# Conditional Independence Test

Xiaoliu Wu, James Sharpnack

Department of Statistics
UC Davis

# Table of Contents

1. Literature Review
   - Background
   - Problem Formalization
   - Hardness of the Conditional Independence (CI) Test
2. Proposed Method and Simulation
   - Review of Some Existing Methods
   - Description of Our Method
   - Simulation
   - Myth: How to get the P-value?
   - Future Work

# Literature Review

# Background

- CI problems appear in many context of statistics and causal inference e.g.:
  - Shoe size $\perp\!\!\!\perp$ Vocabulary size $|$ Age ?
  - Drug administration $\perp\!\!\!\perp$ Cure of a disease $|$ Patients' background ?
- Many previous works employ Gaussian assumption [1] or graphical models [2].
  - Gaussianity?
  - How to obtain the graph?
- Aim to develop a model-based method directly test the hypothesis using neural nets.

# Problem Formalization

- $X, Y$ are binary random variables (i.e. indicator of drug use).
- $Z$ is a (continous or discrete) vector such as an image or vectorized text data, which consists of possible confounding variables.
- We want to test if $X$ and $Y$ are indpendent given $Z$.
- $X, Y \in \{-1, 1\}$ and $Z \in \mathcal{Z}$.

$$
\begin{aligned}
H_0 &: X \perp\!\!\!\perp Y | Z \\
H_1 &: X \not\!\perp\!\!\!\perp Y | Z
\end{aligned}
$$

# Hardness of the CI Test

- Shah and Peters [3] first showed that for continuous $X$, $Y$ and $Z$, no test with valid level can have power against the alternative hypothesis.
- Formally, Let $\mathcal{E}_{0,M}$ be the set of distributions for $(X, Y, Z)$ on $\mathbb{R}^{d_X+d_Y+d_Z}$ and the support is contained within an $L_\infty$ ball of radius $M$. Define $\mathcal{P}_{0,M} \subset \mathcal{E}_{0,M}$ be the set of distributions which are CI $(X \perp\!\!\!\perp Y|Z)$ and $\mathcal{Q}_{0,M} = \mathcal{E}_{0,M} \setminus \mathcal{P}_{0,M}$. Let $d = d_X + d_Y + d_Z$.

**Continuous No-Free-Lunch Theorem.**

Consider the testing problem where $H_0 : P_{X,Y,Z} \in \mathcal{P}_{0,M}$ $H_1 : P_{X,Y,Z} \in \mathcal{Q}_{0,M}$. Given any $n \in N$, $\alpha \in (0,1)$, $M \in (0,\infty]$ and a potentially randomized test $\phi_n : \mathbf{R}^{nd} \times [0,1] \to \{0,1\}$, that has valid level $\alpha$ for the null hypothesis $\mathcal{P}_{0,M}$, we have that $P_Q(\phi_n = 1) \leq \alpha$ for all $Q \in \mathcal{Q}_{0,M}$.

# Hardness of the Conditional Independence (CI) Test

- Neykov et al. [1] presented a "simpler" proof of the Continuous No-Free-Lunch Theorem and proved the theorem for the cases when $Z$ is continuous and $X$, $Y$ are discrete.
- The root of the hardness is the continuity of $Z$.
- The denseness of $\mathcal{P}_{0,M}$ shown in their paper provides intuition of the hardness.

# Hardness of the Conditional Independence (CI) Test

## Denseness of $\mathcal{P}_{0,M}$

### Definition: Wasserstein Distance between Probability Measures.

Let $P_p(\mathbb{R}_d)$ denote the set of measures on $(\mathbb{R}_d, |\Delta|_2)$,For two probability measures, $P$ and $Q$ in $P_p(\mathbb{R}_d)$ the $p$th Wasserstein distance between $P$ and $Q$ is defined as

$$W_p(P,Q) = \left( \inf_{\gamma \in \Gamma(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} ||x - y||_2^p d\gamma(P,Q) \right)^{\frac{1}{p}}$$

where $\Gamma(P,Q))$ is set of all couplings between the measures P and Q, i.e., all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$, with marginals $P$ and $Q$.

### Lemma: Denseness of $\mathcal{P}_{0,M}$.

Take any distribution $Q \in \mathcal{E}_{0,M}$ for some $M > 0$. Then for any $p \geq 1$ and any $\epsilon > 0$, there exists a distribution $P \in \mathcal{P}_{0,M}$ such that $W_p(P,Q) \leq \epsilon$ .

# Hardness of the Conditional Independence (CI) Test

## Sketch of the Proof of Denseness of $\mathcal{P}_{0,M}$ in 1$d$

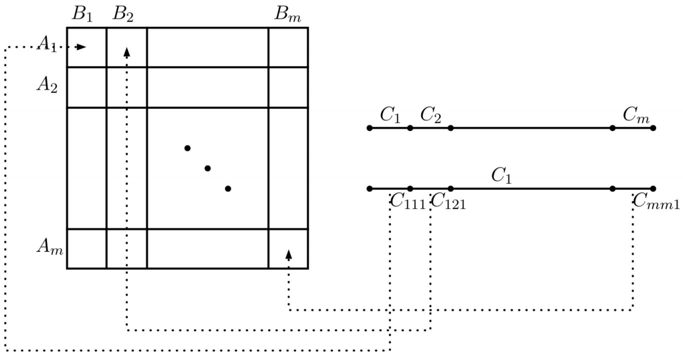


Figure 1: This schematic describes the construction of $Q$ from $P$. $[-M, M]$ is divided in intervals $\{A_1, \ldots, A_m\}$, $\{B_1, \ldots, B_m\}$ and $\{C_1, \ldots, C_m\}$. Next each interval $C_k$ is sub-divided into $m^2$ smaller sub-intervals. The interval $C_1$ is displayed along with its sub-divisions of $C_{ij1}$ for $i, j \in [m]$. Each little interval $C_{ij1}$ corresponds to a pair $(A_i, B_j)$ or equivalently to a cell $A_i \times B_j$ in $[-M, M]^2$.

# Hardness of the Conditional Independence (CI) Test

## Sketch of the Proof of Denseness of $\mathcal{P}_{0,M}$ 1$d$

- Given a draw $(X, Y, Z) \sim Q \in \mathcal{E}_{0,M}$, we construct a $(\tilde{X}, \tilde{Y}, \tilde{Z}) \sim P \in \mathcal{P}_{0,M}$ as follows.
- First construct $P_{\tilde{X},\tilde{Y},\tilde{Z}|X,Y,Z}$.
- Suppose $X \in A_i$, $Y \in B_j$ and $Z \in C_k$.
- Generate $(\tilde{X}, \tilde{Y}) \in A_i \times B_j$ and $\tilde{Z} \in C_{ijk}$ uniformly.
- Note that the density of $(\tilde{X}, \tilde{Y}, \tilde{Z})$ is proportional to

$$\sum_{i,j,k} \mathbb{1}\{\tilde{X} \in A_i, \tilde{Y} \in B_j, \tilde{Z} \in C_{ijk}\} P(X \in A_i, Y \in B_j, Z \in C_k).$$

- Job done. $\tilde{X} \perp\!\!\!\perp \tilde{Y} | \tilde{Z}$ and $W_p(P, Q)^P \leq \mathrm{EE}_{(\tilde{X},\tilde{Y},\tilde{Z})|(X,Y,Z)} ||(X, Y, Z) - (\tilde{X}, \tilde{Y}, \tilde{Z})||_2^p \leq (\sqrt{3}\frac{2M}{m})^p$

# Hardness of the Conditional Independence (CI) Test

## Take-Home Message

Due to this depressing theorem, we cannot expect to distinguish the full composite null and alternative hypotheses, and must restrict the model. This motivates us to use a model-based testing approach to limit the space of distribution we considered.

# Proposed Method and Simulation

# Review of Some Existing Methods

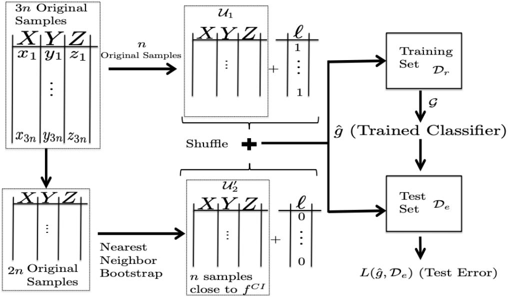
- We construct a simple stratified Chi-squared test as a benchmark. The test uses k-means algorithm to stratify $Z$ into $k$ bins and runs a Chi-squared goodness of fit test on each stratum. Effectively, it is testing if $X \perp\!\!\!\perp Y|Z \in bin\ i$. The test statistic is

$$\sum_{j\in\{bin\}} \sum_{i=1}^{2} \frac{(O_{ji} - E_{ji})^2}{E_{ji}}.$$

- Shah and Peters [3] proposed using the covariance of residuals from estimating $\mathrm{E}[X|Z]$ and $\mathrm{E}[Y|Z]$ as the test statistic. They developed the asymptotic distribution of the test statistic assuming the conditional means can be learnt well enough,

# Review of Some Existing Methods

Several other tests are based on the idea of the permutation test and use various methods to simulate the data under $H_0$.



- Sen et al. [4] suggest a 1-nearest-neighbor algorithm to permute one coordinate of $(X, Y)$ if two samples are close in terms of $Z$. The algorithm creates a sample which is almost CI. Then they suggests train a classifier to classify the simulated data from the original data.

# Review of Some Existing Methods

- Candès et.al [7] and Berrett et.al [6] both use the fact that if $H_0$ is true $X|Y, Z \overset{d}{=} X|Z$. They attempt to estimate the conditional distribution of $X|Z$ $(P(X|Z))$, then generate $(X^*, Y, Z)$ using $P^n(X|Z)$. For any test-statistic, one can use $(X^*, Y, Z)$ to compute p-values as in the usual permutation test.
- Our re-sampling strategy turns out to be similar to theirs to some extents.
- Disclosure: We didn't review these two papers until last week.

# Description of Our Method

- Let $(J_X(Z), J_Y(Z), J_{XY}(Z))$ be the 3-vector output of a NNet.
- We will model the joint distribution of $X, Y|Z$ with an ising model which has negative log-likelihood of

$$-\log P(X, Y|Z) = -J_X(Z) \cdot X - J_Y(Z) \cdot Y - J_{XY}(Z) \cdot XY + \psi(J(Z)).$$

- Express the hypotheses as $H_0 : J_{XY} = 0$, $H_1 : J_{XY} \neq 0$.
- Train the model on the training data and compute the test statistic

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{J}_{XY}(Z_i))^2$$

on the test set.

# Description of Our Method

## Motivation / Justification for Our Model

- $-\log f(X, Y)$ is a reasonable model to approximate the distribution of the marginal binary data $(X, Y)$ because higher-order terms are irrelevant.
- NNets can express a large class of functions. So, conditionally, the Ising model can represent a big class of distribution.
- Our method is very flexible with $Z$.
- We have a natural parameter and test statistic to capture the CI.
- We hope that the test statistic can have a high power against a huge class of distributions in $H_1$.

# Simulation

## Simulation Scheme

- We consider 4 sample sizes–$50, 100, 500, 1000$.
- For each sample size, generate 1000 trials of data under $H_0$ and 1000 trials under $H_1$.
- Apply testing methods on each trial.
- Plot the ROC curve.

## Simulation Scheme–Data Generating Distribution

First, generate $Z$ use a three-dimensional multivariate normal. Next, we consider two ways of generating $X$, $Y$ condition on Z

1. Ising model.
   - We use an a 1-layer neural network with 100 hidden nodes and $J_X, J_Y, J_{XY}$ as output to generate the parameter of the Ising model.
2. Mixture model.
   - Under $H_0$, if $||z_i||_2 > r$, $X_i \sim Bernoulli(0.5), Y = X$. Else, generate $(X_i, Y_i)$ uniformly on $\{-1, 1\} \times \{-1, 1\}$.
   - Under $H_1$, $X_i \sim Bernoulli(0.5)$. If $||z_i||_2 > r$, $Y_i = -X_i$. Else $Y_i = X_i$.
   - Note that if we ignore $Z$, $X$ and $Y$ are dependent under $H_0$. Under $H_1$, $X$ and $Y$ are independent.

# Simulation

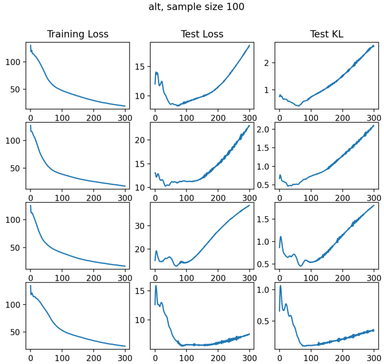
## Simulation Scheme–Test Methods

4 testing methods are implemented.

1. Chi-squared Goodness-of-Fit Test. $\sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$.
2. Stratified Chi-Sqarued Test.
3. CCIT.
   - We use the package default classifier–XGBoost.
4. Ising Model (our proposed method).
   - During each trial, we reserve 10% of data as test data; fit the Ising model on training data and compute $\hat{S}_n$ on the test set.
   - On Ising data, we fit a NNet with the exact architecture. On the mixture data, after extensive "grad student descent", we use a 1-layer NNet with 200 hidden nodes.

# Simulation

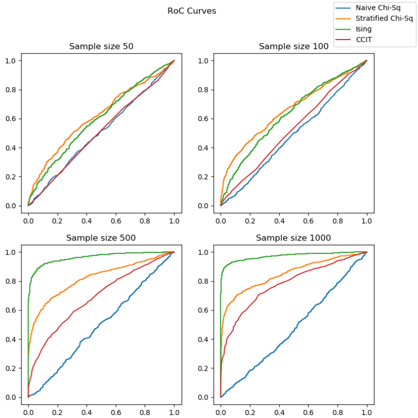
## NNet Parameter Tuning



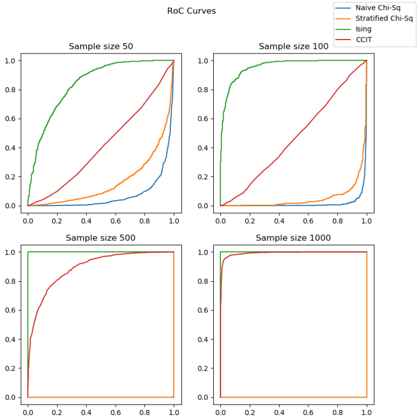
- We track the negative log-likelihood and KL-divergence on the test set.
- Unsurprisingly, likelihood tracks well with KL.
- We pick the network architecture and a fixed training epoch for each sample size based on the test KL.

# Simulation

## Results



ROC Curves under Ising Data.



ROC Curves under Mixture Data

# Myth: How to get the P-value?

## One Honorable Mention

- We have experimented with two approaches, results so far are mixed.
- Horel and Giesecke [5] showed under the regression setting when the squared loss is used, the one-layer neural network estimator converges to the maximum of a Gaussian process.
- We tried to adapt their methods to our setting and the p-value is always 0.

# Myth: How to get the P-value?

## Parametric Bootstrap P-value

Our proposal is using the Ising model to perform a parametric bootstrap. Denote

$$-\log P(X, Y|Z) = -J_X(Z) \cdot X - J_Y(Z) \cdot Y - J_{XY}(Z) \cdot XY + \psi(J(Z)).$$

as the full model and

$$-\log P_0(X, Y|Z) = -J_X(Z) \cdot X - J_Y(Z) \cdot Y + \psi(J(Z)).$$

as the reduced model.

### Parametric Bootstrap P-value Algorithm (PBPA)

1. Fit the full model and compute $\hat{S}_n$ as usual.
2. Fit the reduced model.
3. Use the fitted null model to generate $(X_i^B, Y_i^B)$ and compute $\hat{S}_n^B$ on the Bootstrap sample.
4. Repeat step 3 many times.
5. Retrun $p^B = P(\hat{S}_n^B > \hat{S}_n)$.

# Myth: How to get the P-value?

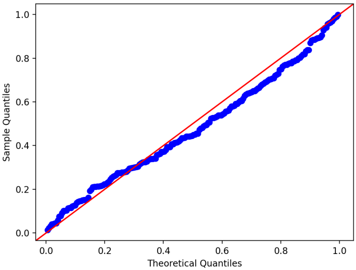
## Motivation of PBPA

- Need to resample under $H_0$.
- Already have a model to model data under $H_0$.
- Should have high power against $H_1$.

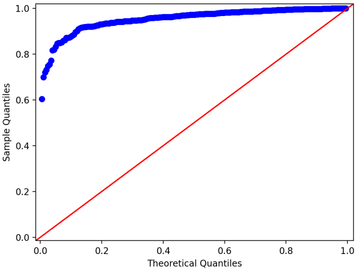
# Myth: How to get the P-value?

## Simulation Study of PBPA

On the null Ising data, the p-value is valid if the exact architecture is used; if the model is misspecified, the p-value is too conservative. On the mixture data, the p-value gets closer and closer to 0 as the sample size increases. We show the simulation results on 200 trials of null Ising data with sample size 100. The number of bootstrap samples is 1000.



Q–Q Plot of ”P-values using True Architecture”.



Q–Q Plot of ”P-values using Misspecified Architecture

# Myth: How to get the P-value?

## Connection to Previous Works

- As in the Candès and Berrett's methods, PBPA estimate the conditional distribution which is used to resample data.
- Their methods works with $P(X|Z)$, whereas PBPA play with $P(X, Y|Z)$.
- Berrett et.al provides an upper bound for the p-value if the estimation of $P(X|Z)$ is sufficiently well.
- A simple modification will make the proof work for our bootstrap p-value ($p^B$).
- This suggests that miscalibration of p-value may stem from the poor approximation of $P(X, Y|Z)$.

# Myth: How to get the P-value

## Upper Bound for Bootstrap P-value

### Definition: Total Variation Distance.

For any two distribution $Q_1$ and $Q_2$ defined on the same probability space. The total variation distance $d_{TV}(Q_1, Q_2) = sup_A|Q_1(A) - Q_2(A)|$, where $A$ is any measurable set in the $\sigma$-algebra.

### Theorem: Upper Bound for $p^B$.

Assume $H_0$ is true, and the conditional distribution of $X, Y|Z$ is given by $P(X, Y|Z)$. For a fixed integer $M \geq 1$, let $(X^{(1)}, Y^{(1)}), \ldots, (X^{(M)}, Y^{(M)})$ be copies of $(X, Y)$ generated by the reduced model $P_0(X, Y|Z)$ in the PBPA. Then $\forall \alpha \in [0, 1]$,

$$P(p^B \leq \alpha|Z) \leq \alpha + d_{TV}(P_0(X, Y|Z), P(X, Y|Z)).$$

# Future Work

There are two possible paths ahead of us.

1. Obtaining the asymptotic distribution of the test statistic.
    - Pro: Computationally easy.
    - Con: Theories of NNets are hard.
2. Continue on the path of PBPA.

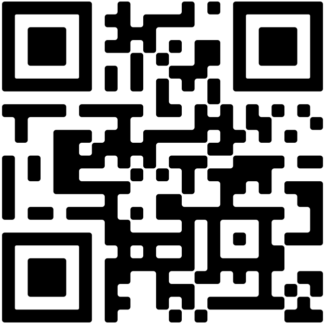
If we choose path 2, ideally, we want to prove the following:

- The p-value is conservative with high probability under the mis-specified model.
- The test has a high power.

Finally, if possible, we want to extend the method to multivariate $X, Y$ and pursue applications to electronic health records.

# Code and Data

The code and data used for the project is publicly available. Scan the QR code to access the repository.



# Bibliography

Neykov, Matey, et al. “Minimax Optimal Conditional Independence Testing.” ArXiv:2001.03039 [Math, Stat], Jan. 2020. arXiv.org, http://arxiv.org/abs/2001.03039.

Wang, Yuhao, et al. “Direct Estimation of Differences in Causal Graphs.” ArXiv:1802.05631 [Stat], Nov. 2018. arXiv.org, http://arxiv.org/abs/1802.05631.

Shah, Rajen D., and Jonas Peters. “The Hardness of Conditional Independence Testing and the Generalised Covariance Measure.” ArXiv:1804.07203 [Math, Stat], July 2020. arXiv.org, http://arxiv.org/abs/1804.07203.

Sen, Rajat, et al. “Model-Powered Conditional Independence Test.” ArXiv:1709.06138 [Cs, Math, Stat], Sept. 2017. arXiv.org, http://arxiv.org/abs/1709.06138.

# Bibliography

Horel, Enguerrand, and Kay Giesecke. “Significance Tests for Neural Networks.” ArXiv:1902.06021 [Cs, Math, Stat], Mar. 2020. arXiv.org, http://arxiv.org/abs/1902.06021.

Berrett, Thomas B., et al. The Conditional Permutation Test for Independence While Controlling for Confounders. July 2018. arxiv.org, https://arxiv.org/abs/1807.05405v2.

Candès, Emmanuel, et al. Panning for Gold: Model-X Knockoffs for High-Dimensional Controlled Variable Selection. Oct. 2016. arxiv.org, https://arxiv.org/abs/1610.02351v4.