Testing for Conditional Independence

Outline:
(i) Hardness of CI Testing
(ii) Kernel-based Testing (KCIT)
(iii) Conditional Randomization Testing (CRT)
(iv) Classifier based Testing (CCIT)
(v) FOCI (Azadkia and Chatterjee)
(vi) Note on Tigramite (Runge et al).

(i) $\boxed{\text{Hardness of CI Testing}}$

Ref: The Hardness of CI Testing and the Generalized Covariance Measure, R. Shah and J. Peters, March 2021 (arXiv: 1804.07203v5).

Setup: (X, Y, Z) a triplet of random vectors s.t.

$$X \in \mathbb{R}^{d_x}, \quad Y \in \mathbb{R}^{d_y}, \quad Z \in \mathbb{R}^{d_z}, \quad \text{and}$$

are continuous r.vs, i.e., their induced measure is absolutely continuous w.r.t. the lebesgue measure.

$$\mathcal{E}_0 = \left\{ r(\cdot,\cdot,\cdot) : r \text{ is a valid dist. over } (X,Y,Z), \text{ and is abs. cont. w.r.t lebesgue measure} \right\}$$

For any $M > 0$, $\mathcal{E}_{0,M} = \Big\{ r(\cdot,\cdot,\cdot) : r \in \mathcal{E}_0$ and
$\text{support}(X) \subseteq [-M, M]$
$\text{supp}(Y) \subseteq [-M, M]$
$\text{supp}(Z) \subseteq [-M, M] \Big\}$.

cond. indep., continuous,

$$\mathcal{P}_0 = \{ p(\cdot,\cdot,\cdot) : p \in \mathcal{E}_0 \text{ and } p \text{ s.t. } X \perp\!\!\!\perp Y \mid Z \}$$

$$Q_0 = \mathcal{E}_0 \setminus \mathcal{P}_0$$

$$\mathcal{P}_{0,M} = \mathcal{P}_0 \cap \mathcal{E}_{0,M} \quad ; \quad Q_{0,M} = Q_0 \cap \mathcal{E}_{0,M}.$$

<u>Composite Hypothesis Testing</u>: (For any fixed $M \in (0, \infty]$)

(0) Null Hypothesis $(X \perp\!\!\!\perp Y \mid Z)$: $\{ p : p \in \mathcal{P}_{0,M} \}$

(1) Alternative $(X \not\perp\!\!\!\perp Y \mid Z)$: $\{ q : q \in Q_{0,M} \}$.

<u>Goal</u>: Given n iid samples $\{(x_i, y_i, z_i)\}_{i=1}^{n}$, we want to device a test:

$$\varphi_n : \mathbb{R}^{(d_x + d_y + d_z)n} \times [0,1] \longmapsto \{0, 1\}.$$

can be randomized, i.e., $\varphi_n(\{(x_i, y_i, z_i)\}_{i=1}^{n}, U)$

independent Uniform$[0,1]$ r.v.

Type I error

Level of Test: Test has valid level $\alpha \in (0,1)$ at sample size n if

we would like $\alpha \to 0$.

$$\sup_{p \in \mathcal{P}_{0,M}} \mathbb{P}_p(\varphi_n = 1) \leq \alpha$$

$(x_i, y_i, z_i) \sim P$, iid $i=1,\dots,n$

(i.e., this is the analog of $\mathbb{P}(\text{Test is wrong} \mid H_0)$ in binary hyp. testing.

null hypothesis.)

1 − (Type II error)

Power of Test: Test has power $\beta \in (0,1)$ at sample size n if

we would like $\beta \to 1$

$$\inf_{q \in Q_{0,M}} \mathbb{P}_q(\varphi_n = 1) \geq \beta.$$

$(x_i, y_i, z_i) \sim q$, iid $i=1,\dots,n$

(This is the analog of $\mathbb{P}(\text{Test is correct} \mid H_1)$ in binary hyp. testing).

The main hardness theorem below:

(Thm 2 in paper)

Theorem 1: Given any $n \geq 1$, $\alpha \in (0,1)$, $M \in (0,\infty]$ any any (randomized) test ψ_n for which:

$$\sup_{p \in \mathcal{P}_{0,M}} \mathbb{P}_p(\psi_n = 1) \leq \alpha$$

We have: $\sup_{q \in \mathcal{Q}_{0,M}} \mathbb{P}_q(\psi_n = 1) \leq \alpha.$

This say that there is no test for CI that can have low level and high power.

Total Variation (TV) Distance: $p, q \in \mathcal{E}_0$. Then,

$$\|p - q\|_{TV} = \sup_{A \in \mathcal{B}} \Big| \mathbb{P}_p((x,y,z) \in A) - \mathbb{P}_q((x,y,z) \in A) \Big|$$

Borel σ-algebra over $\mathbb{R}^{d_x + d_y + d_z}$

(Prop. 5)

Theorem 2: For each $M \in (0,\infty]$, $\exists\ q \in \mathcal{Q}_{0,M}$ s.t.

$$\inf_{p \in \mathcal{P}_{0,M}} \|p - q\|_{TV} \geq 1/24.$$

Theorem 2 says that there is no p satisfying $X \perp\!\!\!\perp Y \mid Z$ that is close in TV distance to a specific q (construction below).

However, for <u>any</u> q (including the one constructed above), CI testing does not have any reasonable power!

<u>Takeway:</u> TV distance is not capturing hardness of CI Testing. See discussion after Prop. 5 in paper for more details.

Instead of going through the proof, I will present an example below that lies at the heart of their proof.

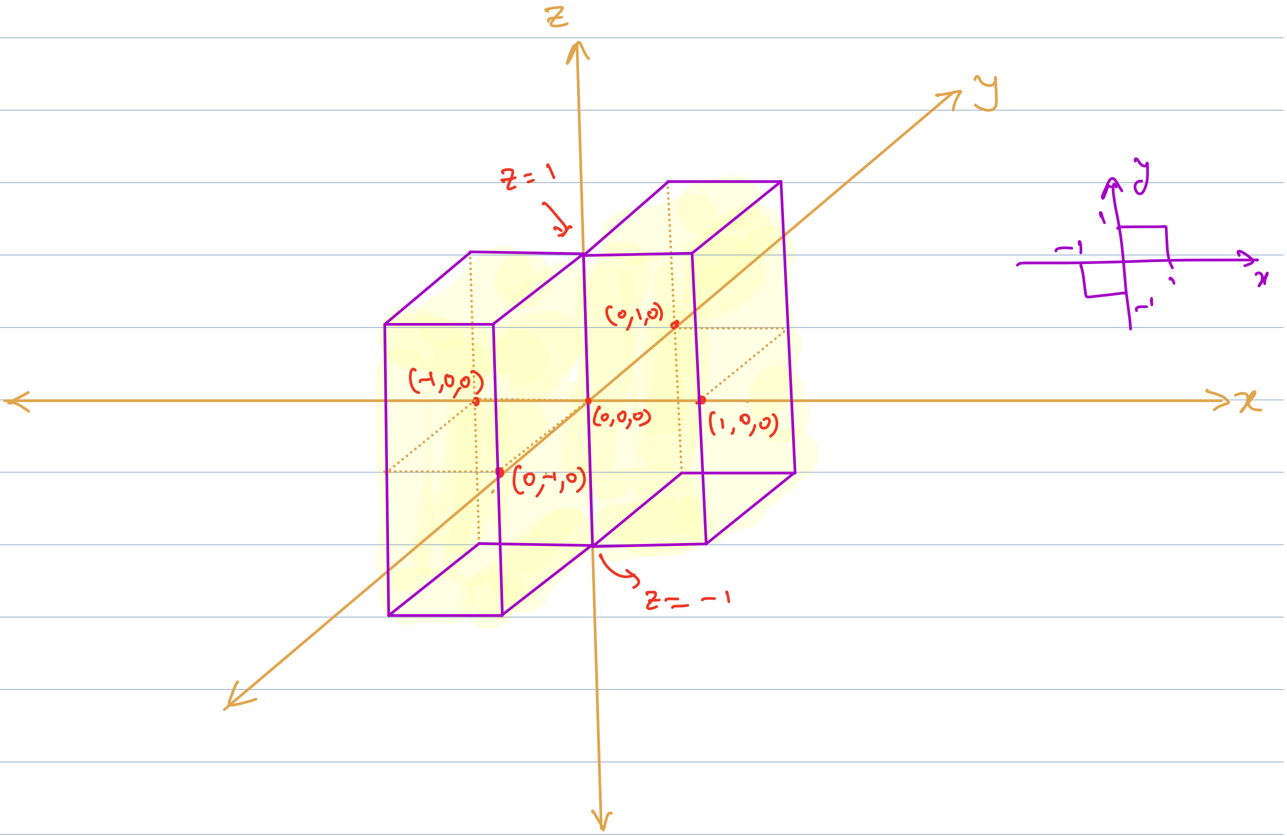
Construction for (X, Y, Z) scalar triplet over $[-1,1]^3$ to "show" Theorem 2 and plausibility of Theorem 1.

X, Y, Z each over $[-1,1]$, with their marginal dists: $X \sim \text{Uniform}[-1,1]$, $Y \sim \text{Unif}[-1,1]$ and $Z \sim \text{Unif}[-1,1]$.

q: $(X \not\perp\!\!\!\perp Y \mid Z)$

$Z \sim \text{Unif}[-1, 1]$

$(X, Y) \perp\!\!\!\perp Z$ with $(X, Y) \sim \text{Uniform}\left([0,1]^2 \cup [-1,0]^2\right)$



For the above q, observe that (X, Y) are correlated — If $X > 0$, then $Y \sim \text{Unif}[0,1]$ else if $X < 0$, then $Y \sim \text{Unif}[-1, 0]$.

$$\text{Cov}(X, Y) = 1/4. \implies X \not\perp\!\!\!\perp Y \mid Z$$

(note: $X \not\perp\!\!\!\perp Y$, and $(X, Y) \perp\!\!\!\perp Z$)

Now, we construct $\tilde{Z}$ a new r.v., with

$\tilde{Z} = f(X, Z)$ as follows: Consider the binary expansion of Z. In the m^{th} bit, delete the value, and replace it by '0' if $X \geq 0$ and '1' if $X < 0$.

ex: Suppose the sample drawn $z_i = 0.010010\ldots$ ($i.e.$, $z_i \approx 1/4 + 1/32 + \cdots \approx 0.2812\ldots$).

m^{th} bit

Then
$$\tilde{z}_i = \begin{cases} 0.010010\ldots & \text{if } x_i \geq 0 \\ 0.010011\ldots & \text{if } x_i < 0. \end{cases}$$

Now consider $(X, Y, Z, \tilde{Z})$.

Then, by choosing m to be large enough, $|z_i - \tilde{z}_i| \ll 1$ (we can make the sample-wise distance to be arbitrarily small with error $= 1/2^m$).

$\therefore$ (a) From samples, it is very hard to distinguish between (X, Y, Z) and $(X, Y, \tilde{Z})$.

(b) The above construction encodes the sign of X noiselessly in $\tilde{Z}$. Furthermore, observe that there is no additional information in X about Y

once we know it's sign, i.e.,

$$X \perp\!\!\!\perp Y \mid \tilde{Z}.$$

(c) Let p s.t. $(X, Y, \tilde{Z}) \sim p$
q s.t. $(X, Y, Z) \sim q$.

The $\|p-q\|_{TV} >$ constant > 0
$\hookrightarrow$ independent of m.

Why? Use a Borel set that focuses only on the m^{th} coordinate of the binary expansion of $Z, \tilde{Z}$, and some large enough set in $X \times Y$ that includes both (I) and (III) quadrant. Then, Z and $\tilde{Z}$ have a constant prob. of differing in the m^{th} component, and thus the TV distance is lower bounded > 0, independent of m.

$\therefore$ Now making m large, we have fixed TV distance lower bound, but the samples gets closer. Thus, it is hard to distinguish between p and q, even with many samples, so Thm 1 is plausible.

In fact, as shown in proof of Prop 2 in the paper, for **any** $p \in \mathcal{P}_{0,1}$, $\|p - q\|_{TV} > 1/24$.

See Figure 1 in paper (Hardness of CI...) for idea of proof of Thm 1.

Remark 1: This crucially relies on Z being a continuous r.v., so that we can hide the sign of X with arbitrarily small perturbation in $|z - \tilde{z}|$.

Remark 2: We saw that without assumptions, the CI composite hypothesis testing problem is untestable. The key issue is non-smoothness of $P(x,y|z)$ in z, i.e., a small perturbation in z drastically changed the joint dist of (X, Y).

In Theorem 2, we saw that we could have a large TV distance in $\|p(x,y,z) - q(x,y,z)\|_{TV}$ and still have issues, so controlling the distance between $p \in \mathcal{P}_{0,M}$ and $q \in \mathcal{Q}_{0,M}$ does not seem useful. Indeed as remarked in Neykov, Balakrishnan and Wasserman 2021, we need to control smoothness of the conditional $p(x,y|z)$. A similar condition was imposed for the CCIT algorithm to have guarantees in Sen, Suresh, Karthikeyan, Dimakis, Shakkottai 2017.

The condn would look something like:

$$\left.\| q(x,y|z) - q(x,y|z')\|_{TV} < L \|z-z'\|_2 \right\} \begin{array}{l}\text{alternate} \\ \text{hypothesis.} \\ (x \not\perp\!\!\!\perp y \mid z)\end{array}$$

$$\left.\begin{array}{l} \| p(x|z) - p(x|z')\|_{TV} < L \|z-z'\|_2 \\ \| p(y|z) - p(y|z')\|_{TV} < L \|z-z'\|_2 \end{array}\right\} \begin{array}{l}\text{null} \\ \text{hypothesis.} \\ (x \perp\!\!\!\perp y \mid z)\end{array}$$

Ref:

Minimax optimal conditional independence testing, M. Neykov, S. Balakrishnan and L. Wasserman, arXiv:2001.03039, 2021.

Model power conditional independence test, R. Sen, A.T. Suresh, K. Shanmugam, A. Dimakis and S. Shakkottai, NeurIPS 2017.

(ii) Kernel-based CI Testing

Ref: Partial associative measures and an application to qualitative regression, J. Daudin, Biometrika 1980

A kernel statistical test for independence, A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Scholkopf, A. Smola, NeurIPS 2007.

Kernel-based conditional independence test and application in causal recovery, K. Zhang, J. Peters, D. Janzing and B. Scholkopf, UAI 2011. (arXiv 1202.3775)

Warm up: Jointly Gaussian (X, Y, Z)

For jointly Gaussian r.vs, CI is equivalent to independence testing of residuals, i.e.,

$$X \perp\!\!\!\perp Y \mid Z \iff R_{XZ} \perp\!\!\!\perp R_{YZ}$$

$$\Updownarrow$$

partial correlation coefficient $\leftarrow$ $\rho_{XY.Z} = 0.$

$$R_{XZ} = X - E[X \mid Z] \quad , \quad R_{YZ} = Y - E[Y \mid Z].$$

Partial correlation coefficient

For jointly Gaussian distributions, Conditional independence can be tested through regression, using partial correlation coefficient.

Recall that for Gaussians, it suffices to check second order statistics for independence, i.e.,

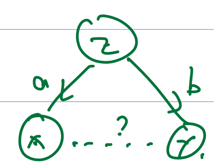
$$X \perp\!\!\!\perp Y \iff \rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = 0.$$

The partial correlation coefficient generalizes this to allow tests for CI for joint Gaussian r.v.s.

$$X \perp\!\!\!\perp Y \mid Z \iff \rho_{XY.Z} = 0.$$

We will define $\rho_{XY.Z}$ below. WLOG, all r.v.s below have zero mean $(E[X] = E[Y] = E[Z] = 0)$

Let $\quad a = \underset{\alpha}{\arg\min}\; E\left[(X - \alpha Z)^2\right]$



$\quad\quad b = \underset{\beta}{\arg\min}\; E\left[(Y - \beta Z)^2\right]$

i.e., regress X and Y each separately on Z.

$\hat{X} = E[X \mid Z] = aZ, \qquad \hat{Y} = E[Y \mid Z] = bZ$

are the associated least-squares estimators for X, Y respectively.

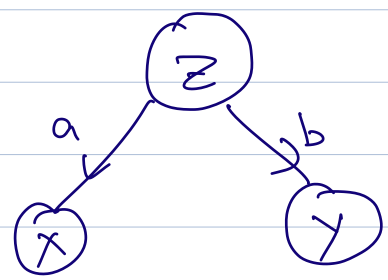
$$R_{XZ} = (X - E[X|Z]) \qquad R_{YZ} = (Y - E[Y|Z])$$

are the associated residuals.

Then, $$\rho_{XY.Z} = \rho_{R_X R_Y} = \frac{Cov(R_{XZ}, R_{YZ})}{\sqrt{Var(R_{XZ})\, Var(R_{YZ})}}$$

Then, $$\rho_{XY.Z} = 0 \iff X \perp\!\!\!\perp Y \mid Z.$$

<u>To see why</u>:



$$Z = N_1$$
$$X = aZ + N_2$$
$$Y = bZ + N_3$$

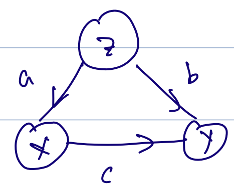
$N_1, N_2, N_3 \sim N(0,1)$ mutually independent

In the model above, it is clear from the discussions in Note 2 (d-separation) that $X \perp\!\!\!\perp Y \mid Z$.

Now, let us compute $\rho_{XY.Z}$. Here $\hat{X} = aZ$, $\hat{Y} = bZ$, $R_{XZ} = (X - aZ)$, $R_{YZ} = (Y - bZ)$.

$$E\left[R_{XZ}R_{YZ}\right] = ab - ab - ab - ab = 0.$$

To check the other way, consider a model with $c \neq 0$, and check that $\rho_{XY.Z} \neq 0$



More generally, if $(Z_1, Z_2, \dots Z_m)$ are the covariates (i.e., possible confounding r.vs), we regress X on $(Z_1, \dots Z_m)$ and also regress Y on $(Z_1, \dots, Z_m)$. Then compute $\rho_{XY.Z}$ be the correlation coefficient of the residuals.

Note: If the r.vs are not jointly Gaussian, then we cannot relate partial correlation to independence in either direction. (see example 7.9 in text)

Beyond Joint Gaussians: A general criterion for CI

(Dawid '80) A useful characterization of CI is the following: $X \in \mathbb{R}^{d_1}$, $Y \in \mathbb{R}^{d_2}$, $Z \in \mathbb{R}^{d_3}$

$f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \longrightarrow \mathbb{R}$, $g : \mathbb{R}^{d_1} \times \mathbb{R}^{d_3} \longrightarrow \mathbb{R}$.

s.t. $E\left[f(X,Z)^2\right] < \infty$, $E\left[g(Y,Z)^2\right] < \infty$

$$\tilde{f}(X,Z) = f(X,Z) - E\left[f(X,Z) \mid Z\right]$$

$$\tilde{g}(X,Z) = g(X,Z) - E\left[g(X,Z) \mid Z\right]$$

$$X \perp\!\!\!\perp Y \mid Z \iff E\left[\tilde{f}(X,Z)\, \tilde{g}(Y,Z)\right] = 0$$

$\forall$ f, g square-integrable s.t. $E\left[f(X,Z) \mid Z\right] = 0$

and $E\left[g(X,Z) \mid Z\right] = 0$ a.s.

Other equivalent criteria: (Daudin 80, Zhang et al 11)

$$\mathcal{E}_{XZ} = \left\{ \tilde{f} \text{ square integrable} : E\left[\tilde{f}(X,Z) \mid Z\right] = 0 \text{ a.s.} \right\}$$
$$\mathcal{E}_{YZ} = \left\{ \tilde{g} \text{ sq. int.} : E\left[\tilde{g}(Y,Z) \mid Z\right] = 0 \text{ a.s.} \right\}$$

$$\mathcal{E}'_{YZ} = \left\{ \tilde{g}' : \tilde{g}'(Y,Z) = g'(Y) - E\left[g'(Y) \mid Z\right], \; g' \text{ sq. int function of } Y \right\}$$

(i) $X \perp\!\!\!\perp Y \mid Z$

(ii) $E\left[\tilde{f}(X,Z)\,\tilde{g}(Y,Z)\right]=0 \quad \forall\ \tilde{f}\in\mathcal{E}_{XZ},\ \tilde{g}\in\mathcal{E}_{YZ}$

(iii) $E\left[\tilde{f}(X,Z)\,g(Y,Z)\right]=0 \quad \forall\ \tilde{f}\in\mathcal{E}_{XZ},$ g sq. int., i.e., $E\left[g(Y,Z)^2\right]<\infty$

(iv) $E\left[\tilde{f}(X,Z)\,\tilde{g}'(Y,Z)\right]=0 \quad \forall\ \tilde{f}\in\mathcal{E}_{XZ},\ \tilde{g}'\in\mathcal{E}'_{YZ}$

(v) $E\left[\tilde{f}(X,Z)\,g'(Y)\right]=0 \quad \forall\ \tilde{f}\in\mathcal{E}_{XZ},$ g' s.t. $E\left[g'(Y)^2\right]<\infty$

$\boxed{\text{Constructing a Test: KCIT}}$ (Operationalizing (iv) above)

$X\in\mathcal{X}=\mathbb{R}^{d_X}$, $Y\in\mathcal{Y}\in\mathbb{R}^{d_Y}$, $Z\in\mathcal{Z}\in\mathbb{R}^{d_Z}$.

k_X positive definite kernel over RKHS $\mathcal{H}_X$, analogous k_Y, k_Z.

Characteristic kernel: k_X is characteristic if $E_{X\sim q}[f(X)] = E_{X\sim p}[f(X)]\ \forall f\in\mathcal{H}_X \implies p=q$

Implication is that we can use a kernel to test for one of the CI criteria above (generalizes residue tests).

The KCIT approach (Zhang et. al. 2011) is the following:

<u>Samples</u>: $x = (x_1, \ldots x_n)$, $y = (y_1, \ldots y_n)$, $z = (z_1, \ldots z_n)$

$$\tilde{K}_x = \left(I - \tfrac{1}{n} 1 1^T\right) K_x \left(I - \tfrac{1}{n} 1 1^T\right)$$

centered sample kernel

$$K_x = \begin{pmatrix} \ddots & \cdots & \\ \vdots & K_x(x_i, x_j) & \\ \vdots & & \ddots \end{pmatrix}$$

$$= V_x \Lambda_x V_x^T$$

non-negative eigenvalues

$$\Psi_x = (\psi_{x,1} \quad \psi_{x,2} \quad \cdots \psi_{x,n}),$$

where $\psi_{x,i} = \sqrt{\lambda_{x,i}}\; V_{x,i}$

Similarly for y, z and $\ddot{x} = (x, z)$

Fix $\varepsilon > 0$ (regularizer parameter).

$$R_z = \varepsilon \left(\tilde{K}_z + \varepsilon I\right)^{-1}, \quad \tilde{K}_{\ddot{x}|z} = R_z \tilde{K}_{\ddot{x}} R_z$$

$$\tilde{K}_{y|z} = R_z \tilde{K}_y R_z$$

Constructing Test Statistic with finite samples, with asymptotic characterization

$$T_{CI}^{(n)} = \frac{1}{n} \operatorname{trace}\left(\tilde{K}_{\ddot{X}|Z} \cdot \tilde{K}_{Y|Z}\right)$$

Theorem (Prop 5 in KCIT): Under null hypothesis $(X \perp\!\!\!\perp Y \mid Z)$, $T_{CI}^{(n)}$ has the same asymptotic dist. (i.e. conv. in distribution) as $\check{T}_{CI}^{(n)}$, where

$$\check{T}_{CI}^{(n)} = \frac{1}{n} \sum_{k=1}^{n^2} \mathring{\lambda}_k r_k^2 \quad , \text{ where}$$

$\mathring{\lambda}_k$ are eigenvalues of $\check{w}\check{w}^T$, and $\check{w}$ defined by

$$\check{w} = \begin{pmatrix} \psi_{\ddot{x}|z} \\ \psi_{y|z} \end{pmatrix}$$

vectors defined analogous as above for ψ_x, but for the appropriate centred kernel matrices.

and $\{r_k\}$ are iid $N(0,1)$ r.v. (or equivalently, r_k^2 are iid χ^2-distributed r.v.s).

$P(T \geq t \mid H_0) = p < \alpha$

Punchline: We can operationalize the above for p-value testing. Simulate $\check{T}_{CI}^{(n)}$ and generate samples. Compute test statistic $T_{CI}^{(n)}$ from data. See where $T_{CI}^{(n)}$ falls in $\check{T}_{CI}^{(n)}$ to compute significance level to reject null hypothesis.

Remark (and caution): All the properties and associated tests are based on population (i.e., exact distribution) based characterization of conditional independence. Unlike the "usual" proofs where finite sample versions follow from "standard" concentrations, CI is quite different.

A "generic" way to break a CI test based on exact distributions/ population statistics is the following: Recall the example above, where the sign of X was embedded in the m^{th} bit, thus slightly perturbing Z to $\tilde{Z}$ through a sample-path coupling.

H_0: $\{P_m\}$: $X \perp\!\!\!\perp Y \mid Z$	P_m : encode sign of X in m^{th} bit
H_1: q : $Z \perp\!\!\!\perp (X,Y)$ with $X \not\perp\!\!\!\perp Y \mid Z$	$X \perp\!\!\!\perp Y \mid \tilde{Z}_m$

But q and P_m are very close in samples, i.e., $(x,y,z) \sim q$ and $(x,y,\tilde{z}_m) \sim P_m$ are only $1/2^m$ apart in $|z-\tilde{z}_m|$.

Goal: Given a tester with a CI characterization that is based on infinite samples/ exact-dist., we want to construct a hard counter-example that will break it for any finite n samples.

<u>Given:</u> n samples, CI tester.
<u>Adversary:</u> Picks one of q or p_n (i.e., sign of x in n^{th} bit).
<u>Problem:</u> Impossible to reliably distinguish between these with n samples, as signal strength ($1/2^n$) is buried in sampling noise. Thus, for any fixed n, tester fails.

<u>Work around:</u> Restrict null and alternatives to have Lipschitz continuous conditionals. Then, all the algorithms that we discuss based on asymptotic properties will "work".

(iii) $\boxed{\text{Conditional Randomization Test (CRT)}}$

<u>Ref:</u> Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection, E. Candes, Y. Fan, L. Janson and J. Lv, Journal of the Royal Statistical Society, Series B, 2018 (arXiv: 1610.02351).

(Motivation is feature selection in regression).

Setting: $(X_1, X_2, \cdots, X_d, Y)$

Goal: Want to check if: $X_1 \perp\!\!\!\perp Y \mid (X_2, \cdots, X_d)$.

Assume: We have exact knowledge of the joint dist. of $(X_1, \cdots, X_d)$, i.e., $p(x_1, x_2, \cdots, x_d)$, and we can generate samples from this dist.

Suppose that $T = g(X_1, X_2, \cdots, X_d, Y)$ is a test statistic

e.g. lasso

(e.g.: T is the (regularized) regression coefficient of X_1, when we regress Y on $(X_1, \cdots, X_d)$.)

Auxillary Variable: Let $X_1^* \sim p(\cdot \mid x_2, \cdots x_d)$ be a conditionally indep. r.v., i.e., given $X_2 = x_2, X_3 = x_3, \cdots, X_d = x_d$, we generate a sample from the dist. $p(z \mid x_2, \cdots, x_d)$. With this construction, observe that:

dummy variable.

$$X_1^* \mid (X_2, \cdots, X_d, Y) \overset{d}{=} X_1 \mid (X_2, \cdots, X_d)$$

The idea is that if $X_1 \perp\!\!\!\perp Y \mid (X_2, X_3, \cdots, X_d)$, then $T = g(X_1, X_2, \cdots X_d, Y)$ and $T^* = g(X_1^*, X_2, \cdots X_d, Y)$ would have identical dist. for every value of $(X_2, X_3, \cdots, X_d, Y)$. This is summarized in their lemma (Lemma 4.1 in Candes et. al. 2018) below:

Key Observation (Lemma 4.1 in paper): Under the null hypothesis,

$$T \,\Big|\, (X_2, X_3, \cdots X_d, Y) \overset{d}{=} T^* \,\Big|\, (X_2, X_3, \ldots, X_d, Y)$$

This provides an asympotic test for CI: ⓐ Simulate a large number of samples X_1^* as above; ⓑ Compute the test statistic for each; ⓒ Compute test statistic for the real X_1; ⓓ Compute p-value.

<u>Remark 1</u>: This test is based on a population property under the null hypothesis, namely the original and simulated test statistics are identically distributed. However, when dealing with finite samples, same issue that was discussed re. untestability arises. We need to control smoothness of conditionals to provide any finite sample guarentees (note that the paper

does not address finite sample guarantees).

Remark 2: Candes et al 2018 provide an alternate approach based on knockoff-X random variables, that is computationally much lighter (in the CRT, we need to generate the test statistic for each sample). The power with knockoffs is somewhat worse; please read their paper for details.

$\boxed{\text{Classifier-based Conditional Independence Test}}$

Ref: Model-powered Conditional Independence Test, R. Sen, A. T. Suresh, K. Shanmugam, A. Dimakis, S. Shakkottai, NeurIPS 2017. (arXiv: 1709.06138).

Setting: (X, Y, Z), $X \in \mathbb{R}^{d_x}$, $Y \in \mathbb{R}^{d_y}$, $Z \in \mathbb{R}^{d_z}$

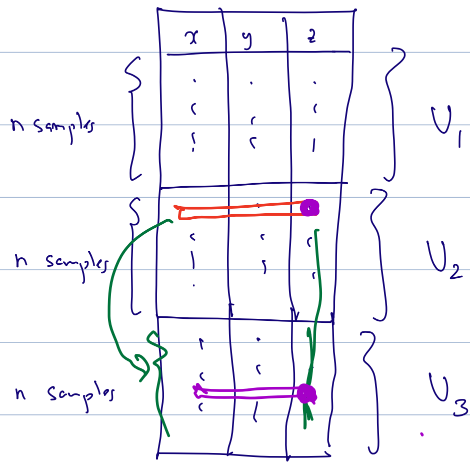
$H_0: X \perp\!\!\!\perp Y \mid Z$ Given: $3n$ iid samples

$H_1: X \not\perp\!\!\!\perp Y \mid Z$. $\{(x_i, y_i, z_i)\}_{i=1}^{3n} \sim p(x,y,z)$

Under H_0: $p(x,y,z) = p^{CI}(x,y,z) \triangleq p(x|z)\, p(y|z)\, p(z)$

Assumptions: Z a continuous r.v. that satisfies smoothness in both $p(z)$ (marginal) and conditional $P(y|z)$.
(characterized through max eigenvalue of fisher information matrix of y w.r.t z)

Algorithm:



① Partition into 3 groups of n-samples each.

② <u>Nearest neighbour bootstrap:</u>

ⓐ For each $u = (x, y, z) \in U_2$, find $(x', y', z') \in U_3$ s.t. $z' \in 1\text{-NN}(z)$

nearest neighbor w.r.t. l_2-norm.

ⓑ Construct $u' = (x, y', z)$ from above step.

(Note that if $z = z'$, then $u' \sim p(x|z)\, p(y|z)\, p(z)$ irrespective of whether H_0 is true or H_1 is true).

Let $\phi(x,y,z)$ s.t. $U' \sim \phi(x,y,z)$.

$$p^{CI}(x,y,z) = p(x|z)\, p(y|z)\, p(z).$$

Note $\phi \approx p^{CI}$ (Theorem 1 in Sen et. al. characterizes TV distance between ϕ and p^{CI} as roughly $\left(\frac{1}{n^{1/2d_z}}\right)$).

So far: * n iid samples $(x,y,z) \sim p$ (labeled U_1)

* n samples $(x,y,z) \sim \phi$ (labeled U_2').

Note:

(i) Samples in U_2' are not iid, because of the 1-NN process, that dips into the same pool U_3 for each $(x,y,z) \in U_2$ to construct (x,y',z).

(ii) Samples in U_2' are only approximately CI, i.e., $\phi(x,y,z) \approx p(x|z)\, p(y|z)\, p(z)$.

③ Train a classifier: Dataset: $U_1 \sim p(x,y,z)$
$U_2' \sim \phi(x,y,z)$

ⓐ label samples in U_1 as '1' (n samples)
samples in U_2' as '0' (n nearly-CI samples)

Intuition: $u \in U_1$, $u' \in U_2'$ are almost identically distributed under H_0 ($X \perp\!\!\!\perp Y \mid Z$). However $u \in U_1$ and $u' \in U_2'$ have different dist. under H_1 (thus, note that this property is again a population property, and thus we need smoothness of conditionals to exploit this property.)

ⓑ Split $D = (\{U_1 : \overset{\text{label}}{1}\} \cup \{U_2' : \overset{\text{label}}{0}\})$ into train and test (D_{tr} and D_{test}), with $|D_{tr}| = |D_{test}| = n$.

Using D_{tr}, train a binary classifier, s.t. the classification function class is rich enough (formally, the risk of the best classifier from this class is close to that under the Baye's optimal classifier, see Thm 2 in Sen et. al.).

ⓒ Using this classifier, evaluate D_{test} using

ERM, with risk function: $R_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{g(u) \neq \ell\}$.

classifier

test sample

label

Result (Implication of Thm 2 in Sen et. al.):

Under $\mathcal{H}_0$ $(X \perp\!\!\!\perp Y \mid Z)$: $R_n \approx 0.5$

with high enough prob.

Under $\mathcal{H}_1$ $(X \not\perp\!\!\!\perp Y \mid Z)$: $R_n \leq 0.5\gamma + o(n)$

$d_{TV}(p, p^{CI}) \geq 1 - \gamma.$