2-Sample and GoF Testing via Regression

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

Joint work with Ilmun Kim, Jing Lei, Nic Dalmasso, Taylor Pospisil, Peter Freeman, Jeff Newman, Rafael Izbicki, Trey McNeely

Motivation and Goals

- The 2-sample regression test [Kim&Lee, 2016 ADA/CMU report; Kim/Lei/Lee, EJS 2019] is closely related to the better known 2-sample classification accuracy test [e.g., Kim et al 2021] but grew out of a different set of problems from astronomy and weather forecasting.
- Let's look at some examples that motivated our work…
  - two versions of 2-sample testing (ex1A&B)
  - two versions of GoF/consistency testing (ex2A&B)

Ex 1A: Comparing Distributions of High-Dimensional Data

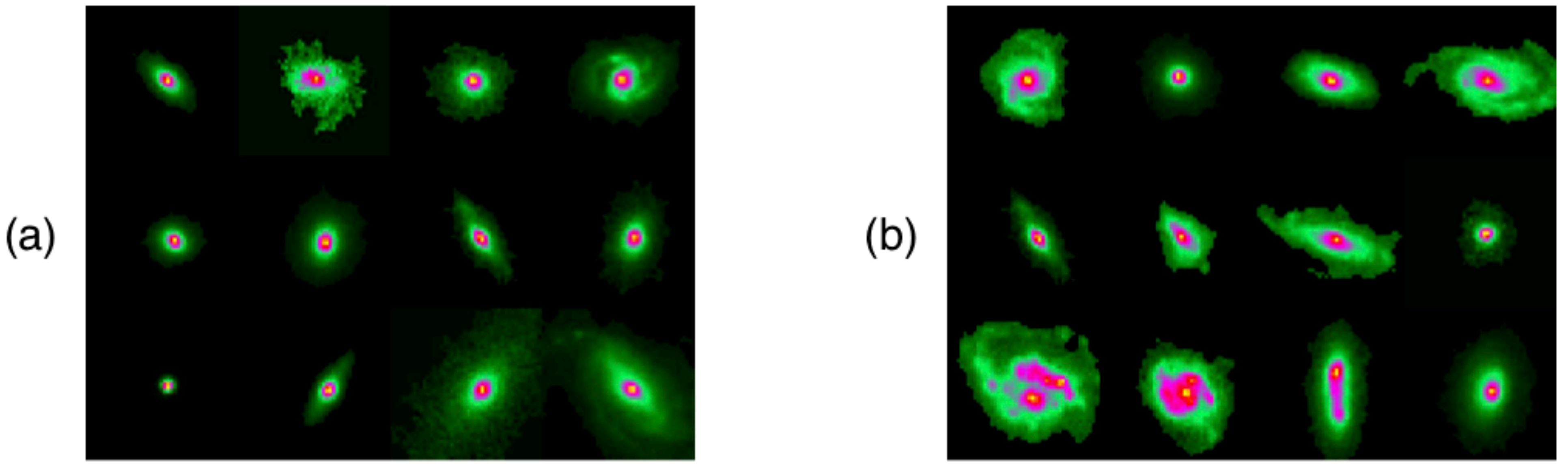
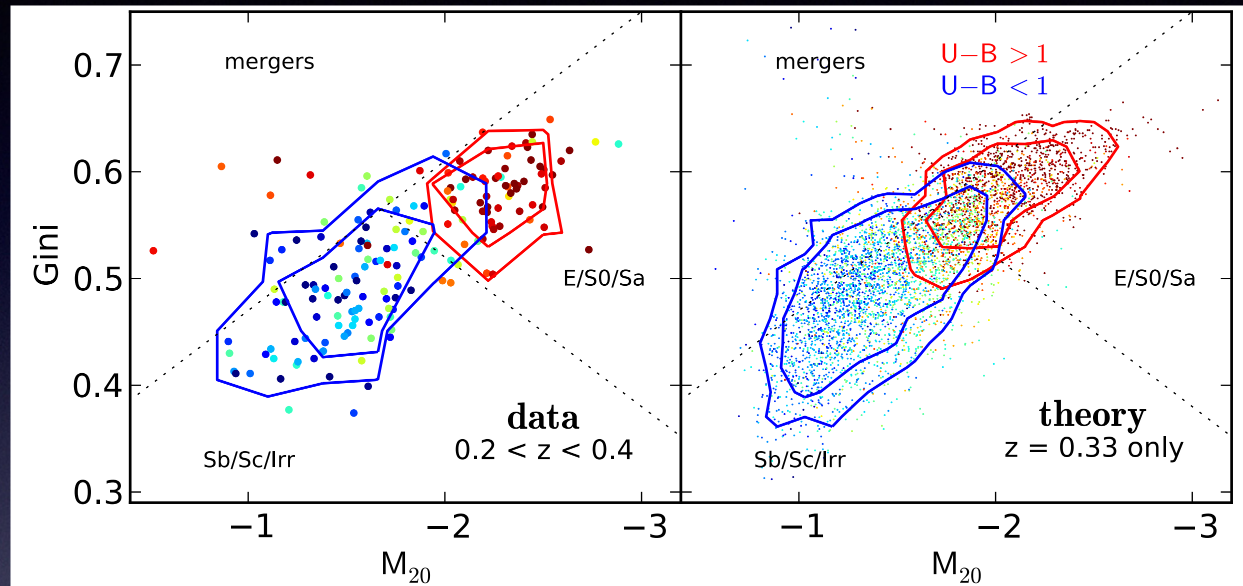


Figure 7: Examples of galaxies from (a) the low-SFR sample $\mathcal{S}_0$ versus (b) the high-SFR sample $\mathcal{S}_1$.

- Morphologies of two galaxy populations
- Can we answer the question if, and if so, how two populations are different beyond looking at low-dim summary statistics?

[Figure credit: Dalmasso et al, 2019]

Snyder et al. (2015)



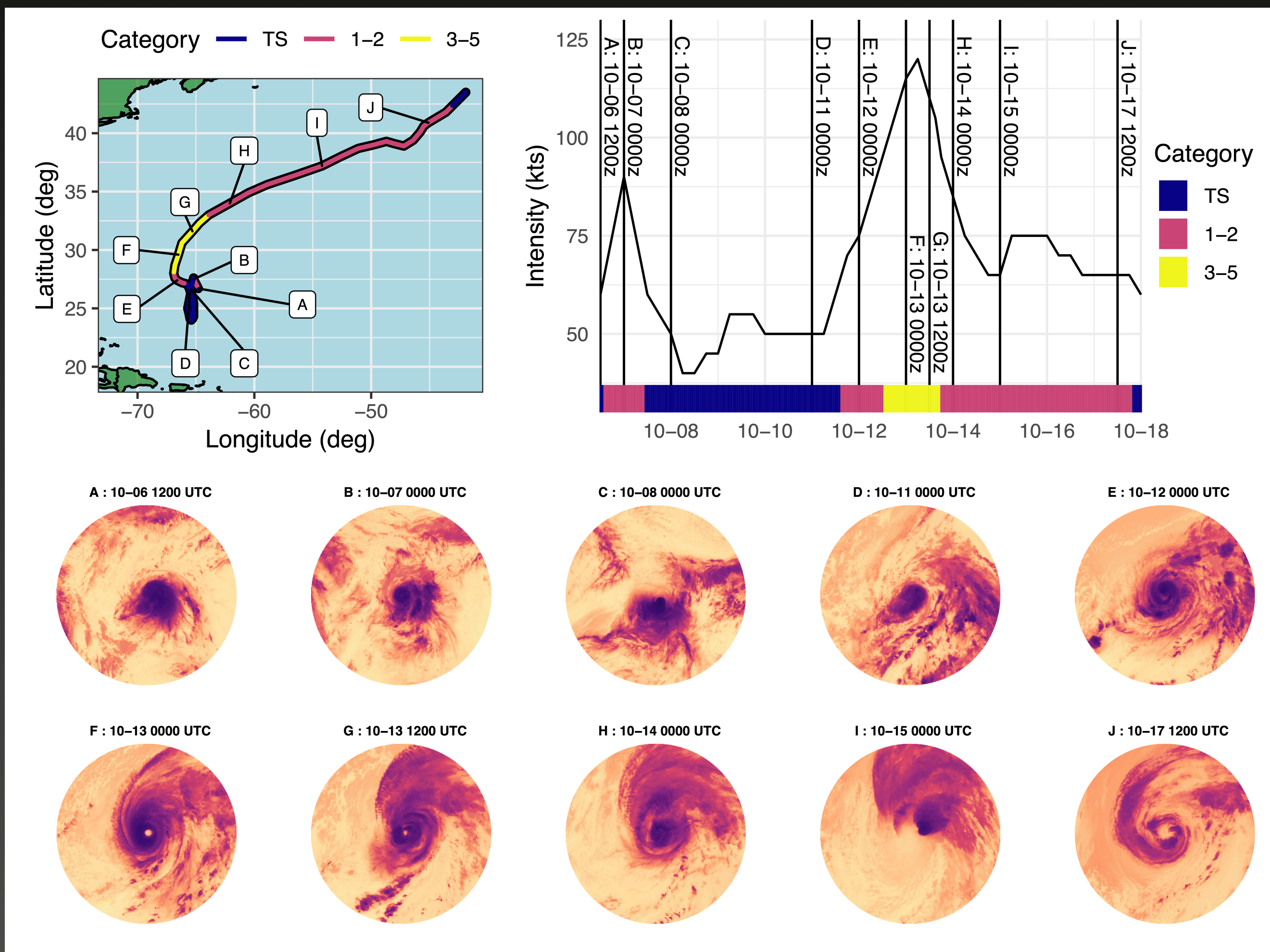
Snyder et al. (2015)

With regard to our first statistical aim, we wish to identify regions in the sample space where the distributions F and G are significantly different and to use this information, e.g., to infer redshift evolution (given two observed samples) or to inform improvements in simulation codes (by comparing simulation output at one wavelength to *HST* data at that same wavelength), etc.

[Slide credit: Peter Freeman]

Ex 1B: Detecting Distributional Differences in Labeled "D.I.D" Sequences of Images

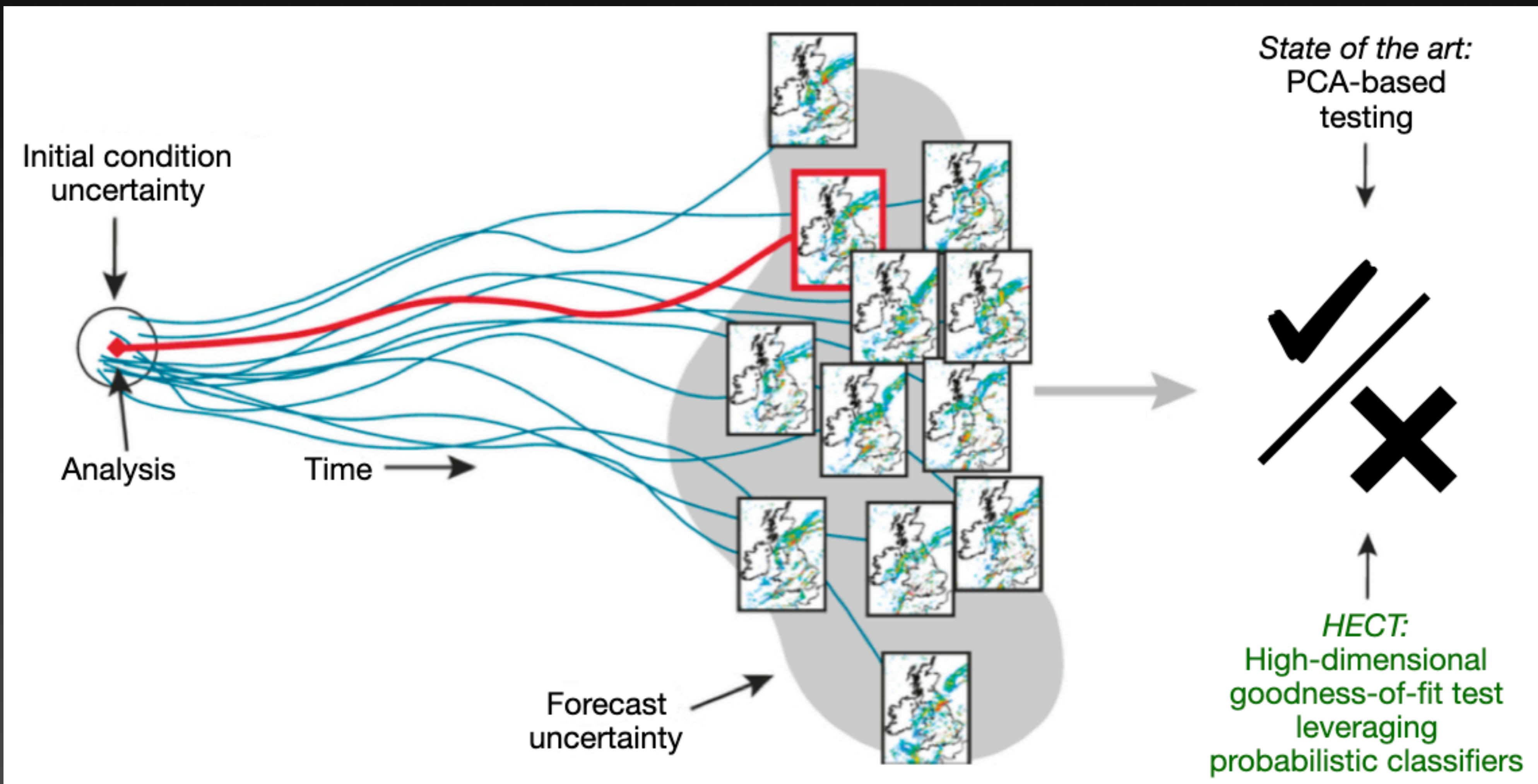
- Are the 24h sequences of satellite imagery preceding a rapid intensity change event (Y=1) vs a non-event (Y=0) different, and if so how?



Ref: McNeely et al; AOAS 2023

	NAL	ENP	Total
6-h HURDAT2 entries	8,438	8,080	16,518
⩾ 50-kt HURDAT2 entries	4,111	3,400	7,511
24-h history available	4,017	3,339	7,356
⩾ 250 km from land	2,225	2,627	4,852
GOES + SHIPS available	**1,236**	**1,575**	**2,811**
RI Observations	361	602	963
RW Observations	221	587	808
Unique TCs	154	206	360
RI Events	71	103	174
RW Events	56	106	162

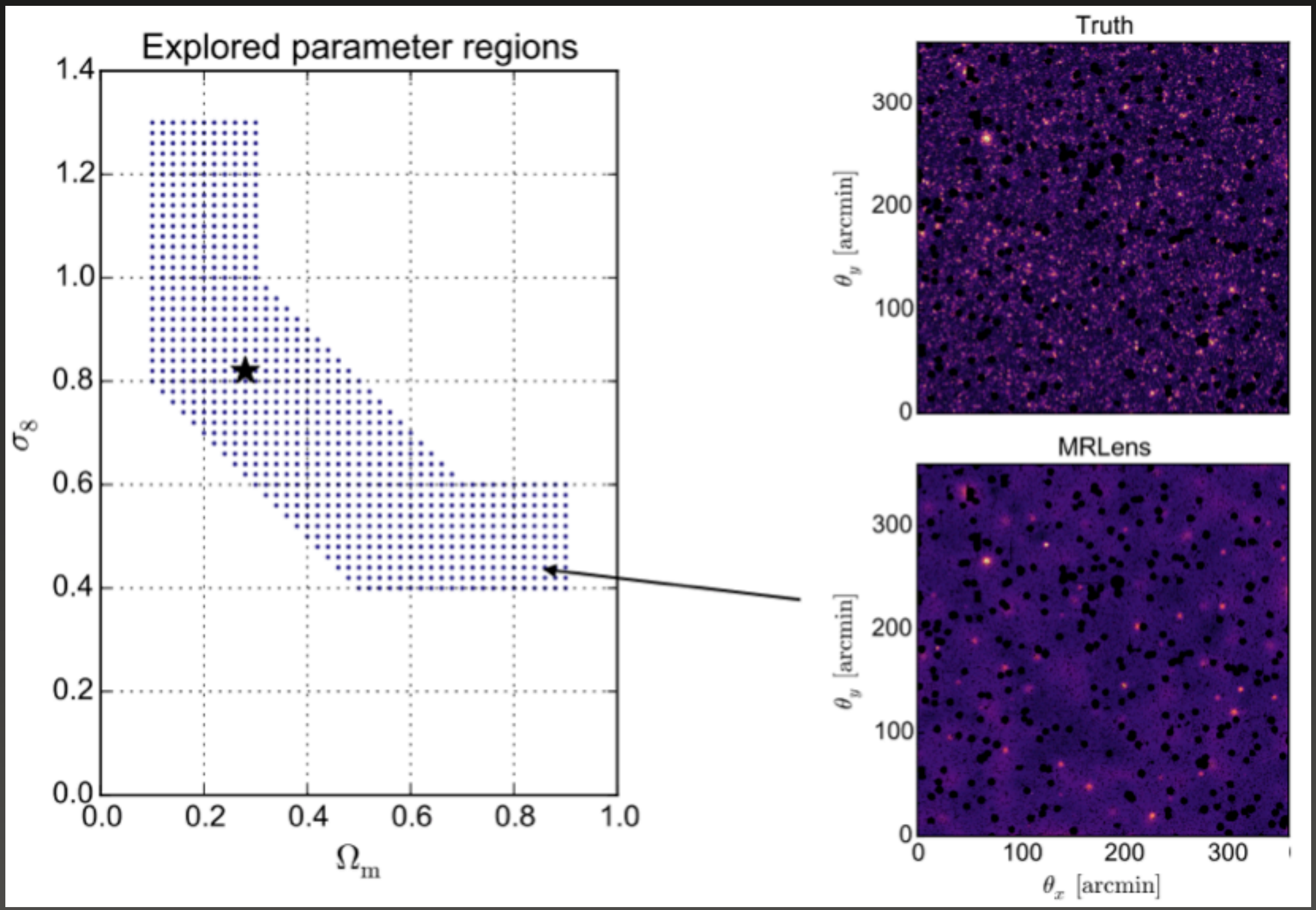
Ex 2A: Quality Assurance of Simulations by Ensemble Consistency Testing (ECT) for Climate Models



 [Dalmasso, Vincent, Hammerling and Lee 2020]

Ex 2B: Validation of Approximate Likelihood Models $\widehat{L}(x;\theta)$ fit to Computationally Intensive Simulations

- Simulate weak lensing data to constrain parameters of the Lambda CDM model in "Big Bang" cosmology.



[Image credit: Lin & Kilbinger, 2018]

Two-Sample and GoF Tests for I.I.D Data (today's talk)

Electronic Journal of Statistics
Vol. 13 (2019) 5253–5305
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1648

[Kim, Lee & Lei; EJS 2019]

Global and local two-sample tests via regression

Ilmun Kim, Ann B. Lee, and Jing Lei

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS 471, 3273–3282 (2017) doi:10.1093/mnras/stx1807
Advance Access publication 2017 July 18

[Freeman, Kim & Lee; MNRAS 2017]

Local two-sample testing: a new tool for analysing high-dimensional astronomical data

P. E. Freeman,★ I. Kim and A. B. Lee
Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

Validation of Approximate Likelihood and Emulator Models for Computationally Intensive Simulations

Niccolò Dalmasso,[1] Ann B. Lee,[1] Rafael Izbicki,[2] Taylor Pospisil,[3] Ilmun Kim,[1] Chieh-An Lin[4]

[Dalmasso et al; AISTATS 2020]

HECT: High-Dimensional Ensemble Consistency Testing for Climate Models

https://arxiv.org/abs/2010.04051 (NeurIPS Workshop 2020)

Niccolò Dalmasso*,[1] Galen Vincent*,[1] Dorit Hammerling[2] Ann B. Lee[1]
[1]Department of Statistics & Data Science, Carnegie Mellon University
[2]Department of Applied Mathematics and Statistics, Colorado School of Mines
ndalmass@stat.cmu.edu

Basic Setting: Two-Sample Testing

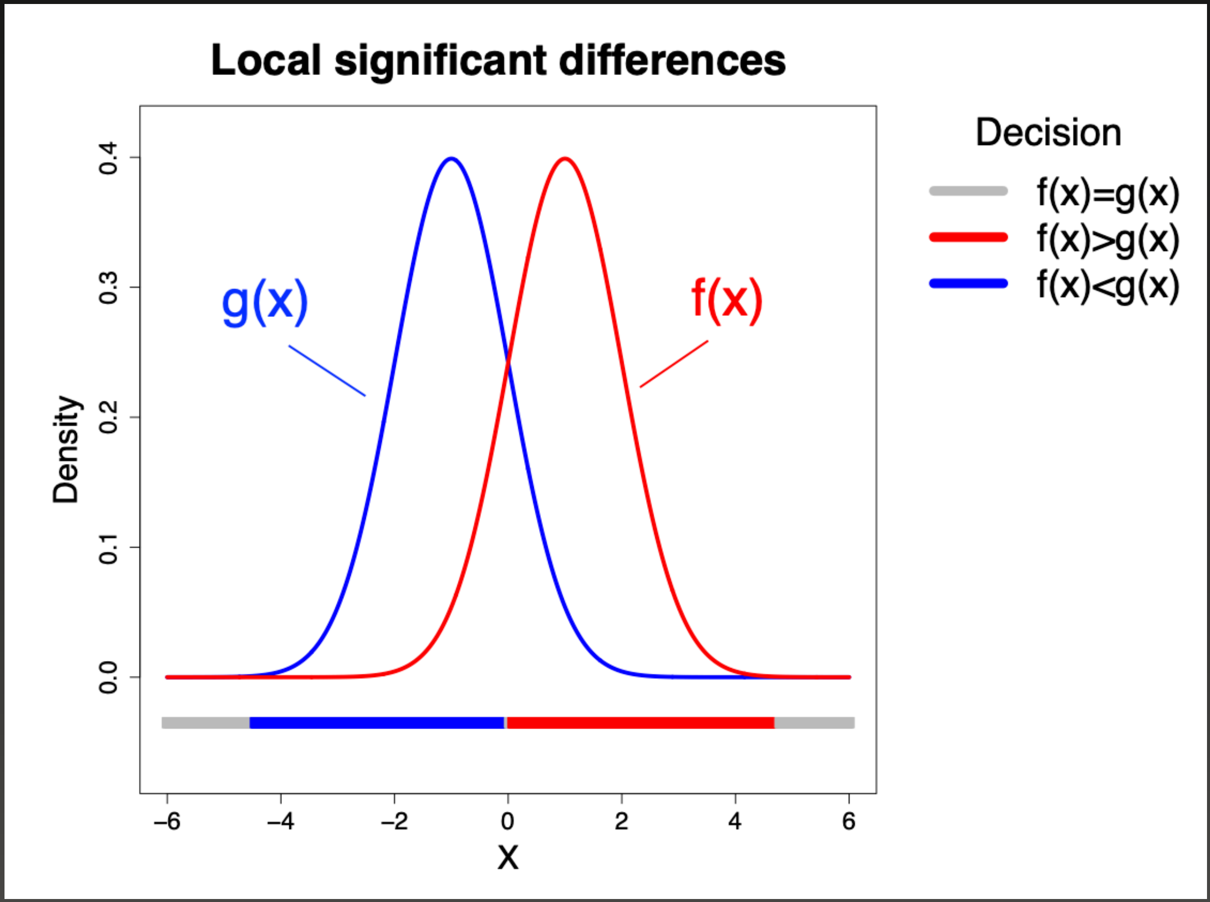
Suppose we have two samples:

$$\mathbf{X}_1^0, \ldots, \mathbf{X}_{n_0}^0 \sim P_0 \quad \text{and} \quad \mathbf{X}_1^1, \ldots, \mathbf{X}_{n_1}^1 \sim P_1$$

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

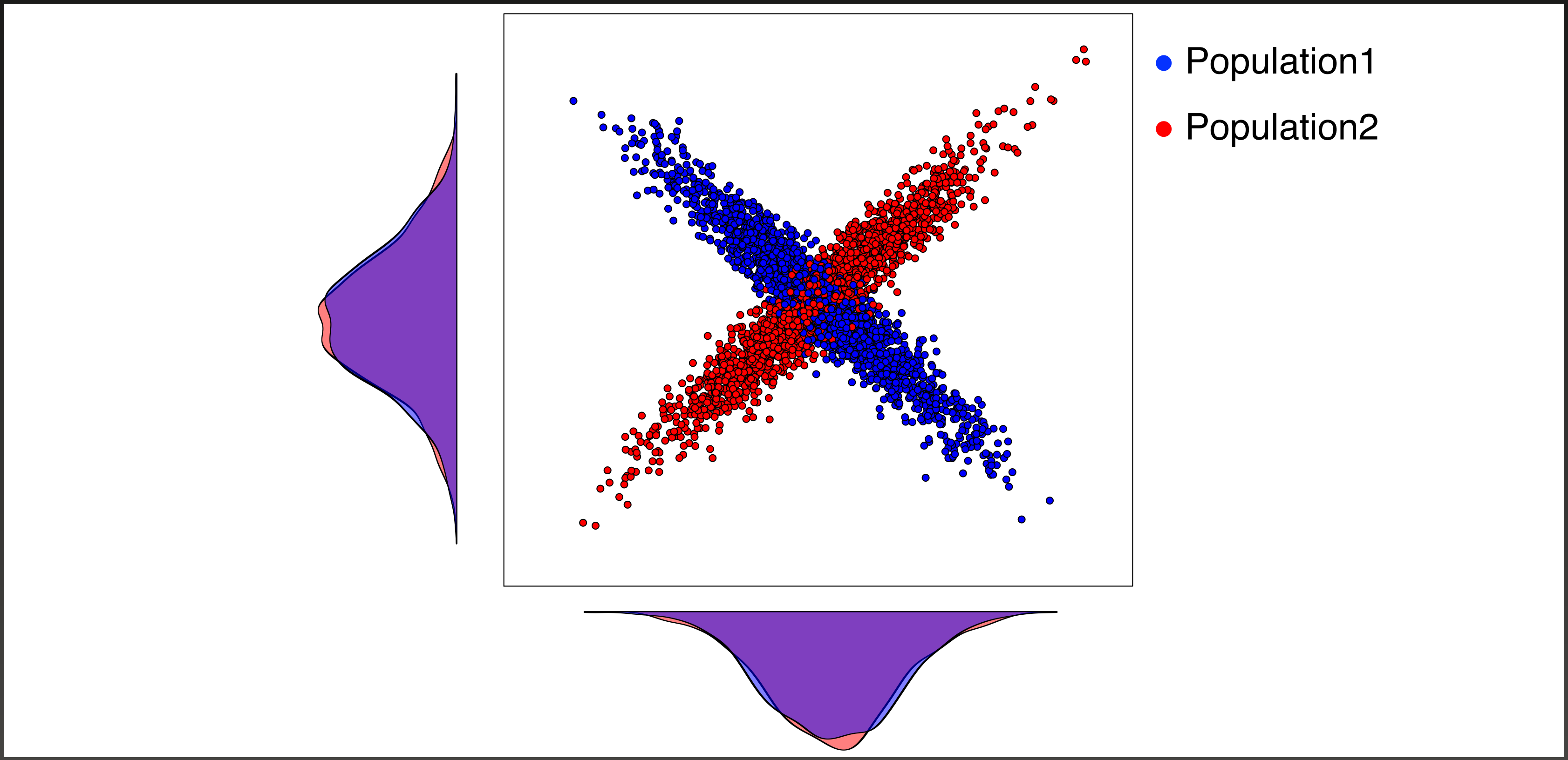
$$H_0 : f(\mathbf{x}|Y = 0) = f(\mathbf{x}|Y = 1), \quad \text{for all } \mathbf{x} \in \mathcal{X}$$

1. We are looking for regions in the state space where the two populations have significantly different densities



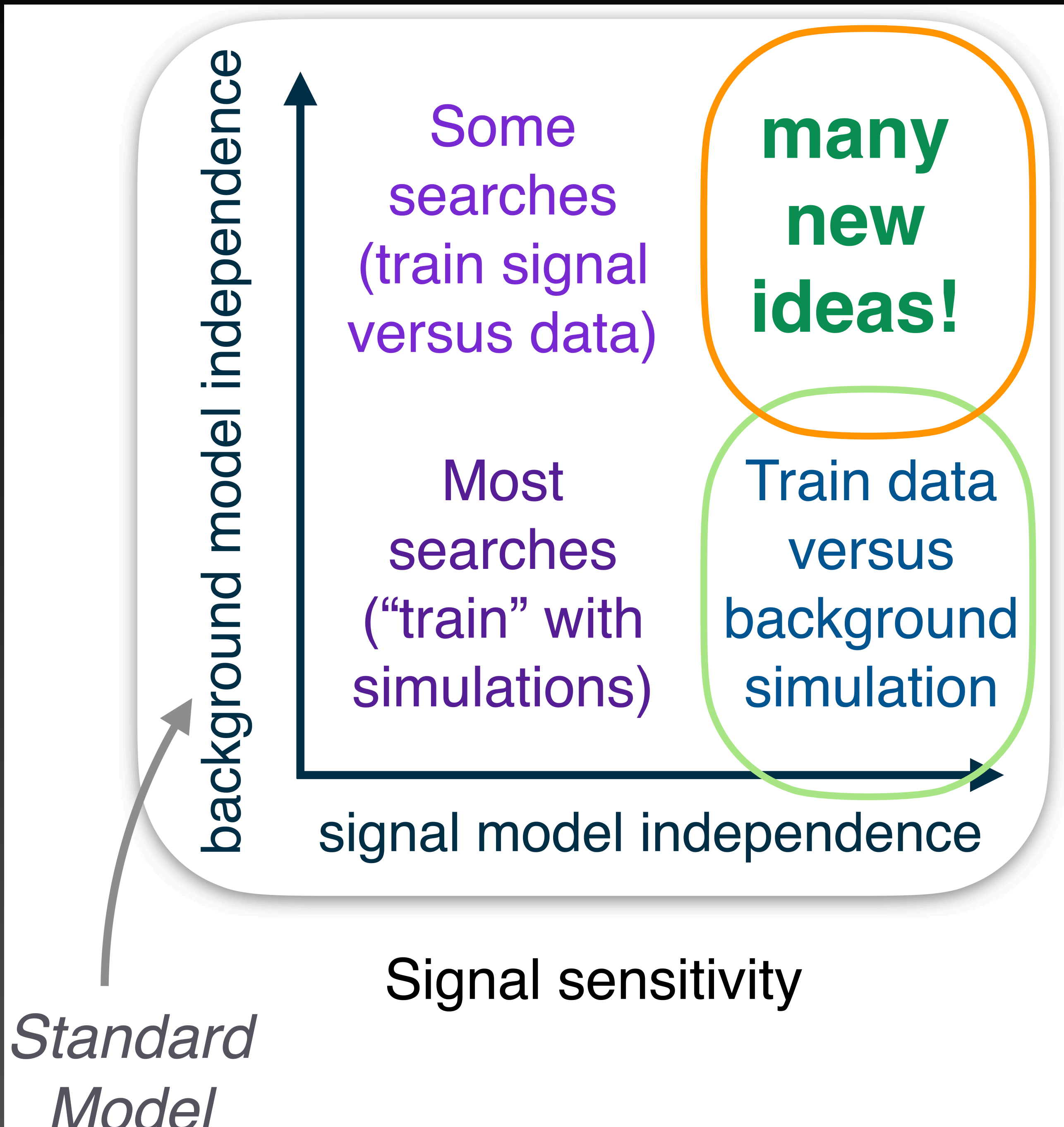
[Figure credit: Ilmun Kim 2016]

2. We are searching for differences in high-dimensional space (e.g., each data point could represent an image or a sequence of images)



[Figure credit: Ilmun Kim 2016]

3. We are Targeting Model Independent Searches



Source: Ben Nachman, "*Landscape of model-independent Searches*" PhyStat-Anomalies 2022

Two-Sample Test via Regression

[Freeman/Kim/Lee MNRAS 2017; Kim/Lee/Lei EJS 2019]

Suppose we have two samples:

$$\mathbf{X}_1^0, \ldots, \mathbf{X}_{n_0}^0 \sim P_0 \quad \text{and} \quad \mathbf{X}_1^1, \ldots, \mathbf{X}_{n_1}^1 \sim P_1$$

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

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

Two-Sample Test via Regression

[Freeman/Kim/Lee MNRAS 2017; Kim/Lee/Lei EJS 2019]

Suppose we have two samples:

$$\mathbf{X}_1^0, \ldots, \mathbf{X}_{n_0}^0 \sim P_0 \quad \text{and} \quad \mathbf{X}_1^1, \ldots, \mathbf{X}_{n_1}^1 \sim P_1$$

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

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

$$\Updownarrow$$

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 2-samples testing to a regression problem

Our null and alternative hypotheses are

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

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 a train set $\mathcal{D} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n \subset \mathcal{X}$.
Let $\widehat{\pi}_1 = \frac{1}{n}\sum_{i=1}^n I(Y_i = 1)$.

Convert 2-samples testing to a regression problem

Our null and alternative hypotheses are

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

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 a train set $\mathcal{D} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n \subset \mathcal{X}$.
Let $\widehat{\pi}_1 = \frac{1}{n}\sum_{i=1}^n I(Y_i = 1)$.

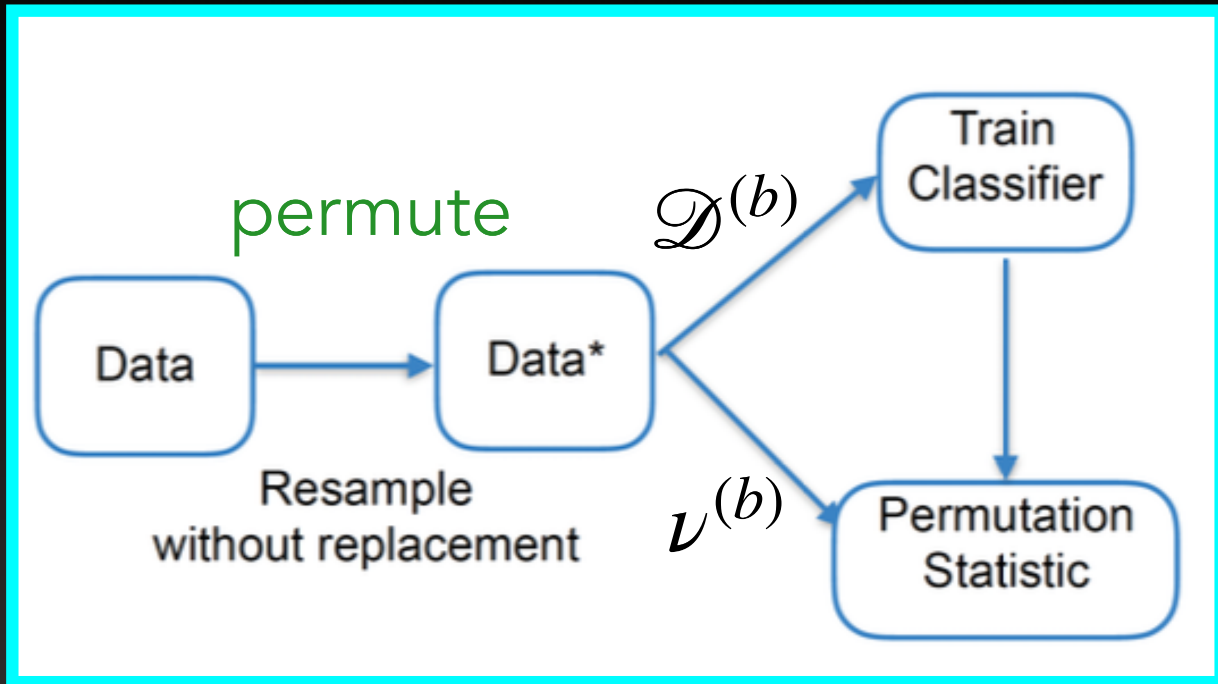
Compute the "local posterior difference" (LPD) at evaluation points $\mathcal{V} \subset \mathcal{X}$:

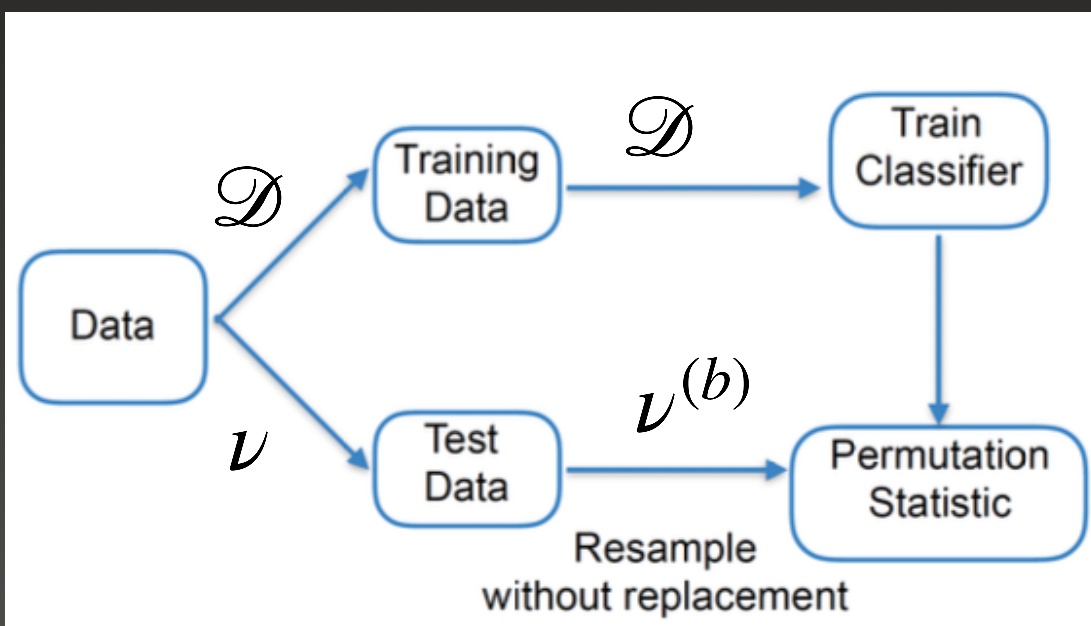
$$\lambda(\mathbf{x}) := \widehat{m}(\mathbf{x}) - \widehat{\pi}_1$$

We define our global test statistic as

$$\widehat{\mathcal{T}} = \frac{1}{|\mathcal{V}|}\sum_{\mathbf{x}\in\mathcal{V}} \lambda(\mathbf{x})^2$$

Compute p-values by permutations (b=1,…B)





[Adapted from: Chakravarati @BanffSystematics2023]

Both full (top) and half (bottom) permutation yield finite-n validity
[Heinerik and Gorman, 2018; Kim et al, 2021]

Why Two-Sample Test via Regression?

$$H_0: \; f(\mathbf{x}|Y=0) = f(\mathbf{x}|Y=1), \quad \text{for all } \mathbf{x} \in \mathcal{X}$$
$$H_1: \; f(\mathbf{x}|Y=0) \neq f(\mathbf{x}|Y=1), \quad \text{for some } \mathbf{x} \in \mathcal{X}$$

$$\widehat{\mathcal{T}} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{x} \in \mathcal{V}} \left(\widehat{m}(\mathbf{x}) - \widehat{\pi}_1\right)^2.$$

- Can adapt to any structure in X for which there is a suitable regression technique
- The power of the regression 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 the state space

Why Two-Sample Test via Regression?

$$H_0: \; f(\mathbf{x}|Y=0) = f(\mathbf{x}|Y=1), \quad \text{for all } \mathbf{x} \in \mathcal{X}$$
$$H_1: \; f(\mathbf{x}|Y=0) \neq f(\mathbf{x}|Y=1), \quad \text{for some } \mathbf{x} \in \mathcal{X}$$

$$\widehat{\mathcal{T}} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{x} \in \mathcal{V}} \left(\widehat{m}(\mathbf{x}) - \widehat{\pi}_1 \right)^2 .$$

- Can adapt to any structure in X for which there is a suitable regression technique
- The power of the regression 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 the state space

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*

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

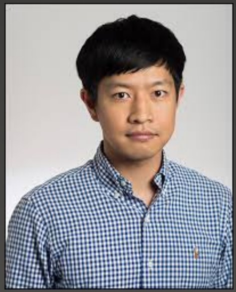
where C_0 is a positive constant, $\delta_n = o(1)$, $\delta_n \geq n^{-1}$, and $\mathcal{M}$ is a class of regressions $m(\mathbf{x})$ containing constant functions. Let t_α^* be the upper α quantile of the permutation distribution of the test statistic $\widehat{\mathcal{T}}'$ on validation data.[1]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

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

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.



Ref: Kim, Lee & Lei; EJS 2019

Practical implication: We should choose a regression method that predicts the "class membership" Y well

Our null and alternative hypotheses are

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

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 a train set $\mathcal{D} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n \subset \mathcal{X}$.
Let $\widehat{\pi}_1 = \frac{1}{n}\sum_{i=1}^n I(Y_i = 1)$.

What about Classification Accuracy Tests?

- Regression tests are very similar to the better known classification accuracy tests [Kim et al 2021].

$$H_0: \; f(\mathbf{x}|Y=0) = f(\mathbf{x}|Y=1)$$

$$\Updownarrow$$

$$H_0: \; \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) = \mathbb{P}(Y=1)$$

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

$$h(x) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) > \alpha, \\ 0 & \text{otherwise} \end{cases}$$

$$H_0: \mathbb{P}\left(\{h(\mathbf{X}) \neq Y\}\right) = \frac{1}{2}$$

$$\widehat{\mathcal{T}}_{acc} = \frac{1}{|\mathcal{V}|}\sum_{\mathbf{X}_i \in \mathcal{V}} \mathbb{I}\left(\widehat{h}(\mathbf{X}_i) \neq Y_i\right)$$

Classification Accuracy Tests Usually Have Similar or Slightly Lower Power

- Consider e.g. a normal means problem where we test for mean differences between two multivariate normals:

$$\mathbf{X}|Y = 0 \sim N(\mu_0, \Sigma), \quad \mathbf{X}|Y = 1 \sim N(\mu_1, \Sigma)$$

$$H_0 : \mu_0 = \mu_1 \quad \text{versus} \quad H_1 : \mu_0 \neq \mu_1$$

- The classification accuracy test with Fisher's LDA is typically underpowered compared to Hotelling's T^2 test [Ramdas et al 2016; Rosenblatt et al 2016]. In contrast, a regression test with Fisher's LDA has optimal asymptotic power [Kim et al 2019].

Power comparisons for finite n0=n1=100 (D=5, 20):

The regression test based on Fisher's LDA has comparable power to Hotelling's T^2 test. Accuracy Tests have less power.

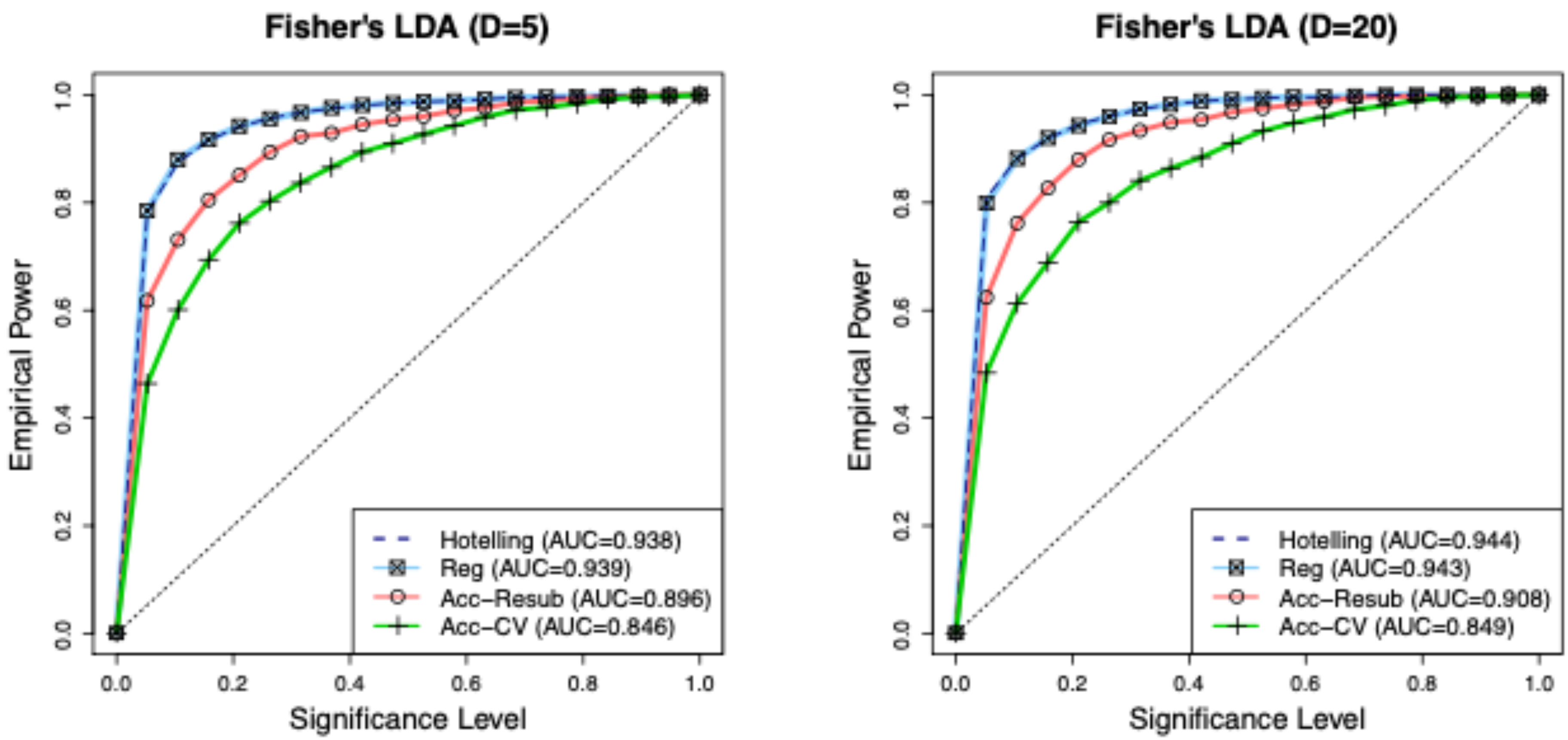


Fig 2: Power comparisons between Hotelling's T^2 (Hotelling), $\widehat{\mathcal{T}}_{\mathrm{LDA}}$ (Reg), the in-sample accuracy (Acc-Resub), and the cross-validated accuracy (Acc-CV) via Fisher's LDA.

Ref: Kim, Lee & Lei; EJS 2019

Why Two-Sample Test via Regression?

$$H_0: \; f(\mathbf{x}|Y=0) = f(\mathbf{x}|Y=1), \quad \text{for all } \mathbf{x} \in \mathcal{X}$$
$$H_1: \; f(\mathbf{x}|Y=0) \neq f(\mathbf{x}|Y=1), \quad \text{for some } \mathbf{x} \in \mathcal{X}$$

$$\widehat{\mathcal{T}} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{x} \in \mathcal{V}} \left(\widehat{m}(\mathbf{x}) - \widehat{\pi}_1\right)^2 .$$

- Can adapt to any structure in X for which there is a suitable regression technique
- The power of the regression 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, two distributions are different in state space

Let's return to the galaxy morphology example (ex1A)

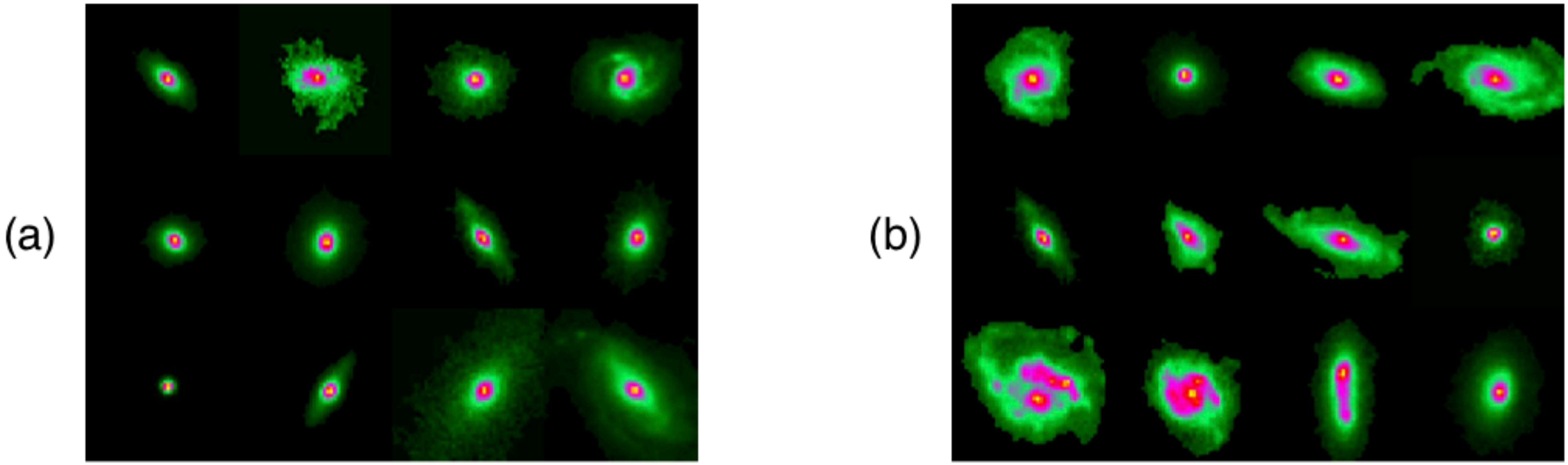


Figure 7: Examples of galaxies from (a) the low-SFR sample $\mathcal{S}_0$ versus (b) the high-SFR sample $\mathcal{S}_1$.

- Divide 2736 galaxies from the CANDELS program into two populations based on SFR: "Low SFR" vs "High SFR" sample
- Consider seven morphology summary statistics jointly
- Are the morphologies the same or not (compared to chance) for the two populations?

Regression Test to Identify If and How Two Distributions Differ in 7-Dim Feature Space

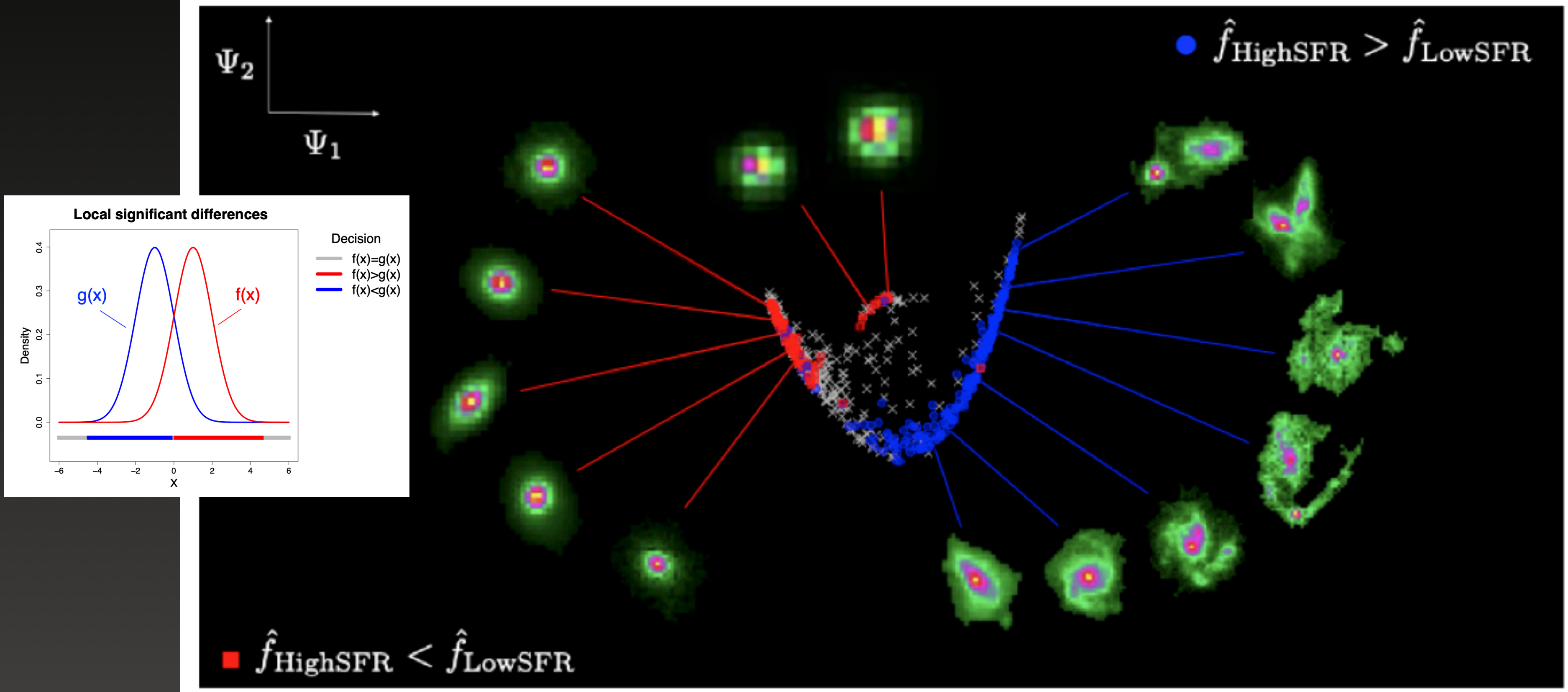
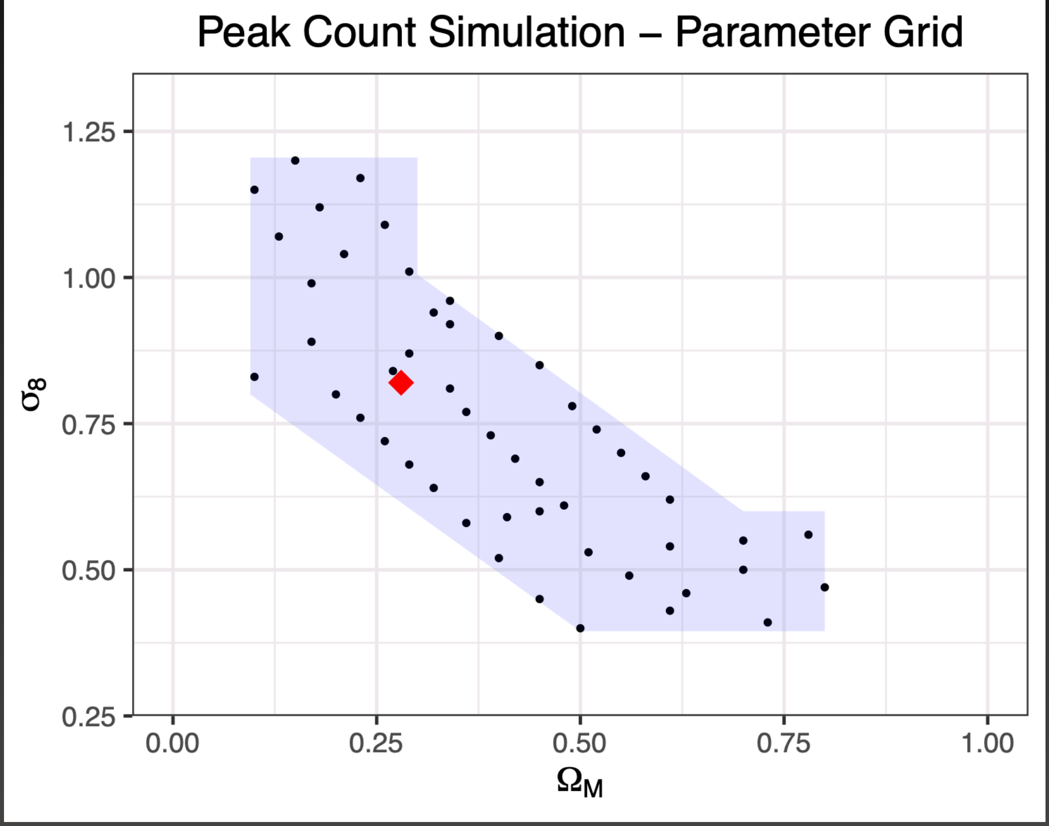


Figure 9: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR galaxies. The test points are visualized via a two-dimensional diffusion map. Figure adapted from [49].

Back to Example 2A: Validation of Approximate Likelihood Models for Weak Lensing Data



- Use CAMELUS [Lin & Kilbinger 2015] to simulate weak lensing convergence maps
  $\Rightarrow$ binned peak counts $x \in \mathbb{N}^7$
- Batch of 200 train + 200 test simulations at 50 different cosmologies/parameter settings.
- Fit 3 different approximate likelihood models: Gaussian, Poisson, Masked Autoregressive Flows (MAFs)

We can use the regression test to validate approximate likelihood (emulator) models for computationally intensive simulations

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

- Our framework can help answer:
  - IF one needs to improve the emulator model
  - WHERE in parameter space Θ the fit might be poor
  - HOW the distributions of emulated and high-fidelity simulated data may differ in observable space χ

We can use the regression test to validate approximate likelihood (emulator) models for computationally intensive simulations

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

- Our framework can help answer:
  - IF one needs to improve the emulator model
  - WHERE in parameter space Θ the fit might be poor
  - HOW the distributions of emulated and high-fidelity simulated data may differ in observable space χ

Two-Step Procedure: 2-Sample Test at Each Parameter.
(IF) Global test of Uniformity of Local p-Values

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

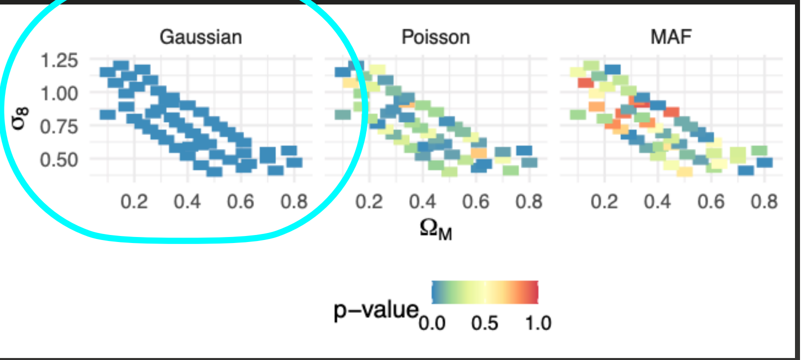
Output: p-value p for testing if $L(\mathbf{x};\theta) = \widehat{L}(\mathbf{x};\theta)$ for every $\mathbf{x}$ and θ

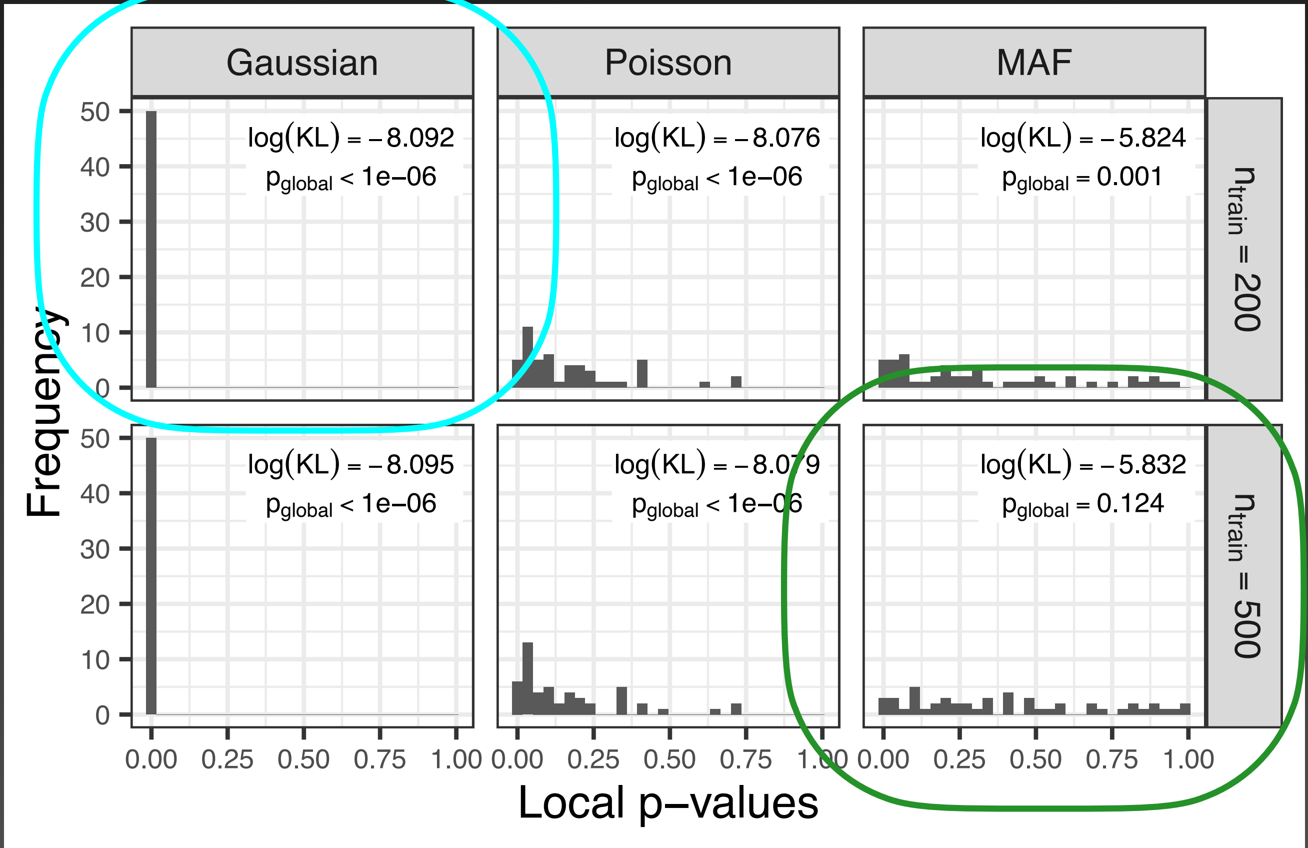
1: **for** $i \in \{1, \ldots, B\}$ **do**
2: sample $\theta_i \sim r(\theta)$
3: compute p_{θ_i} using Algorithm 1
4: **end for**
5: Compute p-value p for testing if $(p_{\theta_i})_{i=1}^{B}$ has a uniform distribution.
6: **return** p

- For the local test, our regression test allows us to accommodate any data type with interpretable diagnostics.
- Global test is consistent against all alternatives if the local test is consistent.

WHERE: Do we need more simulations to fit the data well? If so, where in parameter space?

- 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 parameter values





HOW: Even if it's not feasible to simulate more data, our regression test provides valuable diagnostics...

$$H_0: \ f(\mathbf{x}|Y=0) = f(\mathbf{x}|Y=1), \quad \text{for all } \mathbf{x} \in \mathcal{X}$$
$$H_1: \ f(\mathbf{x}|Y=0) \neq f(\mathbf{x}|Y=1), \quad \text{for some } \mathbf{x} \in \mathcal{X}$$

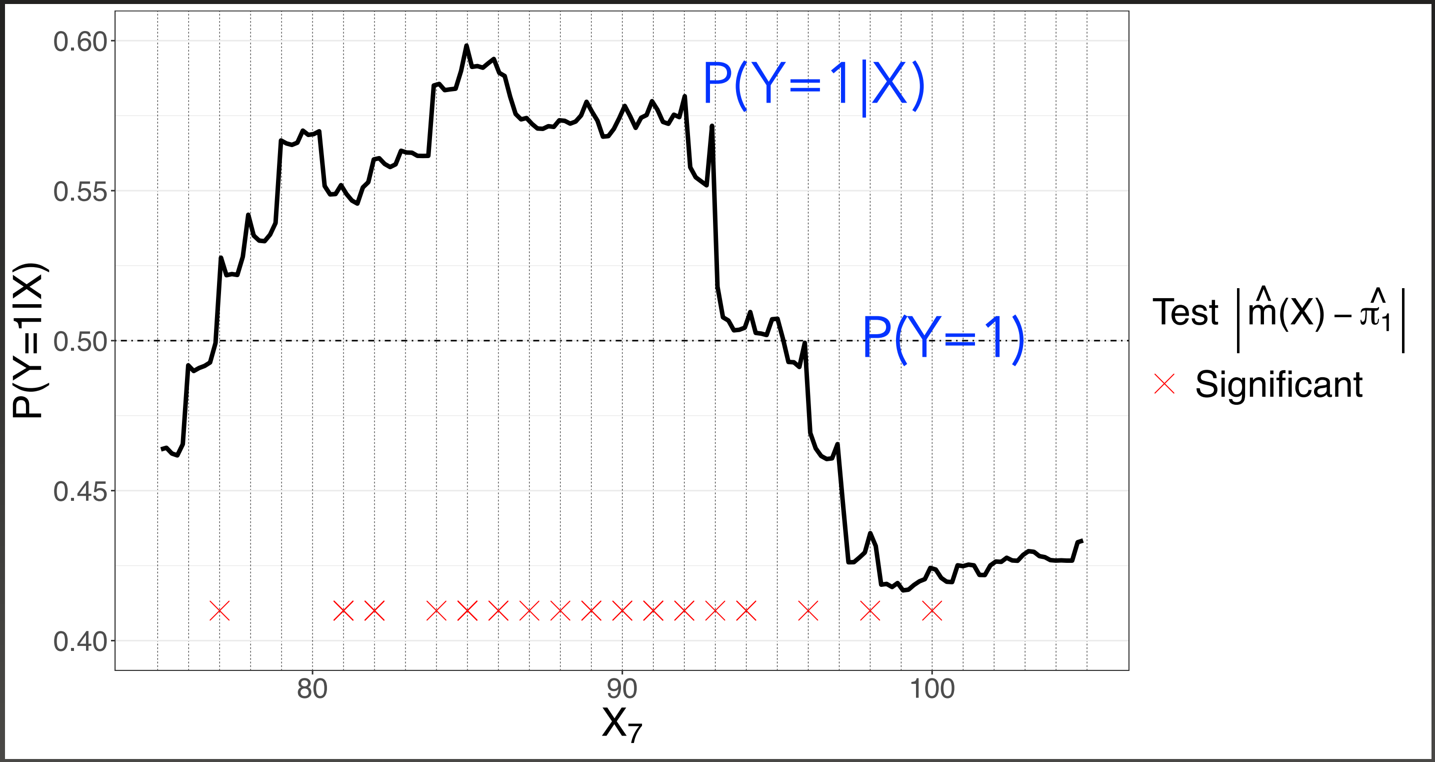
$$\widehat{\mathcal{T}} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{x} \in \mathcal{V}} \left(\widehat{m}(\mathbf{x}) - \widehat{\pi}_1\right)^2.$$

- the difference $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ provides information on how well the emulator fits the simulator in feature space: we can test whether $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ is statistically significantly higher!

Emulator diagnostics: Our regression test tells us how the two samples are different in $\mathbb{N}^7$

- According to our random forest regression, bins with low counts (e.g. bin X_7) contribute the most to the rejection of the Gaussian model.

Partial dependence plot for variable X_7. The regression test is distinguishing between the discrete true distribution and the approximate Gaussian continuous distribution.



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 2-sample test via regression)

Open Problems: 2-Sample and GoF Testing

- Q1: What if we don't have an ensemble/batch setting?

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

- Q2: Is the regression 2-sample test locally valid?

- Q3: Can we increase power of a GoF test via MC sampling from the emulator model (reference distribution) to check consistency with a sample from the ``trusted'' model?

Q1. What if we don't have an ensemble setting?

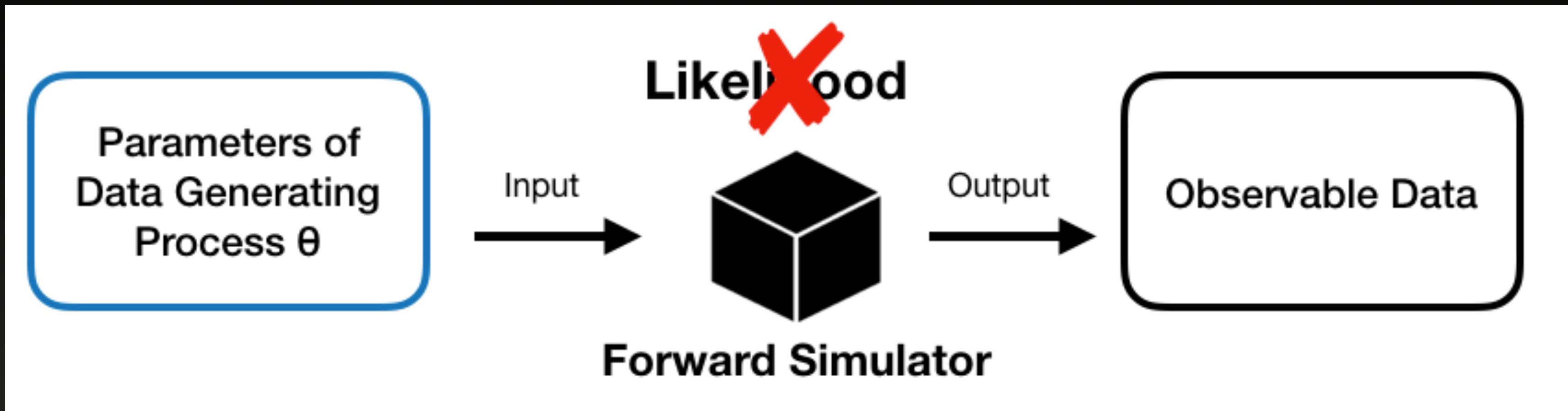


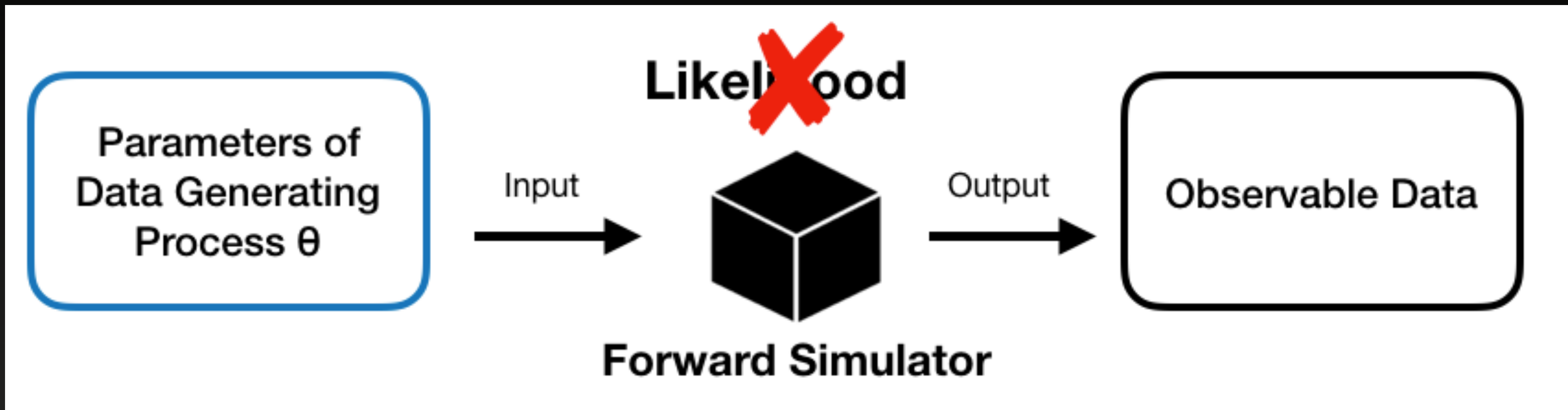
Image credit: Nic Dalmasso

$$\mathcal{S} = \{(\theta_1, \mathbf{X}_1), (\theta_2, \mathbf{X}_2), \ldots, (\theta_n, \mathbf{X}_n)\}, \text{ where } \theta \sim r(\theta), \;\; \mathbf{X}|\theta \sim \mathcal{L}(\mathbf{x};\theta)$$

$$\mathcal{S}_e = \{(\theta'_1, \mathbf{X}'_1), (\theta'_2, \mathbf{X}'_2), \ldots, (\theta'_{n_e}, \mathbf{X}'_{n_e})\}, \;\; \text{where } \theta' \sim r'(\theta), \;\; \mathbf{X}|\theta' \sim \widehat{\mathcal{L}}(\mathbf{x};\theta')$$

$$\mathrm{H}_0 : \;\; \widehat{\mathcal{L}}(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta), \quad \text{for every } \mathbf{x} \in \mathcal{X} \text{ at fixed } \;\; \theta \in \Theta$$

Regress Y on Both x and θ



$$H_0(\theta): \ \widehat{\mathcal{L}}(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta), \quad \text{for every } \mathbf{x} \in \mathcal{X} \text{ at fixed } \ \theta \in \Theta$$

$$\Updownarrow$$

$$H_0(\theta): \ \mathbb{P}(Y=1|\mathbf{x},\theta) = \mathbb{P}(Y=1|\theta), \quad \text{for every } \mathbf{x} \in \mathcal{X} \text{ at fixed } \ \theta \in \Theta$$

$$\widehat{\mathcal{T}} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{x}\in\mathcal{V}} \left(\widehat{m}(\mathbf{x},\theta) - \widehat{\pi}_1(\theta)\right)^2 .$$

Q2. Is the regression 2-sample test valid locally?

$$H_0: \; f_0(\mathbf{x}) = f_1(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathcal{X}$$
$$H_1: \; f_0(\mathbf{x}) \neq f_1(\mathbf{x}), \quad \text{for some } \mathbf{x} \in \mathcal{X}$$

$$\widehat{\mathcal{T}} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{x} \in \mathcal{V}} \left(\widehat{m}(\mathbf{x}) - \widehat{\pi}_1 \right)^2 .$$

$$H_0(\mathbf{x}): f_0(\mathbf{x}) = f_1(\mathbf{x}), \quad \text{at fixed } \mathbf{x} \in \mathcal{X}$$
$$H_1(\mathbf{x}): f_0(\mathbf{x}) \neq f_1(\mathbf{x}), \quad \text{at fixed } \mathbf{x} \in \mathcal{X}$$

$$\widehat{\mathcal{T}}_{local}(\mathbf{x}) = \widehat{m}(\mathbf{x}) - \widehat{\pi}_1$$

Approximate Validity of Local p-Values

Assumption 1 (Local regression estimator). *There exists* $\epsilon > 0$ *such that* $\widehat{m}$ *only uses the sample points in* $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ *with* $\mathbf{X}_i \in \mathcal{B}(\mathbf{x}; \epsilon)$*, where* $\mathcal{B}(\mathbf{x}; \epsilon)$ *is a ball in* $\mathcal{X}$ *of radius* ϵ *centered at* $\mathbf{x}$.

Theorem 1. *Under the null hypothesis*

$$H_0^{\epsilon}(\mathbf{x}) : f_0(\mathbf{x}') = f_1(\mathbf{x}') \text{ for all } \mathbf{x}' \in \mathcal{B}(\mathbf{x}; \epsilon)$$

and under Assumption 1, for any $0 < \alpha < 1$

$$\lim_{B \longrightarrow \infty} \mathbb{P}\left(p_{local}(\mathbf{x}) \leq \alpha\right) = \alpha.$$

Q3. Can we increase the power by MC sampling?

Suppose we have i.i.d. sample

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \sim P$$

from some unknown distribution P with density f.

Collective anomaly detection: Want to detect whether the collection of these data points deviate from what is anticipated under the assumed model P_0 with density f_0.

$$H_0 : f(\mathbf{x}) = f_0(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathcal{X}$$

Suppose we have i.i.d. sample

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \sim P$$

from some unknown distribution P with density f.

Collective anomaly detection: Want to detect whether the collection of these data points deviate from what is anticipated under the assumed model P_0 with density f_0.

$$H_0 : f(\mathbf{x}) = f_0(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathcal{X}$$

- Suppose we can sample from P_0. As suggested by Friedman (2004), may achieve higher power than the 2-sample permutation test by repeated MC sampling from the reference distribution P_0.

- See Dalmasso (2019, Appendix D) for procedure and theory https://http://proceedings.mlr.press/v108/dalmasso20a.html/abs/1905.11505

Take Away

- We can leverage regression methods (probabilistic classifiers) to identify if and how two samples differ.

$$\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F \quad \text{and} \quad \mathbf{X}_1^*, \ldots, \mathbf{X}_n^* \sim F^*$$

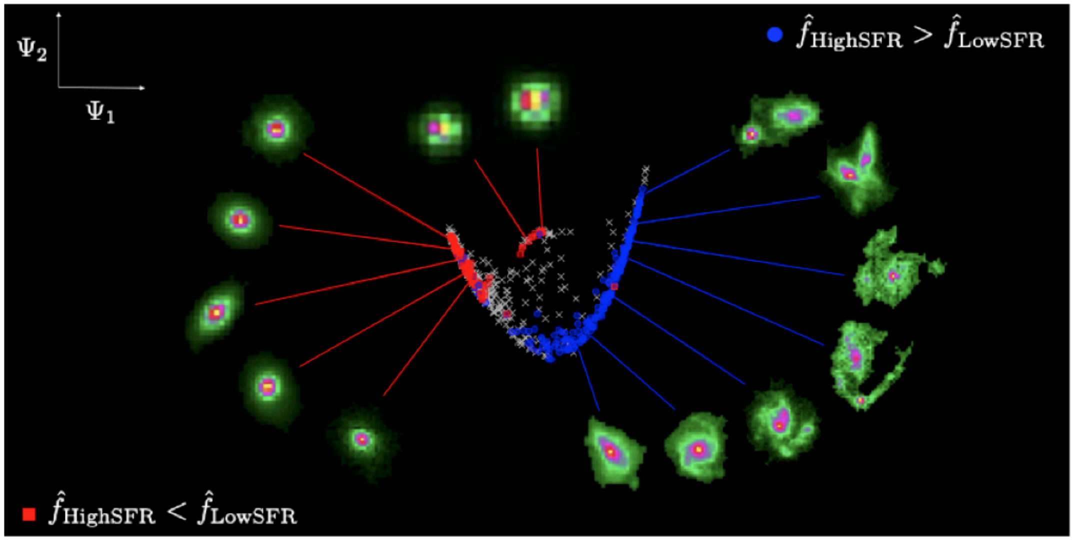
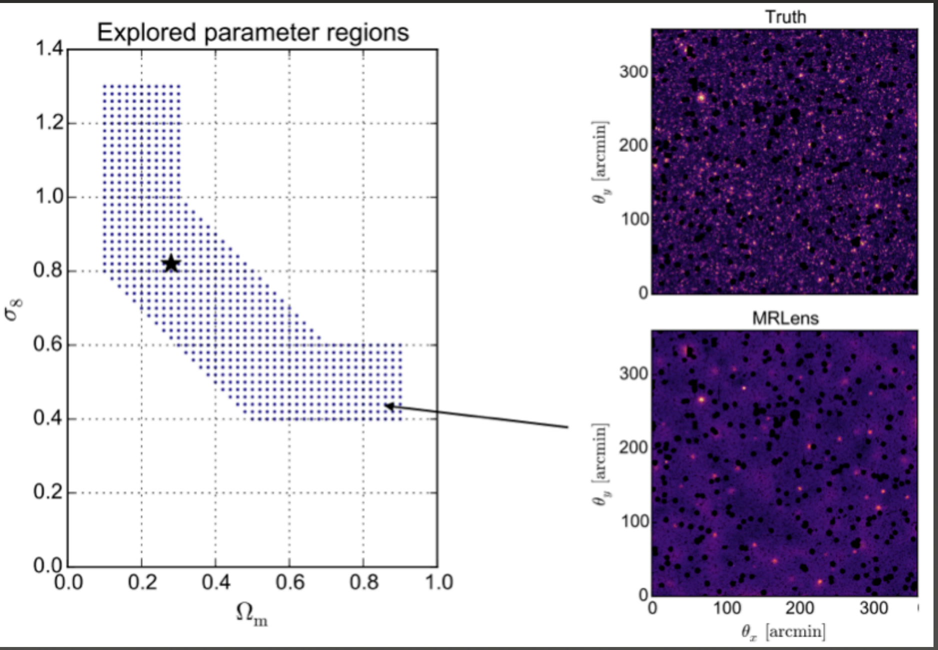


Figure 9: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR galaxies. The test points are visualized via a two-dimensional diffusion map. Figure adapted from [49].



Open Problems

- Local test: validity and power
- GoF tests and MC sampling: how to best simulate (best statistical performance at lowest computational cost)

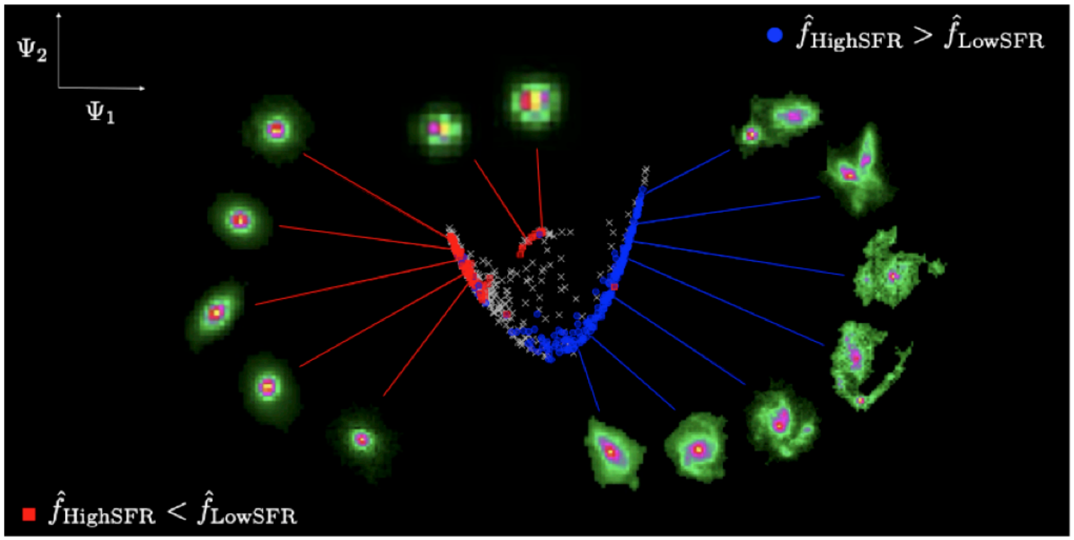
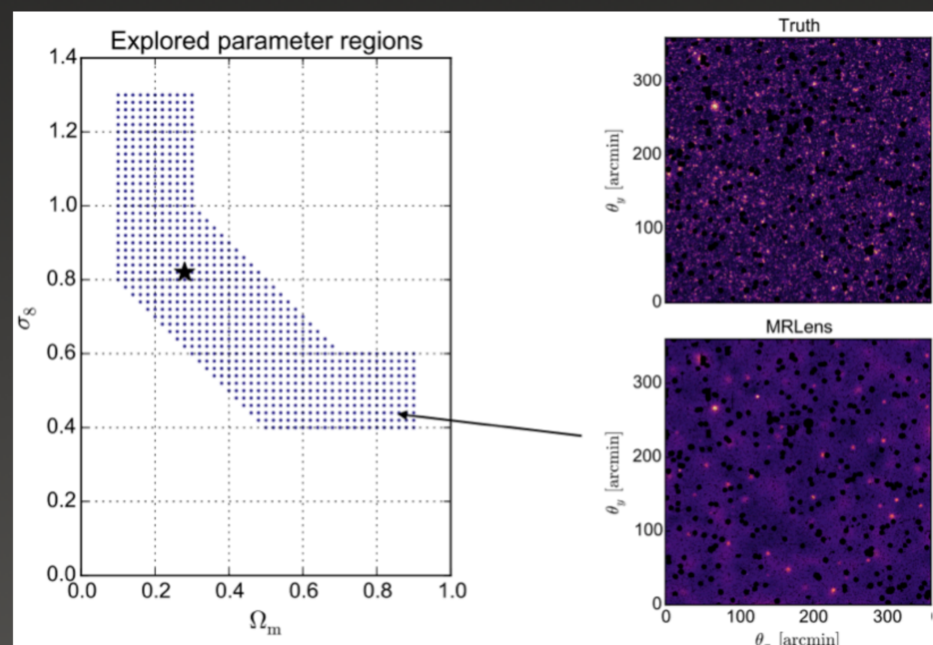


Figure 9: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR galaxies. The test points are visualized via a two-dimensional diffusion map. Figure adapted from [49].



EXTRA SLIDES

2-Sample Regression Test via Permutations

Algorithm 1: Two-Sample Regression Testing 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 B; a regression method $\widehat{m}$

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

1: Define an augmented sample $\{\mathbf{X}_i, Y_i\}_{i=1}^n$, where $\{\mathbf{X}_i\}_{i=1}^n = \mathcal{S}_0 \cup \mathcal{S}_1$, and $Y_i = I(\mathbf{X}_i \in \mathcal{S}_1)$.

2: Calculate the global test statistic $\widehat{\mathcal{T}}_{global}$ (by, e.g., training the regression on the first half of the sample, and then evaluating the test statistic on the second half)

3: Randomly permute $\{Y_1, \ldots, Y_n\}$. Refit $\widehat{m}$ and calculate the test statistic on the permuted data (again by, e.g., training the regression on the first half of the sample and evaluating the test statistic on the second half)

4: Repeat the previous step B times to obtain $\{\widehat{\mathcal{T}}_{global}^{(1)}, \ldots, \widehat{\mathcal{T}}_{global}^{(B)}\}$.

5: Approximate the permutation p-value by

$$p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} I\left(\widehat{\mathcal{T}}_{global}^{(b)} > \widehat{\mathcal{T}}_{global}\right)\right)$$

6: **return** p

DOI: 10.1214/19-EJS1648

GoF Regression Test via MC Sampling

Algorithm 2: Goodness-of-Fit Regression Testing via MC Sampling

Input: i.i.d. sample $\mathcal{S}$ of size n from distribution with density f; reference model with density f_0; size of Monte Carlo sample n_0; number of additional Monte Carlo samples M; a regression method $\widehat{m}$

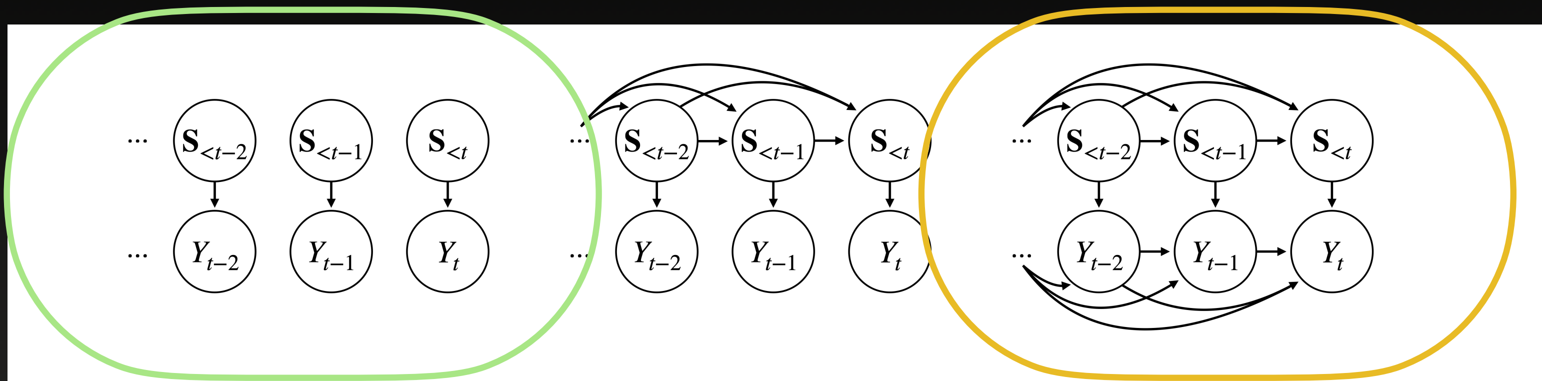
Output: p-value for testing if $f(\mathbf{x}) = f_0(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$

Let $n_{tot} = n + n_0$.
Sample $\mathcal{S}_0 = \{\mathbf{X}_1^*, \ldots, \mathbf{X}_{n_0}^*\}$ from f_0.
Define an augmented sample $\{\mathbf{X}_i, Y_i\}_{i=1}^{n_{tot}}$, where $\{\mathbf{X}_i\}_{i=1}^{n_{tot}} = \mathcal{S} \cup \mathcal{S}_0$, and $Y_i = I(\mathbf{X}_i \in \mathcal{S})$.
Calculate the global test statistic $\widehat{\mathcal{T}}$ in Equation 10.
for $b \in \{1, \ldots, B\}$ **do**
    Sample $\mathcal{S}^{(b)} = \{\mathbf{X}_1^{(b)}, \ldots, \mathbf{X}_n^{(b)}\}$ from f, under the null hypothesis $H_0 : f = f_0$.
    Sample $\mathcal{S}_0^{(b)} = \{\mathbf{X}_1^{*(b)}, \ldots, \mathbf{X}_{n_0}^{*(b)}\}$ from f_0.
    Define a new augmented sample $\{\mathbf{X}_i, Y_i\}_{i=1}^{n_{tot}}$, where $\{\mathbf{X}_i\}_{i=1}^{n} = \mathcal{S}^{(b)} \cup \mathcal{S}_0^{(b)}$, and $Y_i = I(\mathbf{X}_i \in \mathcal{S}^{(b)})$.
    Refit $\widehat{m}$ and calculate the test statistic on the new augmented sample to obtain $\widehat{\mathcal{T}}^{(b)}$ from the null distribution $f = f_0$.
end for
Compute the MC p-value by $p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} I(\widehat{\mathcal{T}}^{(b)} > \widehat{\mathcal{T}})\right)$.
return p

https://http://proceedings.mlr.press/v108/dalmasso20a.html/abs/1905.11505

Detecting Distributional Differences in Labeled "DID" Sequences of Images

[Ref: McNeely et al AOAS 2023]



(a) Setting A: $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$ with no temporal dependence between pairs $(\mathbf{S}_{<t}, Y_t)$ for different t.

(b) Setting B: Y_t conditionally independent of Y_{t-1} given $\mathbf{S}_{<t}$; $\mathbf{S}_{<t}$ is autocorrelated.

(c) Setting C: Y_t conditionally dependent on Y_{t-1} given $\mathbf{S}_{<t}$; $\mathbf{S}_{<t}$ and Y_t are each autocorrelated.

Fig 2: **Dependence settings.** Directed acyclic graphs (DAGs) illustrating the three dependence structures we explore. Note that each variable $\mathbf{S}_{<t}$ can itself represent a temporal sequence of high-dimensional functions or images, as in Figure 3.