enough (answered by our fully consistent global procedure)
- WHERE in parameter space one, if needed, should propose the next batch of simulations (answered by our local procedure)
- HOW emulated and high-resolution simulated data are different in high-dimensional observable space (answered by our regression test)

What's Next?

- Connect our validation procedure to high-resolution simulations and emulator.
- If you have an application of interest or an emulator and want to join forces, then please contact me!

annlee@cmu.edu

EXTRA SLIDES START HERE

Algorithm 2 Two-Sample Regression Test via Permutations

Input: two i.i.d. samples $\mathcal{S}_0$ and $\mathcal{S}_1$ from distributions with resp. densities f_0 and f_1; number of permutations M; a regression method

Output: p-value for testing if $f_0(\mathbf{x}) = f_1(\mathbf{x})$ for every $\mathbf{x}$

1: Define an augmented sample $\{\mathbf{X}_i, Y_i\}_{i=1}^n$, where $\{\mathbf{X}_i\}_{i=1}^n = \mathcal{S}_0 \cup \mathcal{S}_1$, and $Y_i = I(\mathbf{X}_i \in \mathcal{S}_1)$.
2: Calculate the test statistic $\widehat{\mathcal{T}}$ in Equation 1.
3: Randomly permute $\{Y_1, \ldots, Y_n\}$. Refit $\widehat{m}$ and calculate the test statistic using the permuted data.
4: Repeat the previous step M times to obtain $\{\widehat{\mathcal{T}}^{(1)}, \ldots, \widehat{\mathcal{T}}^{(M)}\}$.
5: Approximate the permutation p-value by

$$p = \frac{1}{M+1}\left(1 + \sum_{m=1}^{M} I(\widehat{\mathcal{T}}^{(m)} > \widehat{\mathcal{T}})\right).$$

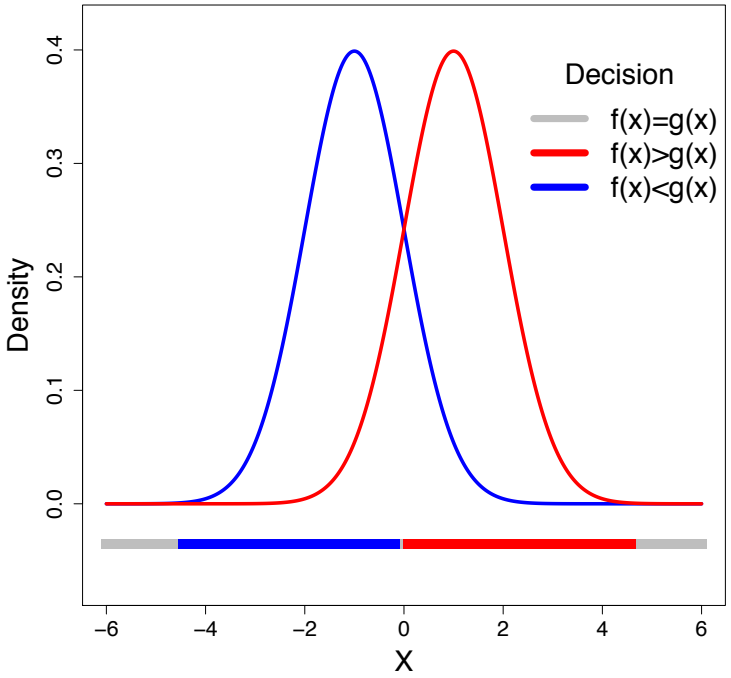
6: **return** p

How Two Samples are Different in Feature Space

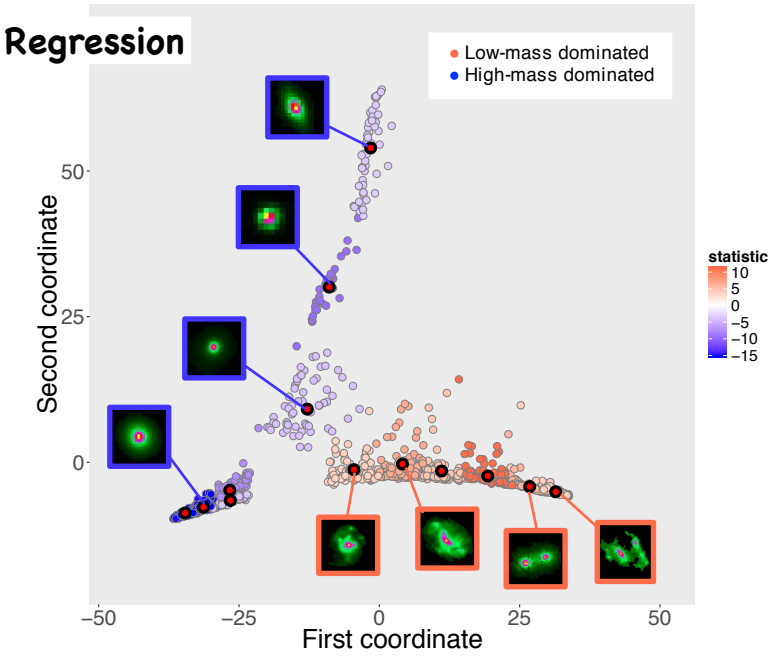
Interests: Galaxy morphology

- Comparing distributions of galaxy morphologies between two populations (high-mass vs. low-mass, old vs. new and high SFR vs. low SFR)
- Main interest is to know how two populations are locally different in a multivariate space of morphology statistics such as M, I, D, Gini, M_{20}, C and A.

Local significant differences (1D toy example)



Binomial Tests for Mass (7-dim feature space)



Ilmun Kim[1] , Ann Lee[1], Peter Freeman[1] and Jeffrey Newman[2]

[1]Department of Statistics
Carnegie Mellon University

[2]Department of Physics and Astronomy
University of Pittsburgh

Power of Local Test (Sparse Structure)

$X_1 \sim Bernoulli(\theta)$

$X_1 \sim N(0, \theta)$

$X_1 \sim \frac{1}{2}N(-\theta, 1) + \frac{1}{2}N(\theta, 1)$