Towards Causal Discovery with Statistical Guarantees

S. Prakash[1*], F. Xia[2], E. Erosheva[1]
1 University of Washington, Seattle, Washington; *shreyap1@uw.edu
2 University of California San Francisco, San Francisco, California

INTRODUCTION

- Functional causal discovery methods aim to infer causal direction from the data given certain distributional assumptions.
- There exists no diagnostic tool to assess assumption violations and their impact on detecting the causal direction.
- We propose the Causal Direction Detection Rate (CDDR) diagnostic to address this need.
- Key observation: Impacts of assumption violations on inferred directionality depends on sample size:
  - Small sample sizes may lead to indeterminate results due to insufficient information about causal directionality.
  - Large sample sizes with subtle assumption violations may obscure detecting the causal direction signal.

METHODS

Our proposed Causal Direction Detection Rate (CDDR) diagnostic

- Measures ***uncertainty*** in causal direction as a function of sample size
- ***Applicable*** to any functional causal discovery method
- Is ***consistent*** and exhibits ***CLT*** properties under some assumptions

Causal Discovery Methods

- Linear Non-Gaussian Acyclic Model (LiNGAM)[1] and the Test-based Approach
- Additive Noise Models (ANM)
- Post-Nonlinear Causal Model (PNL)

Test-based

- Uses hypothesis tests to determine the causal direction:

$$H = \begin{cases} H_Y^0: X \to Y, H_Y^1: Y \to X \\ H_X^0: Y \to X, H_X^1: X \to Y \end{cases} \Rightarrow H^* = \begin{cases} H_Y^0: X \perp \epsilon, H_Y^1: X, \epsilon \text{ dependent} \\ H_X^0: Y \perp \delta, H_X^1: Y, \delta \text{ dependent} \end{cases}$$

- Compares p-value estimated from H^* to significance level

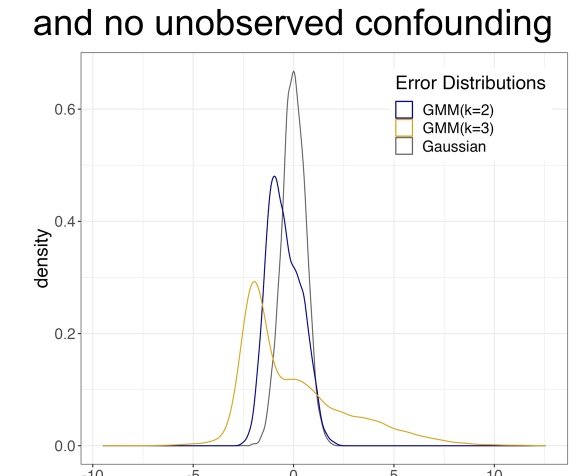
- Assumes
  1. Linearity
  2. Non-Gaussianity
  3. I.I.D data
  4. Acyclicity
  5. No unobserved confounding
- Uses linear regression to decide between
  1. $X \to Y \Rightarrow Y = \beta X + \epsilon, X \perp \epsilon$
  2. $Y \to X \Rightarrow X = \gamma Y + \eta, Y \perp \eta$

- Compares "test-statistics" (e.g. mutual information) between directions

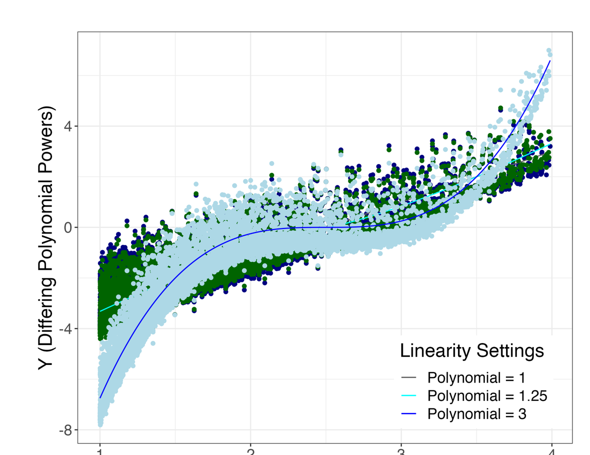
LiNGAM

SIMULATION SETUP

- Demonstrate the CDDR diagnostic applied to LiNGAM and test-based approach for varying levels of linearity and non-Gaussianity assumption violations
- Correct direction is $X \to Y$, $N = 10000$, subsample size ranges from 20 to 1699
- CDDR diagnostic interpretation assumes consistent direction, acyclicity, i.i.d data, and no unobserved confounding



Simulation settings for varying levels of non-Gaussianity. GMM(k=3) corresponds to non-Gaussian. GMM(k=2) corresponds to slightly non-Gaussian. Gaussian corresponds to Gaussian setting.



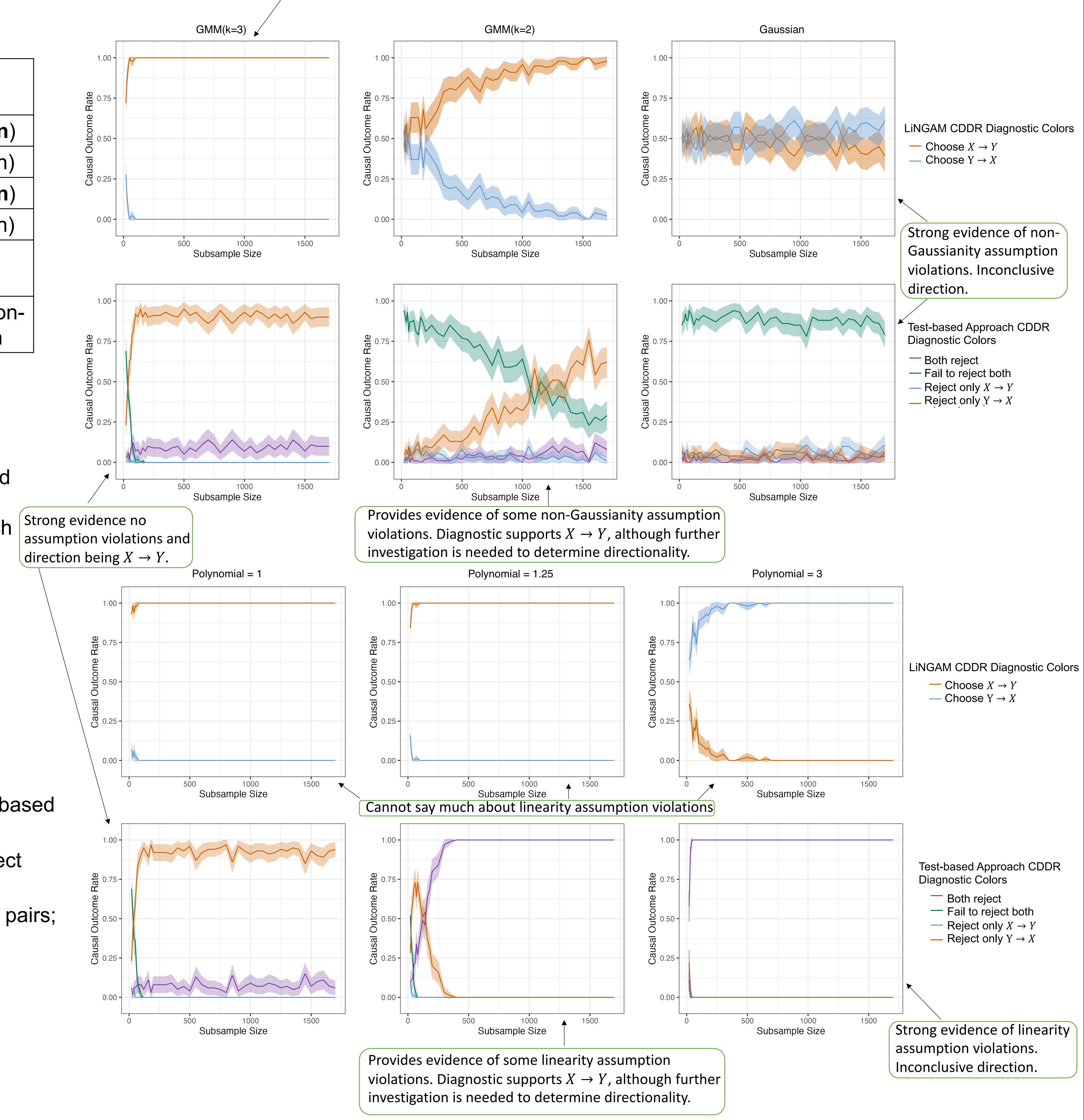
Simulation settings for varying levels of linearity. Polynomial = 1 corresponds to linear setting. Polynomial = 1.25 corresponds to slightly nonlinear. Polynomial = 3 corresponds to nonlinear.

RESULTS

CDDR Diagnostic Interpretation

Method	CDDR Diagnostic Colors	Description
LiNGAM with HSIC	orange	Detects $X \to Y$ (**correct direction**)
	blue	Detects $Y \to X$ (incorrect direction)
Test-based Approach	orange	Detects $X \to Y$ (**correct direction**)
	blue	Detects $Y \to X$ (incorrect direction)
	purple	Indicates linearity assumption violation
	green	Indicates small sample size or non-Gaussianity assumption violation

Simulation Results



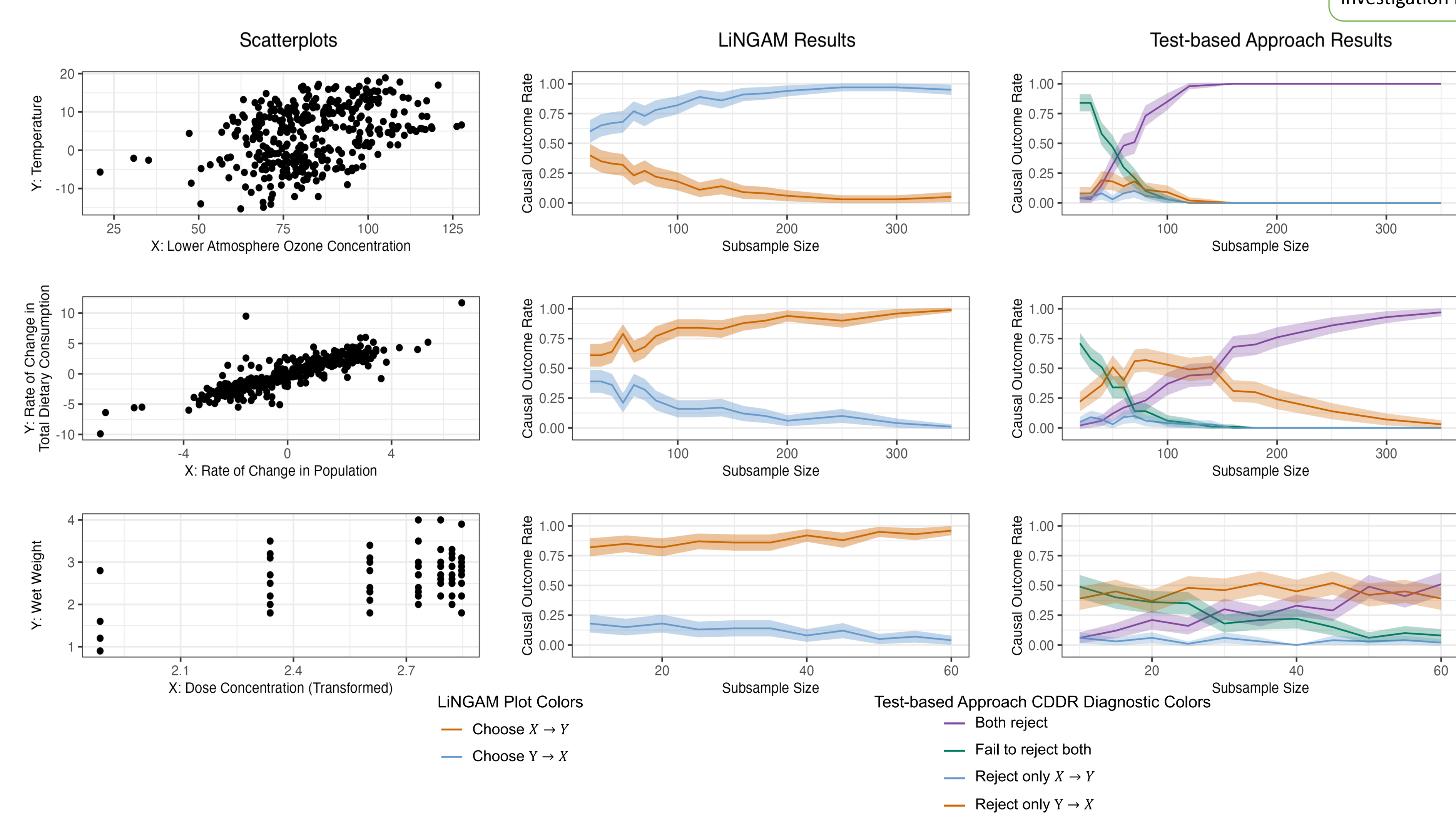
Simulation Study: Conclusions

- Non-Gaussianity assumption violations:
  - CDDR diagnostic provides information about the existence and extent of violations while providing evidence in favor of a causal direction for both LiNGAM and the Test-based Approach
- Linearity assumptions violations:
  - CDDR diagnostic for LiNGAM provides little information.
  - CDDR diagnostic for the Test-based Approach provides information about the existence and extent of violations while providing evidence in favor of a causal direction.

Examples: Real Data

- Demonstrate CDDR diagnostic applied to LiNGAM and the test-based approach on 3 real datasets where casual direction is known:
  1. Ozone and Temperature dataset[3] (from Tübingen cause-effect pairs; known direction is $\text{Temperature} \to \text{Ozone}$)
  2. Population and Food Consumption dataset[3] (from Tübingen pairs; known direction is $\text{Population} \to \text{Food Consumption}$)
  3. Rainbow Trout Dose-Response dataset[4] (known direction is $\text{Dose Concentration} \to \text{Wet Weight}$)

Real Data Results



Real Data CDDR Diagnostic: Conclusions

1. Ozone and Temperature dataset
   - LiNGAM favors incorrect direction due to assumption violations.
   - Detects moderate to severe linearity assumption violations; inconclusive direction for Test-based Approach.
2. Population and Food Consumption dataset
   - Both methods favor correct direction.
   - Evidence of linearity assumption violations with the Test-based Approach.
   - No assumption violations detected with LiNGAM.
3. Rainbow Trout Dose-Response dataset
   - Both methods support correct direction
   - With the Test-based Approach, detects minor linearity assumption violations due to the inevitable non-linearities in real data.
   - No assumption violations detected with LiNGAM

CONTRIBUTIONS

- CDDR Diagnostic: first diagnostic tool for causal discovery to evaluate assumption violations as a function of sample size.
- Applicable to any bivariate functional causal discovery method.
- CDDR diagnostic is especially effective when paired with a causal discovery method that provides more than just a deterministic direction such as our proposed Test-based Approach.

REFERENCES

1 Shimizu et al. *Journal of Machine Learning Research*. 2006.
2 Sen & Sen. *Biometrika*. 2014.
3 Mooji et al. *Journal of Machine Learning Research*. 2016.
4 Ritz et al. *PloS one*. 2015.