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

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 7: Examples of galaxies from (a) the low-SFR sample $S_0$ versus (b) the high-SFR sample $S_1$.

[Figure credit: Dalmasso et al, 2019]
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.
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
Ex 2A: Quality Assurance of Simulations by Ensemble Consistency Testing (ECT) for Climate Models
Ex 2B: Validation of Approximate Likelihood Models \( \hat{L}(x; \theta) \) 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)

Global and local two-sample tests via regression

Ilmun Kim, Ann B. Lee, and Jing Lei

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 Lin⁴

HECT: High-Dimensional Ensemble Consistency Testing for Climate Models

Niccolò Dalmasso¹, Galen Vincent¹, Dorit Hammerling², Ann B. Lee¹

¹Department of Statistics & Data Science, Carnegie Mellon University
²Department of Applied Mathematics and Statistics, Colorado School of Mines

Basic Setting: Two-Sample Testing

Suppose we have two samples:

\[ X_1^0, \ldots, X_{n_0}^0 \sim P_0 \quad \text{and} \quad X_1^1, \ldots, X_{n_1}^1 \sim P_1 \]

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

\[ H_0 : f(x|Y = 0) = f(x|Y = 1), \quad \text{for all} \ x \in \mathcal{X} \]
1. We are looking for **regions** in the state space where the two populations have significantly different densities.
2. We are searching for differences in high-dimensional space (e.g., each data point could represent an image or a sequence of images)
3. We are Targeting **Model Independent** Searches

Source: Ben Nachman, “Landscape of model-independent Searches” PhyStat-Anomalies 2022
Suppose we have two samples:

\[ X_1^0, \ldots, X_{n_0}^0 \sim P_0 \quad \text{and} \quad X_1^1, \ldots, X_{n_1}^1 \sim P_1 \]

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

\[ H_0 : f(x|Y = 0) = f(x|Y = 1), \quad \text{for all } x \in \mathcal{X} \]
Two-Sample Test via Regression

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

Suppose we have two samples:

\[ X_1^0, \ldots, X_{n_0}^0 \sim P_0 \quad \text{and} \quad X_1^1, \ldots, X_{n_1}^1 \sim P_1 \]

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

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

By Bayes rule, this is equivalent to testing

\[ H_0 : \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 1) \quad \text{for all } x \in \mathcal{X} \]
Convert 2-samples testing to a regression problem

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 \( m(x) = \mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}) \).

Let \( \hat{m}(\mathbf{x}) \) 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 \( \hat{n}_1 = \frac{1}{n} \sum_{i=1}^n I(Y_i = 1) \).
Convert 2-samples testing to a regression problem

Our null and alternative hypotheses are

\[ H_0 : \ P(Y = 1 | \mathbf{X} = \mathbf{x}) = P(Y = 1), \ for \ all \ \mathbf{x} \in \mathcal{X} \]
\[ H_1 : \ P(Y = 1 | \mathbf{X} = \mathbf{x}) \neq P(Y = 1), \ for \ some \ \mathbf{x} \in \mathcal{X} \]

Define the regression function \( m(\mathbf{x}) = P(Y = 1 | \mathbf{X} = \mathbf{x}) \). Let \( \hat{m}(\mathbf{x}) \) 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 \( \hat{\pi}_1 = \frac{1}{n} \sum_{i=1}^{n} I(Y_i = 1) \).

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

\[ \lambda(\mathbf{x}) := \hat{m}(\mathbf{x}) - \hat{\pi}_1 \]

We define our global test statistic as

\[ \hat{T} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{x} \in \mathcal{V}} \lambda(\mathbf{x})^2 \]
Compute p-values by permutations (b=1,…B)

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?

- 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 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?

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

\[ \widehat{T} = \frac{1}{|\mathcal{V}|} \sum_{x \in \mathcal{V}} (\widehat{m}(x) - \widehat{\pi_1})^2. \]

- 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 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 \( \hat{m}(x) \) is a linear smoother satisfying

\[
\sup_{m \in M} \mathbf{E} \int_\chi (\hat{m}(x) - m(x))^2 \, dP_X(x) \leq C_0 \delta_n, \tag{2}
\]

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

- **Type I error:** \( P_0 \left( \hat{T}' \geq t^*_\alpha \right) \leq \alpha \), and

- **Type II error:** \( \sup_{m \in M(C_1 \delta_n)} P_1 \left( \hat{T}' < t^*_\alpha \right) \leq \beta \)

against the class of alternatives \( M(C_1 \delta_n) \) defined by

\[
\left\{ m \in M : \int_\chi (m(x) - \pi_1)^2 \, dP_X(x) \geq C_1 \delta_n \right\},
\]

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

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

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

Define the regression function $m(x) = P(Y = 1|X = x)$. Let $\hat{m}(x)$ be an estimate of $m(x)$ based on a train set $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n \subset \mathcal{X}$. Let $\hat{\pi}_1 = \frac{1}{n} \sum_{i=1}^n I(Y_i = 1)$. 
What about Classification Accuracy Tests?

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

\[ H_0 : f(x|Y = 0) = f(x|Y = 1) \]

\[ H_0 : \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 1) \]

\[ \hat{T} = \frac{1}{|\mathcal{V}|} \sum_{x_i \in \mathcal{V}} (\hat{m}(x_i) - \hat{\pi}_1)^2 \]

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

\[ H_0 : \mathbb{P}(\{h(X) \neq Y\}) = \frac{1}{2} \]

\[ \hat{T}_{acc} = \frac{1}{|\mathcal{V}|} \sum_{x_i \in \mathcal{V}} \mathbb{I}(\hat{h}(X_i) \neq Y_i) \]
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:

\[
X|Y = 0 \sim N(\mu_0, \Sigma), \quad X|Y = 1 \sim N(\mu_1, \Sigma)
\]

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

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 $n_0=n_1=100$ ($D=5, 20$): 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), $\hat{T}_{LDA}$ (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?

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

\[ \hat{T} = \frac{1}{|\mathcal{V}|} \sum_{x \in \mathcal{V}} (\hat{m}(x) - \hat{\pi}_1)^2. \]

- 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)

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?

Figure 7: Examples of galaxies from (a) the low-SFR sample $S_0$ versus (b) the high-SFR sample $S_1$. 
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
  \[ \text{\Rightarrow binned peak counts } x \in \mathbb{N}^7 \]

- 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.

Test $H_0 : \widehat{\mathcal{L}}(\mathbf{x}; \theta) = \mathcal{L}(\mathbf{x}; \theta)$ for every $\mathbf{x} \in \mathcal{X}$ and $\theta \in \Theta$

versus $H_1 : \widehat{\mathcal{L}}(\mathbf{x}; \theta) \neq \mathcal{L}(\mathbf{x}; \theta)$ for some $\mathbf{x} \in \mathcal{X}$ and $\theta \in \Theta$

Our framework can help answer:

- IF one needs to improve the emulator model
- WHERE in parameter space $\Theta$ the fit might be poor
- HOW the distributions of emulated and high-fidelity simulated data may differ in observable space $\mathcal{X}$
We can use the regression test to validate approximate likelihood (emulator) models for computationally intensive simulations.

\[
\begin{align*}
\text{Test} & \quad H_0 : \hat{L}(x; \theta) = L(x; \theta) \text{ for every } x \in \mathcal{X} \text{ and } \theta \in \Theta \\
\text{versus} & \quad H_1 : \hat{L}(x; \theta) \neq L(x; \theta) \text{ for some } x \in \mathcal{X} \text{ and } \theta \in \Theta
\end{align*}
\]

Our framework can help answer:

- If one needs to improve the emulator model
- Where in parameter space \( \Theta \) the fit might be poor
- How the distributions of emulated and high-fidelity simulated data may differ in observable space \( \mathcal{X} \)
Two-Step Procedure: 2-Sample Test at Each Parameter.

(IF) Global test of Uniformity of Local p-Values

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...

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

\[ \hat{T} = \frac{1}{|\mathcal{V}|} \sum_{x \in \mathcal{V}} (\hat{m}(x) - \hat{\pi}_1)^2. \]

- the difference \(|\hat{m}(x) - \hat{\pi}_1|\) provides information on how well the emulator fits the simulator in feature space: we can test whether \(|\hat{m}(x) - \hat{\pi}_1|\) 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?

Test \( H_0 : \hat{L}(x; \theta) = L(x; \theta) \) for every \( x \in \mathcal{X} \) and \( \theta \in \Theta \)

versus \( H_1 : \hat{L}(x; \theta) \neq L(x; \theta) \) for some \( x \in \mathcal{X} \) and \( \theta \in \Theta \)

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?

\[ S = \{ (\theta_1, X_1), (\theta_2, X_2), \ldots, (\theta_n, X_n) \} \], where \( \theta \sim r(\theta) \), \( X|\theta \sim \mathcal{L}(x; \theta) \)

\[ S_e = \{ (\theta'_1, X'_1), (\theta'_2, X'_2), \ldots, (\theta'_{n_e}, X'_{n_e}) \} \], where \( \theta' \sim r'(\theta) \), \( X|\theta' \sim \hat{\mathcal{L}}(x; \theta') \)

\[ H_0 : \hat{\mathcal{L}}(x; \theta) = \mathcal{L}(x; \theta) \], for every \( x \in \mathcal{X} \) at fixed \( \theta \in \Theta \)

Image credit: Nic Dalmasso
Regress Y on Both x and θ

\[ H_0(\theta) : \hat{\mathcal{L}}(x; \theta) = \mathcal{L}(x; \theta), \quad \text{for every } x \in \mathcal{X} \text{ at fixed } \theta \in \Theta \]

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

\[ \hat{T} = \frac{1}{|\mathcal{V}|} \sum_{x \in \mathcal{V}} (\hat{m}(x, \theta) - \hat{\pi}_1(\theta))^2. \]
Q2. Is the regression 2-sample test valid locally?

\[ H_0 : f_0(x) = f_1(x), \text{ for all } x \in \mathcal{X} \]
\[ H_1 : f_0(x) \neq f_1(x), \text{ for some } x \in \mathcal{X} \]

\[ \hat{T} = \frac{1}{|\mathcal{V}|} \sum_{x \in \mathcal{V}} (\hat{m}(x) - \hat{\pi}_1)^2. \]

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

\[ \hat{T}_{local}(x) = \hat{m}(x) - \hat{\pi}_1 \]
Approximate Validity of Local p-Values

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

Theorem 1. Under the null hypothesis

\[
H^\epsilon_0(x) : f_0(x') = f_1(x') \text{ for all } x' \in B(x; \epsilon)
\]

and under Assumption 1, for any \( 0 < \alpha < 1 \)

\[
\lim_{{B \to \infty}} P(p_{local}(x) \leq \alpha) = \alpha.
\]
Q3. Can we increase the power by MC sampling?

Suppose we have i.i.d. sample

\[ X_1, \ldots, X_n \sim P \]

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 \( f_0 \).

\[ H_0 : f(x) = f_0(x), \text{ for all } x \in \mathcal{X} \]
Suppose we have i.i.d. sample

\[ X_1, \ldots, X_n \sim P \]

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 \( f_0 \).

\[ H_0 : f(x) = f_0(x), \text{ for all } x \in \mathcal{X} \]

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:

Take Away

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

\[ X_1, \ldots, X_m \sim F \quad \text{and} \quad X_1^*, \ldots, X_n^* \sim F^* \]

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 is 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 [162].
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 [12].
EXTRA SLIDES
2-Sample Regression Test via Permutations

Algorithm 1: Two-Sample Regression Testing via Permutations

**Input:** two i.i.d. samples $S_0$ and $S_1$ 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 $f_0(x) = f_1(x)$ for every $x \in \mathcal{X}$

1. Define an augmented sample $\{X_i, Y_i\}_{i=1}^n$, where $\{X_i\}_{i=1}^n = S_0 \cup S_1$, and $Y_i = I(X_i \in S_1)$.
2. Calculate the global test statistic $\widehat{r}_{global}$ (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 $\{\widehat{r}_{global}^{(1)}, \ldots, \widehat{r}_{global}^{(B)}\}$.
5. Approximate the permutation $p$-value by

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

6. **return** $p$
Algorithm 2: Goodness-of-Fit Regression Testing via MC Sampling

**Input:** i.i.d. sample $S$ 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 $\hat{m}$

**Output:** $p$-value for testing if $f(x) = f_0(x)$ for every $x \in X$

1. Let $n_{tot} = n + n_0$.
2. Sample $S_0 = \{X_1^*, \ldots, X_{n_0}^*\}$ from $f_0$.
3. Define an augmented sample $\{X_i, Y_i\}_{i=1}^{n_{tot}}$, where $\{X_i\}_{i=1}^{n_{tot}} = S \cup S_0$, and $Y_i = I(X_i \in S)$.
4. Calculate the global test statistic $\hat{T}$ in Equation [10].
5. for $b \in \{1, \ldots, B\}$ do
6. Sample $S^{(b)} = \{X_1^{(b)}, \ldots, X_n^{(b)}\}$ from $f$, under the null hypothesis $H_0: f = f_0$.
7. Sample $S_0^{(b)} = \{X_1^{*(b)}, \ldots, X_{n_0}^{*(b)}\}$ from $f_0$.
8. Define a new augmented sample $\{X_i, Y_i\}_{i=1}^{n_{tot}}$, where $\{X_i\}_{i=1}^{n} = S^{(b)} \cup S_0^{(b)}$, and $Y_i = I(X_i \in S^{(b)})$.
9. Refit $\hat{m}$ and calculate the test statistic on the new augmented sample to obtain $\hat{T}^{(b)}$ from the null distribution $f = f_0$.
10. end for
11. Compute the MC $p$-value by $p = \frac{1}{B+1} \left(1 + \sum_{b=1}^{B} I(\hat{T}^{(b)} > \hat{T})\right)$.
12. return $p$

Detecting Distributional Differences in Labeled “DID” Sequences of Images

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

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

(c) Setting C: \(Y_t\) conditionally dependent on \(Y_{t-1}\) given \(S_{<t}\); \(S_{<t}\) 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 \(S_{<t}\) can itself represent a temporal sequence of high-dimensional functions or images, as in Figure 3.