A theoretical treatment of conditional independence testing under Model-X

Eugene Katsevich

August 3, 2020

Joint work with Aaditya Ramdas (CMU)

Preprint available at arxiv.org/abs/2005.05506

Conditional independence testing under Model-X

For random variables $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \in \mathbb{R}^{1+1+p}$, would like to test

$$H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$$

based on a sample $(X, Y, Z) = \{(X_i, Y_i, Z_i) : i = 1, \ldots, n\}$.

Conditional independence testing under Model-X

For random variables $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \in \mathbb{R}^{1+1+p}$, would like to test

$$H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$$

based on a sample $(X, Y, Z) = \{(X_i, Y_i, Z_i) : i = 1, \ldots, n\}$.

Model-X assumption (Candès et al., 2018)

$f_{\boldsymbol{X}|\boldsymbol{Z}} = f^*_{\boldsymbol{X}|\boldsymbol{Z}}$ for known f^*

Model-X methodologies

MX knockoffs and the conditional randomization test (CRT) proposed by Candès et al., 2018.

Algorithm: Conditional Randomization Test

Data: Samples (X_i, Y_i, Z_i), test statistic $T(X, Y, Z)$, MX $f^*_{\boldsymbol{X}|\boldsymbol{Z}}$

Compute $T(X, Y, Z)$;

for $b = 1, \ldots, B$ **do**

  Resample $\widetilde{X}_i^b$ from $f^*_{\boldsymbol{X}|\boldsymbol{Z}=Z_i}$ and recompute $T(\widetilde{X}^b, Y, Z)$;

end

Return: $p_{\text{CRT}} = \frac{1}{1+B}\left(1 + \sum_{b=1}^{B} \mathbb{1}(T(\widetilde{X}^b, Y, Z) \geq T(X, Y, Z))\right)$

A variety of knockoffs and CRT extensions are now available.

Common themes in MX methodology

- T usually based on a statistical machine learning method
- Performance of the ML method impacts the power of the test
- Randomness in X used for inference, conditioning on Y and Z

Common themes in MX methodology

- T usually based on a statistical machine learning method
- Performance of the ML method impacts the power of the test
- Randomness in X used for inference, conditioning on Y and Z

Goal of this talk:

Develop a quantitative understanding of these themes.

Most powerful CRT against a point alternative

Given alternative distributions $\bar{f}_{\boldsymbol{Z}}$ and $\bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$, consider testing

$$
\begin{aligned}
H_0 &: (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim f_{\boldsymbol{Z}} f^*_{\boldsymbol{X}|\boldsymbol{Z}} f_{\boldsymbol{Y}|\boldsymbol{Z}} \quad \text{for some } f_{\boldsymbol{Z}}, f_{\boldsymbol{Y}|\boldsymbol{Z}} \\
H_1 &: (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim \bar{f}_{\boldsymbol{Z}} f^*_{\boldsymbol{X}|\boldsymbol{Z}} \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}.
\end{aligned}
$$

Composite null prevents directly deducing the most powerful test.

Most powerful CRT against a point alternative

Given alternative distributions $\bar{f}_{\boldsymbol{Z}}$ and $\bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$, consider testing

$$H_0 : (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim f_{\boldsymbol{Z}} f^*_{\boldsymbol{X}|\boldsymbol{Z}} f_{\boldsymbol{Y}|\boldsymbol{Z}} \quad \text{for some } f_{\boldsymbol{Z}}, f_{\boldsymbol{Y}|\boldsymbol{Z}}$$
$$H_1 : (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim \bar{f}_{\boldsymbol{Z}} f^*_{\boldsymbol{X}|\boldsymbol{Z}} \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}.$$

Composite null prevents directly deducing the most powerful test.

But the CRT ϕ_T is also a *conditionally valid* test:

$$\sup_{H_0} \mathbb{E}[\phi_T(X, Y, Z) | Y = y, Z = z] \leq \alpha \quad \text{for all } y, z.$$

What is the most powerful *conditionally valid* test?

Conditioning reduces a composite null to a point null

Fix realizations Y_i and Z_i for each i. Then,

$$\begin{aligned}
&\text{Under } H_0, \quad X_i | Y_i, Z_i \overset{\text{ind}}{\sim} f^*_{X_i|Z_i}; \\
&\text{Under } H_1, \quad X_i | Y_i, Z_i \overset{\text{ind}}{\sim} f^*_{X_i|Z_i} \frac{\bar{f}_{Y_i|X_i,Z_i}}{\bar{f}_{Y_i|Z_i}}.
\end{aligned}$$

So, conditioning on Y, Z gives a simple hypothesis testing problem.

Neyman-Pearson gives the most powerful CRT

By NP, most powerful conditionally valid test rejects for large

$$T^{\text{opt}}(X, Y, Z) = \prod_{i=1}^{n} \frac{\bar{f}_{Y_i|X_i,Z_i}}{\bar{f}_{Y_i|Z_i}} \propto \prod_{i=1}^{n} \bar{f}_{Y_i|X_i,Z_i}.$$

Neyman-Pearson gives the most powerful CRT

By NP, most powerful conditionally valid test rejects for large

$$T^{\text{opt}}(X, Y, Z) = \prod_{i=1}^{n} \frac{\bar{f}_{Y_i|X_i,Z_i}}{\bar{f}_{Y_i|Z_i}} \propto \prod_{i=1}^{n} \bar{f}_{Y_i|X_i,Z_i}.$$

Theorem[1]

CRT based on T^{opt} is the most powerful conditionally valid test against $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim \bar{f}_{\boldsymbol{Z}} f^{*}_{\boldsymbol{X}|\boldsymbol{Z}} \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$.

[1] All results stated informally

Neyman-Pearson gives the most powerful CRT

By NP, most powerful conditionally valid test rejects for large

$$T^{\text{opt}}(X, Y, Z) = \prod_{i=1}^{n} \frac{\bar{f}_{Y_i|X_i,Z_i}}{\bar{f}_{Y_i|Z_i}} \propto \prod_{i=1}^{n} \bar{f}_{Y_i|X_i,Z_i}.$$

Theorem[1]

CRT based on T^{opt} is the most powerful conditionally valid test against $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim \bar{f}_{\boldsymbol{Z}} f^*_{\boldsymbol{X}|\boldsymbol{Z}} \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$.

The knockoff statistic $T^{\text{opt}}([X, \widetilde{X}], Y) = \prod_{i=1}^{n} \bar{f}_{Y_i|X_i}$ maximizes the probability $\mathbb{P}[T([X, \widetilde{X}], Y) > T([X, \widetilde{X}]_{\text{swap}(j)}, Y)]$.

[1]All results stated informally

Neyman-Pearson gives the most powerful CRT

By NP, most powerful conditionally valid test rejects for large

$$T^{\text{opt}}(X, Y, Z) = \prod_{i=1}^{n} \frac{\bar{f}_{Y_i|X_i,Z_i}}{\bar{f}_{Y_i|Z_i}} \propto \prod_{i=1}^{n} \bar{f}_{Y_i|X_i,Z_i}.$$

Theorem[1]

CRT based on T^{opt} is the most powerful conditionally valid test against $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim \bar{f}_{\boldsymbol{Z}} f^*_{\boldsymbol{X}|\boldsymbol{Z}} \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$.

The knockoff statistic $T^{\text{opt}}([X, \widetilde{X}], Y) = \prod_{i=1}^{n} \bar{f}_{Y_i|X_i}$ maximizes the probability $\mathbb{P}[T([X, \widetilde{X}], Y) > T([X, \widetilde{X}]_{\text{swap}(j)}, Y)]$.

ML methods T used in practice learn approximations to $\bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$.

[1] All results stated informally

Connections

- *Least squares:* If $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma + \epsilon$, the optimal CRT statistic is $\|Y - X\beta - Z\gamma\|^2 - \|Y - Z\gamma\|^2$, akin to the OLS F-statistic.

Connections

- *Least squares:* If $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma + \boldsymbol{\epsilon}$, the optimal CRT statistic is $\|Y - X\beta - Z\gamma\|^2 - \|Y - Z\gamma\|^2$, akin to the OLS F-statistic.
- *Unbiased testing:* In parametric families with nuisance params, MP unbiased test is MP test conditional on nuisance sufficient statistic.

Connections

- *Least squares:* If $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma + \boldsymbol{\epsilon}$, the optimal CRT statistic is $\|Y - X\beta - Z\gamma\|^2 - \|Y - Z\gamma\|^2$, akin to the OLS F-statistic.
- *Unbiased testing:* In parametric families with nuisance params, MP unbiased test is MP test conditional on nuisance sufficient statistic.
- *Holdout randomization test*[2]*:* $\widehat{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$ learned on a training set and CRT based on loss $\sum_i \log \widehat{f}_{Y_i|X_i,Z_i}$ run on a test set.

[2] Tansey et al., 2018, Bates et al., 2020

Impact of ML prediction error on CRT power

Impact of ML prediction error on CRT power

Suppose that

$$\boldsymbol{Y} = \boldsymbol{X}\beta + g(\boldsymbol{Z}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2),$$

and we estimate $\widehat{g}$ based on a separate training set (like HRT).

How does test error in $\widehat{g}$ impact the power of the CRT based on $\widehat{g}$?

Impact of ML prediction error on CRT power

Suppose that

$$\boldsymbol{Y} = \boldsymbol{X}\beta + g(\boldsymbol{Z}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2),$$

and we estimate $\widehat{g}$ based on a separate training set (like HRT).

How does test error in $\widehat{g}$ impact the power of the CRT based on $\widehat{g}$?

In particular, consider the CRT based on[3]

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

where $\mu_i = \mathbb{E}[X_i | Z_i]$.

[3] Related to the *generalized covariance measure* of Shah and Peters (2020); studied in the double robustness literature, e.g. Chernozhukov et al. (2018).

Prediction error impacts asymptotic efficiency of the CRT

For simplicity, suppose $\text{Var}[\boldsymbol{X}|\boldsymbol{Z}] = s^2$ a.s. for some $s^2 > 0$.

Define the test error $\mathcal{E} = \mathbb{E}[(\widehat{g}(\boldsymbol{Z}) - g(\boldsymbol{Z}))^2]$.

Fixing dimension and training set, let $n \to \infty$ and $\beta_n = \frac{h}{\sqrt{n}}$.

Prediction error impacts asymptotic efficiency of the CRT

For simplicity, suppose $\text{Var}[\boldsymbol{X}|\boldsymbol{Z}] = s^2$ a.s. for some $s^2 > 0$.

Define the test error $\mathcal{E} = \mathbb{E}[(\widehat{g}(\boldsymbol{Z}) - g(\boldsymbol{Z}))^2]$.

Fixing dimension and training set, let $n \to \infty$ and $\beta_n = \frac{h}{\sqrt{n}}$.

Theorem

Under the MX assumption and bounded fourth moments,

$$\mathbb{E}\left[\phi_T(X, Y, Z)|Y, Z\right] \to \Phi\left(z_\alpha + \frac{hs}{\sqrt{\sigma^2 + \mathcal{E}}}\right),$$

almost surely in Y, Z.

More connections to linear regression

Note that

$$\boldsymbol{Y} - \widehat{g}(\boldsymbol{Z}) = \boldsymbol{X}\beta + (g(\boldsymbol{Z}) - \widehat{g}(\boldsymbol{Z}) + \boldsymbol{\epsilon}) = \boldsymbol{X}\beta + \boldsymbol{\epsilon}'; \quad \boldsymbol{\epsilon}' \sim (0, \sigma^2 + \mathcal{E}).$$

Estimation error in $\widehat{g}$ inflates the noise level by $\mathcal{E}$.

More connections to linear regression

Note that

$$\boldsymbol{Y} - \widehat{g}(\boldsymbol{Z}) = \boldsymbol{X}\beta + (g(\boldsymbol{Z}) - \widehat{g}(\boldsymbol{Z}) + \boldsymbol{\epsilon}) = \boldsymbol{X}\beta + \boldsymbol{\epsilon}'; \quad \boldsymbol{\epsilon}' \sim (0, \sigma^2 + \mathcal{E}).$$

Estimation error in $\widehat{g}$ inflates the noise level by $\mathcal{E}$.

$T(X, Y, Z) = \frac{1}{\sqrt{n}}(X - \mu)^T(Y - \widehat{g}(Z))$ is unnormalized OLS t-stat.

More connections to linear regression

Note that

$$\boldsymbol{Y} - \widehat{g}(\boldsymbol{Z}) = \boldsymbol{X}\beta + (g(\boldsymbol{Z}) - \widehat{g}(\boldsymbol{Z}) + \boldsymbol{\epsilon}) = \boldsymbol{X}\beta + \boldsymbol{\epsilon}'; \quad \boldsymbol{\epsilon}' \sim (0, \sigma^2 + \mathcal{E}).$$

Estimation error in $\widehat{g}$ inflates the noise level by $\mathcal{E}$.

$T(X, Y, Z) = \frac{1}{\sqrt{n}}(X - \mu)^T(Y - \widehat{g}(Z))$ is unnormalized OLS t-stat.

Under the null ($\beta = 0$),

$$T(\widetilde{X}, Y, Z) | Y, Z \to N(0, s^2(\sigma^2 + \mathcal{E}))$$

More connections to linear regression

Note that

$$\boldsymbol{Y} - \widehat{g}(\boldsymbol{Z}) = \boldsymbol{X}\beta + (g(\boldsymbol{Z}) - \widehat{g}(\boldsymbol{Z}) + \boldsymbol{\epsilon}) = \boldsymbol{X}\beta + \boldsymbol{\epsilon}'; \quad \boldsymbol{\epsilon}' \sim (0, \sigma^2 + \mathcal{E}).$$

Estimation error in $\widehat{g}$ inflates the noise level by $\mathcal{E}$.

$T(X, Y, Z) = \frac{1}{\sqrt{n}}(X - \mu)^T(Y - \widehat{g}(Z))$ is unnormalized OLS t-stat.

Under the null ($\beta = 0$),

$$T(\widetilde{X}, Y, Z) | Y, Z \to N(0, s^2(\sigma^2 + \mathcal{E}))$$
$$T(X, Y, Z) | X, Z \to N(0, s^2(\sigma^2 + \mathcal{E})).$$

More connections to linear regression

Note that

$$\boldsymbol{Y} - \widehat{g}(\boldsymbol{Z}) = \boldsymbol{X}\beta + (g(\boldsymbol{Z}) - \widehat{g}(\boldsymbol{Z}) + \boldsymbol{\epsilon}) = \boldsymbol{X}\beta + \boldsymbol{\epsilon}'; \quad \boldsymbol{\epsilon}' \sim (0, \sigma^2 + \mathcal{E}).$$

Estimation error in $\widehat{g}$ inflates the noise level by $\mathcal{E}$.

$T(X, Y, Z) = \frac{1}{\sqrt{n}}(X - \mu)^T(Y - \widehat{g}(Z))$ is unnormalized OLS t-stat.

Under the null ($\beta = 0$),

$$\begin{aligned} T(\widetilde{X}, Y, Z) | Y, Z &\to N(0, s^2(\sigma^2 + \mathcal{E})) \\ T(X, Y, Z) | X, Z &\to N(0, s^2(\sigma^2 + \mathcal{E})). \end{aligned}$$

If the semiparametric model is true, CRT resampling distribution asymptotically equivalent to OLS null distribution.

Asymptotic validity under a weaker assumption than MX

Asymptotic validity under a weaker assumption than MX

Under the null,

$$\mathrm{Var}\left[\,T_n \middle|\, Y, Z\right] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}[X_i|Z_i](Y_i - \widehat{g}(Z_i))^2 = S_n^2,$$

with S_n^2 known. We can show that, almost surely in (Y, Z),

$$\mathcal{L}\left(S_n^{-1} T_n | Y, Z\right) \to N(0, 1).$$

(S_n, T_n) only involves first and second moments $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}], \mathrm{Var}[\boldsymbol{X}|\boldsymbol{Z}]$.

Asymptotic validity under a weaker assumption than MX

This observation motivates the following:

Definition (MX(2) assumption)

$(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is such that $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}] = \mu(\boldsymbol{Z})$ and $\text{Var}[\boldsymbol{X}|\boldsymbol{Z}] = s^2(\boldsymbol{Z})$, for known functions $\mu(\cdot)$ and $s^2(\cdot)$.

Asymptotic validity under a weaker assumption than MX

This observation motivates the following:

Definition (MX(2) assumption)

$(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is such that $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}] = \mu(\boldsymbol{Z})$ and $\text{Var}[\boldsymbol{X}|\boldsymbol{Z}] = s^2(\boldsymbol{Z})$, for known functions $\mu(\cdot)$ and $s^2(\cdot)$.

There is an asymptotically valid conditional independence test that does not require the MX assumption or resampling:

Theorem

Under MX(2), $\phi = \mathbb{1}(S_n^{-1} T_n > z_{1-\alpha})$ *has uniform asympt. level* α.

Asymptotic validity under a weaker assumption than MX

This observation motivates the following:

Definition (MX(2) assumption)

$(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is such that $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}] = \mu(\boldsymbol{Z})$ and $\text{Var}[\boldsymbol{X}|\boldsymbol{Z}] = s^2(\boldsymbol{Z})$, for known functions $\mu(\cdot)$ and $s^2(\cdot)$.

There is an asymptotically valid conditional independence test that does not require the MX assumption or resampling:

Theorem

Under MX(2), $\phi = \mathbb{1}(S_n^{-1} T_n > z_{1-\alpha})$ *has uniform asympt. level* α.

Related to the double robustness literature, but the latter focuses on approximating the first moments $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$.

Connections to causal inference

The MX setting is like that of a randomized experiment:
$\boldsymbol{X}$ is the treatment; $\boldsymbol{Y}$ is the response; $\boldsymbol{Z}$ are the covariates.

Instead of complete randomization, the treatment $\boldsymbol{X}$ is assigned to units based on the covariates $\boldsymbol{Z}$ using a known mechanism $f^*_{\boldsymbol{X}|\boldsymbol{Z}}$.

Even in the absence of confounding, adjusting for covariates known to reduce variance in estimates of causal effect.

Connections to causal inference

Non-asymptotic tests based on resampling $\boldsymbol{X}$ go back to Fisher (1935) and Rosenbaum (1984). Both treat $\boldsymbol{Y}, \boldsymbol{Z}$ as fixed.

Asymptotic "superpopulation" approach (e.g. Robins et al., 1992) treats $\boldsymbol{Y}$ as random, focused on semiparametric models.

Current work reinforces close links between the two approaches; see also discussion in Rosenbaum (2002).

Summary

In this talk, we

- Identified the CRT most powerful against point alternatives;

Summary

In this talk, we

- Identified the CRT most powerful against point alternatives;
- Expressed CRT's asymptotic power in terms of ML test error;

Summary

In this talk, we

- Identified the CRT most powerful against point alternatives;
- Expressed CRT's asymptotic power in terms of ML test error;
- Weakened the MX assumption, retaining asymptotic validity;

Summary

In this talk, we

- Identified the CRT most powerful against point alternatives;
- Expressed CRT's asymptotic power in terms of ML test error;
- Weakened the MX assumption, retaining asymptotic validity;
- Drew some connections between MX and causal inference.

Future work

Many questions still remain open:

- Are all valid tests under MX also conditionally valid? If not, are all optimal tests conditionally valid?

Future work

Many questions still remain open:

- Are all valid tests under MX also conditionally valid? If not, are all optimal tests conditionally valid?
- Optimality statements against composite alternatives?

Future work

Many questions still remain open:

- Are all valid tests under MX also conditionally valid? If not, are all optimal tests conditionally valid?
- Optimality statements against composite alternatives?
- Extensions of power results to high dimensions? Some results for lasso test statistics available for knockoffs[4] and for CRT[5].

[4] Weinstein et al. (2017, 2020), Fan et al. (2019), Liu and Rigollet (2019)

[5] Celentano et al. (2020)

Future work

Many questions still remain open:

- Are all valid tests under MX also conditionally valid? If not, are all optimal tests conditionally valid?
- Optimality statements against composite alternatives?
- Extensions of power results to high dimensions? Some results for lasso test statistics available for knockoffs[4] and for CRT[5].
- Further connections with causal inference and with existing asymptotic (doubly robust, semiparametric) inference?

[4] Weinstein et al. (2017, 2020), Fan et al. (2019), Liu and Rigollet (2019)
[5] Celentano et al. (2020)

Future work

Many questions still remain open:

- Are all valid tests under MX also conditionally valid? If not, are all optimal tests conditionally valid?
- Optimality statements against composite alternatives?
- Extensions of power results to high dimensions? Some results for lasso test statistics available for knockoffs[4] and for CRT[5].
- Further connections with causal inference and with existing asymptotic (doubly robust, semiparametric) inference?

These lines of inquiry can improve our understanding of MX methodologies and help guide their development in the future.

[4] Weinstein et al. (2017, 2020), Fan et al. (2019), Liu and Rigollet (2019)
[5] Celentano et al. (2020)