Conditional Independence Testing using Generative Adversarial Networks

Alexis Bellot and Mihaela van der Schaar

Motivation

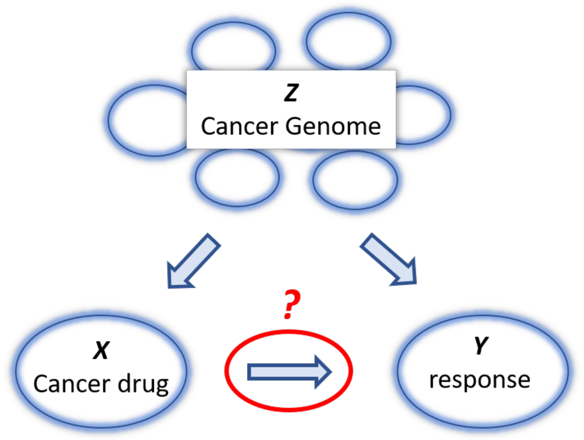
Improve decision-making in higher dimensional data

What is the question? Do variables X and Y behave independently of each other, after accounting for the effect of confounders Z?

• **Examples:**

1) Can we determine whether a new drug is directly related to the disease of interest?

2) Is a government policy effective even after accounting for other economic factors?



Introduction

Problem

Decision-making in high dimensional samples is challenging because **spurious correlations** tend to make X and Y appear independent (conditional on Z) when they are not.

Conditional independence tests

Formulate such questions as a *hypothesis testing* problem:

$$\mathcal{H}_0 : X \perp\!\!\!\perp Y | Z \quad \text{versus} \quad \mathcal{H}_1 : X \not\perp\!\!\!\perp Y | Z$$

- ▷ **Kernel-based** - provide only asymptotic guarantees.
- ▷ **Permutation-based** - can be hard to implement in practice because of the need to maintain the conditional distribution.
- ▷ **Parametric tests** - often impose strong structure on the data.

Impact

• **Technical Significance:** We present a test for conditional independence that relies on a *different set of assumptions*.

- ▷ We show that given only a *viable approximation to a conditional distribution* one can derive conditional independence tests that are approximately valid in *finite samples* and that have non-trivial power.

• **Practical Relevance:** Conditional independence plays a role in many causal discovery algorithms and have been applied in *all areas of science*.

- ▷ This work opens the door to principled statistical testing with *complex data* - images, text, speech etc.

Generative Conditional Independence TEsting (GCIT)

Procedure: We use the following representation under $\mathcal{H}_0$:

$$\boxed{X \perp\!\!\!\perp Y | Z \iff Pr(X|Z,Y) = Pr(X|Z) \sim q_{\mathcal{H}_0}}$$

1) Sample - Assuming access to $q_{\mathcal{H}_0}$ we can sample repeatedly $\tilde{X}$ conditioned on the observed confounders Z. Form an exchangeable sequence of generated triples $(\tilde{X}, Y, Z)$ and original data (X, Y, Z).

2) Summarise - Any function ρ of our data, chosen independently of the values of X applied to the real and generated samples preserves exchangeability. Hence the sequence,

$$\rho(X,Y,Z), \rho(\tilde{X}^{(1)},Y,Z), ..., \rho(\tilde{X}^{(M)},Y,Z)$$

is exchangeable under $\mathcal{H}_0$.

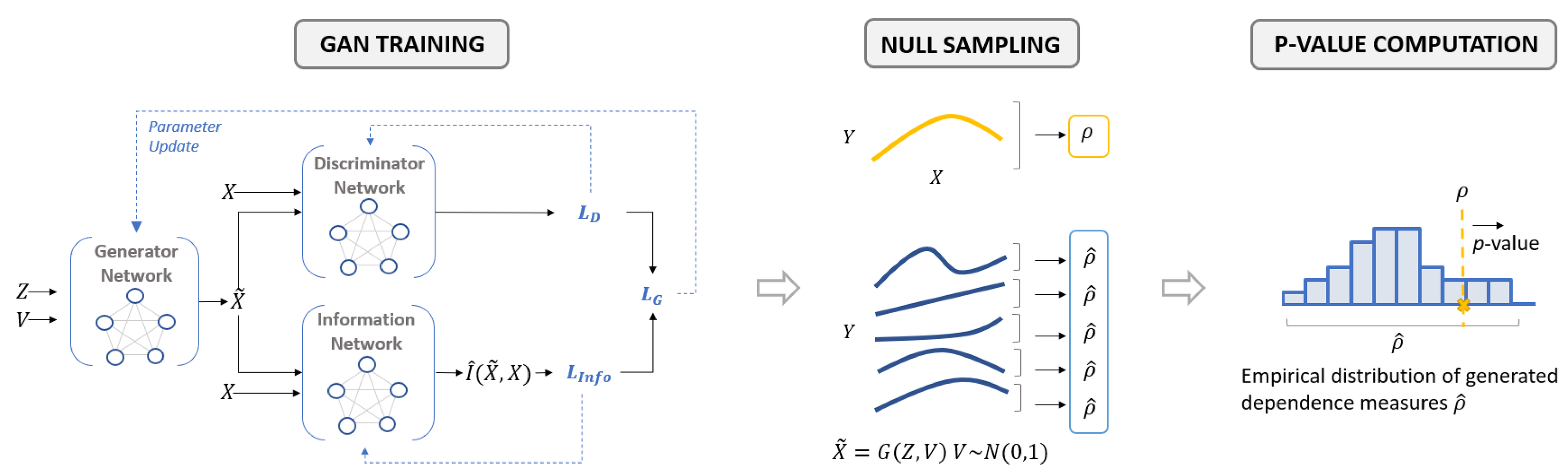
3) Compare - Decisions on the validity of $\mathcal{H}_0$ are based on the p-value. It can be approximated by comparing the generated samples with the observed sample,

$$\sum_{m=1}^{M} \mathbf{1}\{\rho(\tilde{X}^{(m)},Y,Z) \geq \rho(X,Y,Z)\}/M$$

Implementation:

Idea is to design a sampling process such as to enforce equality in distributions but otherwise encourage independence.

- ▷ We adapt GANs to minimize total variation $\mathcal{L}_G$ - *ensures valid p-values*.
- ▷ We encourage low mutual information between X and $\tilde{X}$ - *increases power in high-dimensional samples*.



Guarantees:

- ▷ **Valid p-values** - Generating conditionally independent samples with a neural network preserves *exchangeability* of input samples.
- ▷ **Upper-bound on the error in the worst case** - In practice, we can often only hope to recover an approximation to the true conditional (Theorem).
- ▷ **Weaker assumptions** - No assumptions on the data generating process posed.

Theorem *An optimal discriminator D^* minimizing $\mathcal{L}_D$ exists; and, for any statistic $\hat{\rho} = \rho(X,Y,Z)$, the excess type I error over a desired level α is bounded by $\mathcal{L}_G(D^*)$,*

$$Pr(\hat{\rho} > c_\alpha | \mathcal{H}_0) - \alpha \leq \mathcal{L}_G(D^*) \tag{1}$$

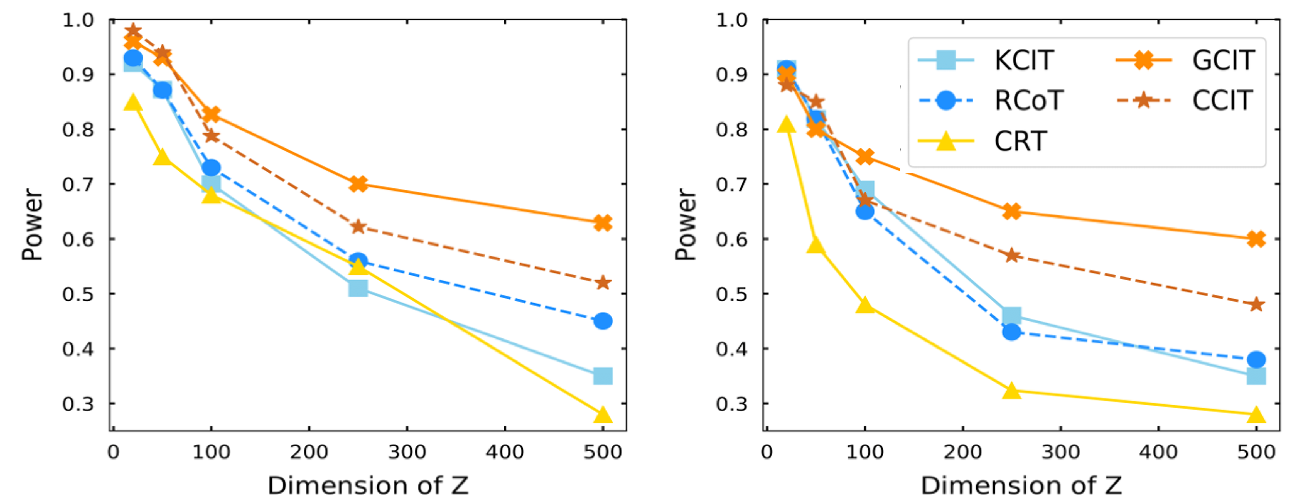
where $c_\alpha := \inf\{c \in \mathbb{R} : Pr(\hat{\rho} > c) \leq \alpha\}$ is the critical value on the test's distribution and $Pr(\hat{\rho} > c_\alpha | \mathcal{H}_0)$ is the probability of making a type I error.

Experiments

• **Synthetic simulations:** Validating on real data is hard because ground truth conditional independence is usually not known.

- ▷ Post nonlinear noise model:

$$\mathcal{H}_0 : \quad X = f(A_f Z + \epsilon_f), \quad Y = g(A_g Z + \epsilon_g)$$
$$\mathcal{H}_1 : \quad Y = h(A_h Z + \alpha X + \epsilon_h)$$



- ▷ **Left panel:** we set f, g and h to be linear before but use a Laplace distribution to generate Z and X.
- ▷ **Right panel:** We set f, g and h to be randomly sampled from $\{x^3, \tanh x, \exp(-x)\}$ and use a Gaussian distributions to generate Z and X.

• **Remarks:** We show good control of type I error in all settings, analyse hyperparameter sensitivity and provide guidelines to optimizing them in practice even without labeled data.

Acknowledgements

This work was supported by the Alan Turing Institute, the ONR and the NSF.