# Conditional Randomization Test

## For Conditional Independence Testing

Abhinav Chakraborty

The Wharton School, University of Pennsylvania

# Table of contents

# Intro

# Problem Statement

We have

- predictor $X \in \mathbb{R}^{d_X}$ (geneotype of at 10 adjacent polymorphic sites)
- response $Y \in \mathbb{R}$ (Cholesterol Level)
- covariate $Z \in \mathbb{R}^{d_Z}$ (genotype at other sites)

# Problem Statement

We have

- predictor $\mathsf{X} \in \mathbb{R}^{d_X}$ (geneotype of at 10 adjacent polymorphic sites)
- response $\mathsf{Y} \in \mathbb{R}$ (Cholesterol Level)
- covariate $\mathsf{Z} \in \mathbb{R}^{d_Z}$ (genotype at other sites)

Want to test

$$H_0 : \mathsf{Y} \perp\!\!\!\perp \mathsf{X} \mid \mathsf{Z} \qquad \text{v/s} \qquad H_1 : \mathsf{Y} \not\!\perp\!\!\!\perp \mathsf{X} \mid \mathsf{Z}$$

using $n$ data points

$$(X, Y, Z) = \{(X_i, Y_i, Z_i\}_{i=1,\ldots,n} \overset{i.i.d}{\sim} \mathcal{L}$$

# Application: Variable Selection

- Given p covariates $X_1, \ldots, X_p$ and a response $Y$
- Determine $X_j$ such that $Y \not\perp\!\!\!\perp X_j \mid \mathbf{X}_{-j}$ by testing

$$H_{0j} : Y \perp\!\!\!\perp X_j \mid \mathbf{X}_{-j} \quad \text{v/s} \quad H_{1j} : Y \not\perp\!\!\!\perp X_j \mid \mathbf{X}_{-j}$$

- Formulate a multiple testing problem on the resulting p-values.

# Model-X framework

# Model-X framework

(MX) assumption

$$\mathcal{L}(\mathsf{X} \mid \mathsf{Z}) = f^*_{\mathsf{X}|\mathsf{Z}} \text{ for some known } f^*_{\mathsf{X}|\mathsf{Z}}$$

Q: Why do we need such an assumption?
A: Shah and Peters (2018) show that without any assumption on $\mathcal{L}$ no asymptotically uniform valid test can have non-trivial power against *any* alternative.

# Model-X framework

(MX) assumption

$$\mathcal{L}(\mathsf{X} \mid \mathsf{Z}) = f^*_{\mathsf{X}|\mathsf{Z}} \text{ for some known } f^*_{\mathsf{X}|\mathsf{Z}}$$

Q: Can we really make such a strong assumption?

- In GWAS example earlier $\mathsf{X} \mid \mathsf{Z}$ reflects the joint distribution of genotype across the genome, which is well described by hidden markov model from population genetics.

# Model-X framework

(MX) assumption

$$\mathcal{L}(X \mid Z) = f^*_{X|Z} \text{ for some known } f^*_{X|Z}$$

Q: Can we really make such a strong assumption?

- In general causal inference settings, the MX assumptions are equivalent to assuming the propensity score to be known. e.g. X-treatment, Y-outcome, Z-covariate (age).

# CRT: Algorithm

1. Input: The distribution of $X \mid Z$, data $(X, Y, Z)$, test statistic function T, number of randomizations $M$.
2. For $m = 1, 2, \ldots, M$ : Sample $X^{(m)}$ from the distribution $\mathcal{L}(\mathrm{X} \mid \mathrm{Z})$, conditionally independently of $X$ and $Y$.
3. Output: CRT p-value

$$\frac{1}{M+1}\left[1 + \sum_{m=1}^{M} \mathbb{1}\{T(X^{(m)}, Y, Z) \geq T(X, Y, Z)\}\right]$$

# Validity of CRT p-values

**Theorem (Candes et al. (2018)[1])**

*The CRT p-value $p(X, Y, Z)$ satisfies*

$$\mathbb{P}_{H_0}(p(X, Y, Z) \leq \alpha) \leq \alpha$$

*for all $\alpha \in [0, 1]$.*

# Model-X(2) Framework

## MX(2) Framework

Q: Can the Model-X assumption be relaxed?

A: Yes, but for a particular choice of test statistic $T$ and that too with asymptotic type-I error guarantees.

# MX(2) Framework

Q: Can the Model-X assumption be relaxed?
A: Yes, but for a particular choice of test statistic $T$ and that too with asymptotic type-I error guarantees.

**MX(2) assumption**

the conditional mean $\mathbb{E}(X \mid Z)$ and variance $\mathrm{Var}(X \mid Z)$ is known.

**Setting 1 (Arbitrary Dimension Asymptotics)**

Data: $(\mathsf{X}, \mathsf{Y}, \mathsf{Z}) \in \mathbb{R}^{d_X+1+d_Z}$, $d_X$ is fixed but $d_Z$ can vary arbitrarily with $n$.

$$\mu_n(Z) = \mathbb{E}(\mathsf{X} \mid \mathsf{Z}) \quad \Sigma_n(Z) = \mathrm{Var}(\mathsf{X} \mid \mathsf{Z}) \qquad \text{known}$$

# MX(2) $F$ test

1. We train an estimate $\hat{g}_n$ of $E[Y|Z]$ on an independent dataset.
2. Set $\widehat{S}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{g}_n(Z_i)\right)^2 \Sigma_n(Z_i)$
3. Set $U_n \equiv \frac{\widehat{S}_n^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_i - \widehat{g}_n(Z_i)\right)\left(X_i - \mu_n(Z_i)\right)$ and $T_n = \|U_n\|^2$ (normalised product of residual statistics)
4. Result: $\mathrm{MX}(2)F$-test asymptotic $p$-value $\widehat{p} \equiv \mathbb{P}\left[\chi^2_{d_X} > T_n\right]$.

We will use this same statistic ($T_n$) in *CRT* for testing independence.

# Asymptotic validity of MX(2) F test

**Theorem (Katsevich et al. (2020) [2])**
*If $\mathcal{L}_n$ and $\hat{g}_n$ satisfy certain moment conditions, then the $U_n(X, Y, Z)$ converges to the standard normal:*

$$U_n(X, Y, Z) \overset{\mathcal{L}_n}{\rightarrow}_d N\left(0, I_{d_X}\right)$$

*Therefore, the $MX(2)F$-test uniformly controls Type-I error asympotically.*

$$\limsup_{n \to \infty} \sup_{\mathcal{L}_n} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathrm{MX}(2)}(X, Y, Z)\right] \leq \alpha$$

# CRT v/s MX(2) F test

# Equivalence of CRT and MX(2) F test

**Theorem (Katsevich et al. (2020) [2])**
*Under the same conditions as the previous theorem. Let* $\phi_n^{\mathrm{CRT}}$ *denote the CRT based on* $T_n$*, with threshold* $C_n(Y, Z)$ *denoting the* $(1-\alpha)$*-th conditional quantile of* $T_n \mid Y, Z$*. The CRT threshold converges in probability to the MX (2) threshold:*

$$C_n(Y, Z) \overset{\mathcal{L}_n}{\rightarrow}_p c_{d_X, 1-\alpha}.$$

*Furthermore, if* $T_n(X, Y, Z)$ *does not accumulate near* $c_{d_X, 1-\alpha}$*, i.e.*

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}_{\mathcal{L}_n}\left[\left|T_n(X, Y, Z) - c_{d_X, 1-\alpha}\right| \leq \delta\right] = 0,$$

*then the CRT is asymptotically equivalent to the MX*$(2)$*F-test:*

$$\lim_{n \to \infty} \mathbb{P}_{\mathcal{L}_n}\left[\phi_n^{\mathrm{MX}(2)}(X, Y, Z) \neq \phi_n^{\mathrm{CRT}}(X, Y, Z)\right] = 0.$$

# Power Analysis of CRT

**Setting 2 (Semiparametric alternatives)**

Under Setting 1, assume $\mathcal{L}_n(Y \mid X, Z)$ is such that

$$Y = (X - \mu_n(Z))^T \beta_n + g_n(Z) + \epsilon; \quad \epsilon \sim N\left(0, \sigma^2\right), \sigma^2 > 0$$

for $\epsilon(X, Y, Z)$. Here, $\beta_n \in \mathbb{R}^{d_X}$ is a coefficient vector, $g_n : \mathbb{R}^{D_Z} \to \mathbb{R}$ a general function, and $\sigma^2 > 0$ the residual variance.

# Asymptotic Power against local alternatives

To test

$$H_0 \ : \ \beta = 0 \quad v/s \quad H_{1n} : \beta_n = \frac{h_n}{\sqrt{n}}$$

where $h_n \to h$. We evaluate

$$\lim_{n\to\infty} \mathbb{E}_{H_{1n}}(\phi_n(X, Y, Z))$$

for a level $\alpha$ test $\phi_n$.

# Asymptotic Power of MX(2) and CRT

In the next theorem we express the asymptotic power of the $\mathrm{MX}(2)F$-test against semiparametric alternatives in terms of the variance-weighted mean square error of $\widehat{g}_n$ :

$$\mathcal{E}_n^2 \equiv \mathbb{E}_{\mathcal{L}_n}\left[\left(\widehat{g}_n(Z) - g_n(Z)\right)^2 \cdot \bar{\Sigma}_n^{-1/2}\Sigma_n(Z)\bar{\Sigma}_n^{-1/2}\right],\ \text{where } \bar{\Sigma}_n \equiv \mathbb{E}_{\mathcal{L}_n}\left[\Sigma_n(Z)\right]$$

**Theorem ([2])**
*Consider the semiparametric alternative in Setting 2. Suppose $\mathcal{L}_n$ satisfies some moment conditions, and that the conditional variance and variance-weighted mean squared error converge:*

$$\bar{\Sigma}_n \to \bar{\Sigma} \quad \text{and} \quad \mathcal{E}_n^2 \to \mathcal{E}^2 \quad \text{as } n \to \infty$$

*Then, we have the following two statements:*

(a) (Consistency) If $\beta_n = \beta \neq 0$ for each $n$, then the *MX*(2)*F*-test and the *CRT* based on the same statistic are consistent:

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathrm{MX(2)}}(X,Y,Z)\right] = \lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathrm{CRT}}(X,Y,Z)\right] = 1$$

(a) (Consistency) If $\beta_n = \beta \neq 0$ for each $n$, then the *MX*(2)*F*-test and the *CRT* based on the same statistic are consistent:

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathrm{MX(2)}}(X,Y,Z)\right] = \lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathrm{CRT}}(X,Y,Z)\right] = 1$$

(b) (Power against local alternatives) If $\beta_n = h_n/\sqrt{n}$ for a convergent sequence $h_n \to h \in \mathbb{R}^d$, then

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathrm{MX(2)}}(X,Y,Z)\right] &= \lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathrm{CRT}}(X,Y,Z)\right] \\ &= \mathbb{P}\left[\chi_d^2\left(\left\|\left(\sigma^2 I_{d_X} + \mathcal{E}^2\right)^{-1/2} \bar{\Sigma}^{1/2} h\right\|^2\right) > c_{d_X,1-\alpha}\right]\end{aligned}$$

# Asymptotic power and Estimation error

- The theorem establishes a direct link between the estimation error in $\hat{g}_n$ and the power of the CRT against local alternatives.
- In particular, the mean-squared error term $\mathcal{E}^2$ contributes additively to the irreducible error term $\sigma^2 I_d$.

# Conclusion

# Summary

- Conditional Independence testing: its hardness!
- Testing under Model X framework
- Relaxing Model X assumption - MX(2); paying a price (no finite sample guarantee)
- Asymptotic equivalence of two tests.
- Asymptotic power of CRT.

# Future Work

# Future Work

- Power analysis of more general models in MX/MX(2) framework!
- Fully letting go of the MX(2) assumption learning about the distribution from in-sample and/or test data.

# My current research

We look at the statistic

$$T(X, Y, Z) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \hat{h}_n(Z_i))(Y_i - \hat{g}_n(Z_i))$$

where $\hat{h}_n(.)$ and $\hat{g}_n(.)$ are learned in-sample.

# My current research

- We learn a model of $\mathcal{L}_n(X|Z) \equiv \hat{\mathcal{L}}_n(X|Z)$.
- Generate samples $\hat{X}^{(1)}, \hat{X}^{(2)}, \ldots, \hat{X}^{(M)} \sim \hat{\mathcal{L}}_n(X|Z)$.
- Return the p-value

$$p_n(X, Y, Z) := \frac{1}{M+1}\left[1 + \sum_{m=1}^{M} \mathbb{1}\{T(\hat{X}^{(m)}, Y, Z) \geq T(X, Y, Z)\}\right]$$

# My current research

**Theorem**

*Informally: under some condition on* $\mathcal{L}_n, \hat{\mathcal{L}}_n, \hat{h}_n$ *and* $\hat{g}_n$.

$$p_n(X, Y, Z) \xrightarrow{d} U[0, 1] \quad \textit{under } H_0$$

**Novelty**

Free of MX/MX(2) assumptions and we use the entire data set for testing!

Questions?

# References i


E. Candes, Y. Fan, L. Janson, and J. Lv.
**Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection.**
*Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

E. Katsevich and A. Ramdas.
**On the power of conditional independence testing under model-x.**
*arXiv preprint arXiv:2005.05506*, 2020.