The conditional permutation test for independence while controlling for confounders

Tom Berrett
CREST, ENSAE, Institut Polytechnique de Paris

Young Data Science Researchers Seminar

May 22nd, 2020

Collaborators

Material in talk based on joint work with



Yi Wang, Rina Foygel Barber and Richard Samworth

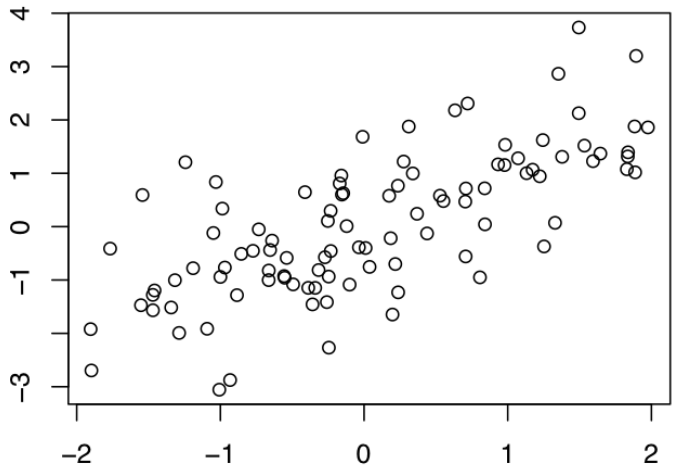
Background

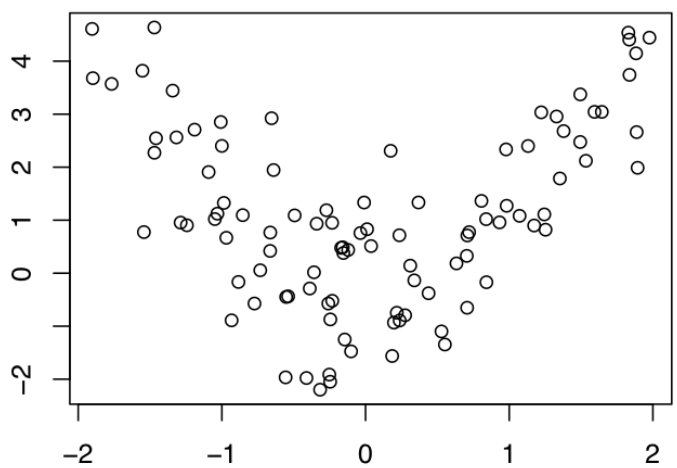
Measuring dependence and testing independence are fundamental problems in statistics, and are essential for model building, certain goodness-of-fit tests, feature selection, independent component analysis and more.

Classical measures include:

- Pearson's correlation (e.g. Pearson, 1920);
- Kendall's tau (Kendall, 1938);
- Hoeffding's D (Hoeffding, 1948).

These are limited to linear or monotonic dependence, or bivariate settings.





Measures of dependence

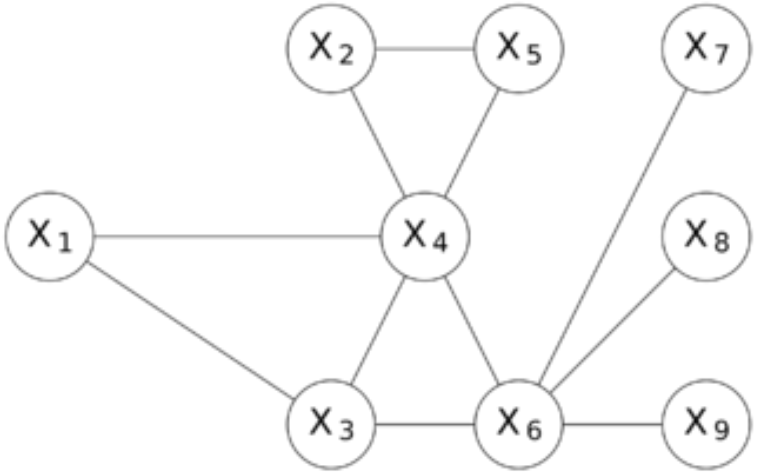
Modern datasets often exhibit complex dependence which is not well captured by these classical measures.

As a result, many new measures and tests have been proposed and studied recently:

- HSIC (Gretton et al., 2005; Sejdinovic et al., 2013; Pfister et al., 2018; Meynaoui et al., 2019);
- Distance covariance (Székely, Rizzo and Bakirov, 2007; Székely and Rizzo, 2013);
- Nearest neighbour methods (B. and Samworth, 2019);
- Multivariate rank-based tests (Weihs et al., 2017; Shi, Drton and Han, 2019; Deb and Sen, 2019);
- Empirical copula processes (Kojadinovic and Holmes, 2009);
- Sample space partitioning (Gretton and Györfi, 2010; Heller et al., 2016).

Conditional dependence

Moreover, in practice, it is often conditional independence that is of primary interest.



In GLMs for a response Y regressed on a high-dimensional feature vector $(X, Z) = (X, Z^1, \ldots, Z^p)$, the regression coefficient of X is zero if and only if $H_0 : X \perp\!\!\!\perp Y|Z$.

Conditional independence tests are well-developed within standard parametric models.

Measures of conditional dependence

In the regression setting, we can model the relationships of X and Y on Z and look for correlation between the residuals (Belloni et al., 2014; Shah and Peters, 2019).

Many tests of independence also have counterparts in the conditional independence setting (partial distance covariance, conditional kernel methods, partial copulas etc.)).

Critical values are typically found through asymptotic theory, which may not be reliable under model misspecification or with smaller sample sizes.

Problem statement

Suppose we observe an i.i.d. sample $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$ taking values in a space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Our aim is to test the null hypothesis

$$H_0 : X \perp\!\!\!\perp Y | Z$$

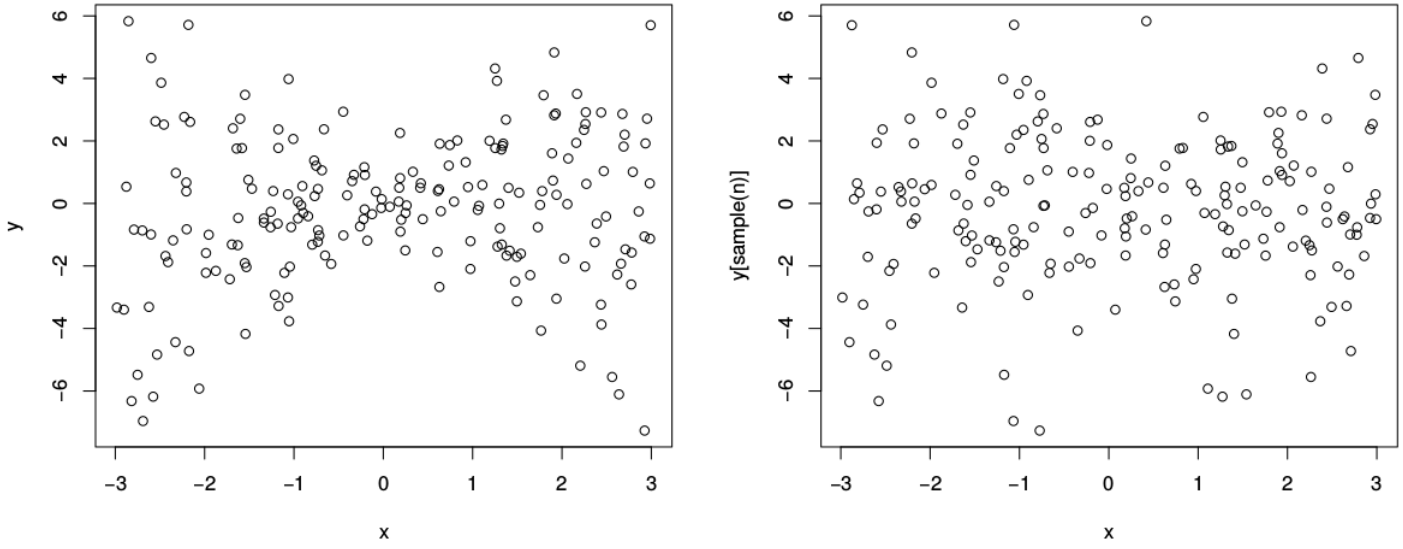
of conditional independence, or, equivalently, that $f_{XYZ} = f_{X|Z} f_{Y|Z} f_Z$.

Our results are very general, but we may think of $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and $\mathcal{Z} = \mathbb{R}^p$ for some (large) p.

Given the data $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and any test statistic $T : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n \to \mathbb{R}$, our procedure will output a p-value.

Permutation tests

In the simpler problem of testing $H_0 : X \perp\!\!\!\perp Y$, a practical and popular approach is to carry out a permutation test (e.g. Pitman, 1938; Fisher, 1935).



For any test statistic $T : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$ and i.i.d. uniformly random permutations $\pi_1, \ldots, \pi_M \in \mathcal{S}_n$ we can set $X_i^{(m)} = X_{\pi_m(i)}$ and calculate the p-value

$$P = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\{T(\mathbf{X}^{(m)}, \mathbf{Y}) \geq T(\mathbf{X}, \mathbf{Y})\}}{1 + M}.$$

Size of permutation tests

Under H_0, the datasets $(\mathbf{X}, \mathbf{Y}), (\mathbf{X}^{(1)}, \mathbf{Y}), \ldots, (\mathbf{X}^{(B)}, \mathbf{Y})$ are exchangeable.

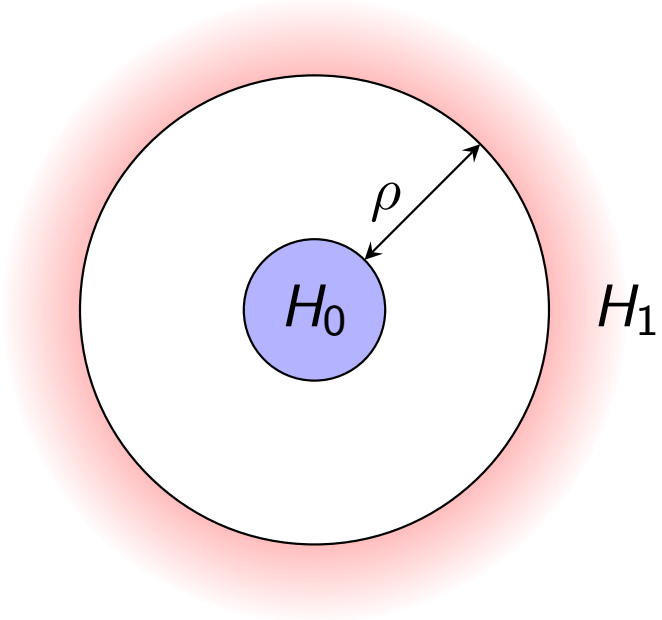
Hence, the statistics $T(\mathbf{X}, \mathbf{Y}), T(\mathbf{X}^{(1)}, \mathbf{Y}), \ldots, T(\mathbf{X}^{(B)}, \mathbf{Y})$ are exchangeable.

The rank of $T(\mathbf{X}, \mathbf{Y})$ among $T(\mathbf{X}, \mathbf{Y}), T(\mathbf{X}^{(1)}, \mathbf{Y}), \ldots, T(\mathbf{X}^{(B)}, \mathbf{Y})$ is uniformly distributed on $\{1, \ldots, B+1\}$, and so $\mathbb{P}(P \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$.

Permutation tests

Without confounding variables, we can easily construct permutation independence tests with exact, *assumption-free* Type I error control.

In fact, such tests with well-chosen test statistics can also have small Type II error. They can sometimes be proved to have desirable asymptotic properties (e.g. Lehmann and Romano, 2005), and can even be minimax rate optimal (B., Kontoyiannis and Samworth, 2020).



$$H_0 : X \perp\!\!\!\perp Y$$

$$H_1 : D(P, P_X P_Y) \geq \rho, s\text{-smooth}$$

$$\rho^* \asymp n^{-\frac{2s}{4s+d}}$$

Hardness of conditional independence testing

This is not the case for conditional independence testing, where it is generally not even possible to control Type I error uniformly.

Let $\mathcal{P}$ be the class of continuous null ($X \perp\!\!\!\perp Y|Z$) distributions of (X, Y, Z) and let $\mathcal{Q}$ be the class of continuous alternatives ($X \not\perp\!\!\!\perp Y|Z$).

Theorem (Shah and Peters, 2019)

For $n \in \mathbb{N}$ and $\alpha \in (0, 1)$, let ψ_n be a test with $\sup_{P \in \mathcal{P}} \mathbb{P}_P(\psi_n = 1) \leq \alpha$. Then

$$\sup_{Q \in \mathcal{Q}} \mathbb{P}_Q(\psi_n = 1) \leq \alpha.$$

If Z is discrete with a small alphabet size we may be able to split the sample and test, but we can see that this is not possible for general distributions of Z.

Model-X framework

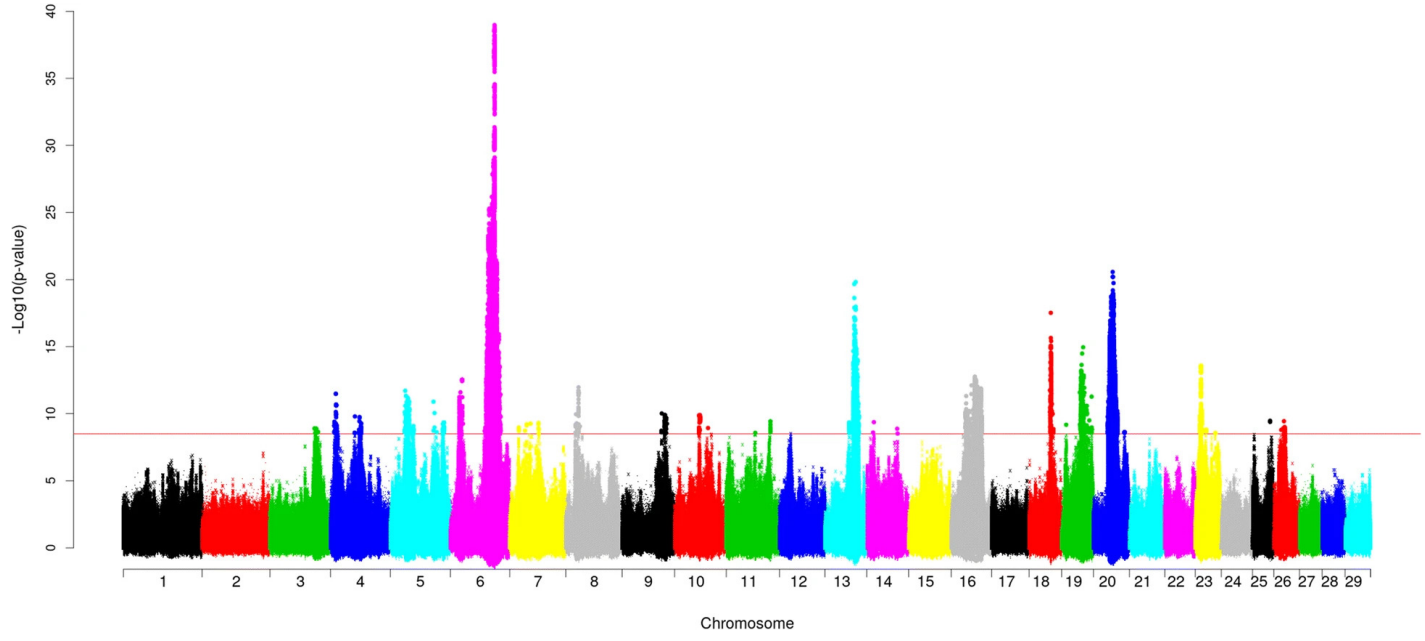
According to Shah and Peters (2019), it is necessary to make assumptions when testing conditional independence.

We adopt the 'Model-X' framework of Candès et al. (2018) and assume that we have an approximation $Q(\cdot|z)$ to the conditional distribution of $X|Z = z$. With this extra information, we will see that we can restore finite-sample Type I error control.

Formally, we will only require our tests to control the Type I error for distributions of (X, Y, Z) that are (approximately) consistent with $X|Z = z \sim Q(\cdot|z)$.

Model-X framework

The Model-X assumption has become popular in settings where we want to avoid making any assumptions on Y (Candès et al., 2018; Barber and Candès, 2018; Tansey et al., 2018; Romano et al., 2019) .



In some cases (X, Z) data is abundant while labeled data (X, Y, Z) is relatively scarce.

1. The conditional permutation test
2. Robustness and comparison with the conditional randomization test
3. Bikeshare data set example
4. Computation

The conditional permutation test

Let $\mathbf{X}_{()} = (X_{(1)}, \ldots, X_{(n)})$ denote the order statistics of $\mathbf{X}$, and let $\Pi \in \mathcal{S}_n$ denote the ranks so that Π satisfies $X_i = X_{(\Pi(i))}$. Then, under H_0,

$$\mathbb{P}(\Pi = \pi | \mathbf{X}_{()}, \mathbf{Y}, \mathbf{Z}) = \frac{\prod_{i=1}^n q(X_{(\pi(i))}|Z_i)}{\sum_{\pi'} \prod_{i=1}^n q(X_{(\pi'(i))}|Z_i)} = \frac{q^n(\mathbf{X}_{(\pi)}|\mathbf{Z})}{\sum_{\pi'} q^n(\mathbf{X}_{(\pi')}|\mathbf{Z})}$$

if $Q(\cdot|z)$ is the true conditional distribution of $X|Z = z$ with density $q(\cdot|z)$.

The conditional permutation test

Let $\mathbf{X}_{()} = (X_{(1)}, \ldots, X_{(n)})$ denote the order statistics of $\mathbf{X}$, and let $\Pi \in \mathcal{S}_n$ denote the ranks so that Π satisfies $X_i = X_{(\Pi(i))}$. Then, under H_0,

$$\mathbb{P}(\Pi = \pi | \mathbf{X}_{()}, \mathbf{Y}, \mathbf{Z}) = \frac{\prod_{i=1}^n q(X_{(\pi(i))}|Z_i)}{\sum_{\pi'} \prod_{i=1}^n q(X_{(\pi'(i))}|Z_i)} = \frac{q^n(\mathbf{X}_{(\pi)}|\mathbf{Z})}{\sum_{\pi'} q^n(\mathbf{X}_{(\pi')}|\mathbf{Z})}$$

if $Q(\cdot|z)$ is the true conditional distribution of $X|Z = z$ with density $q(\cdot|z)$.

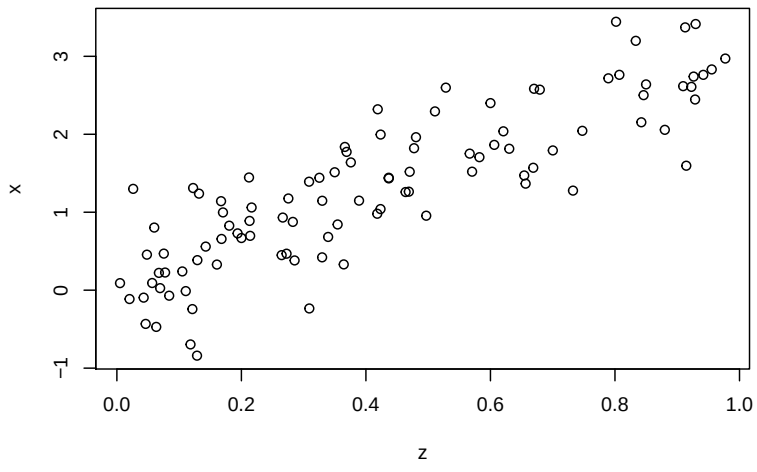
We can draw $\pi_1, \ldots, \pi_M$ independently from the same distribution

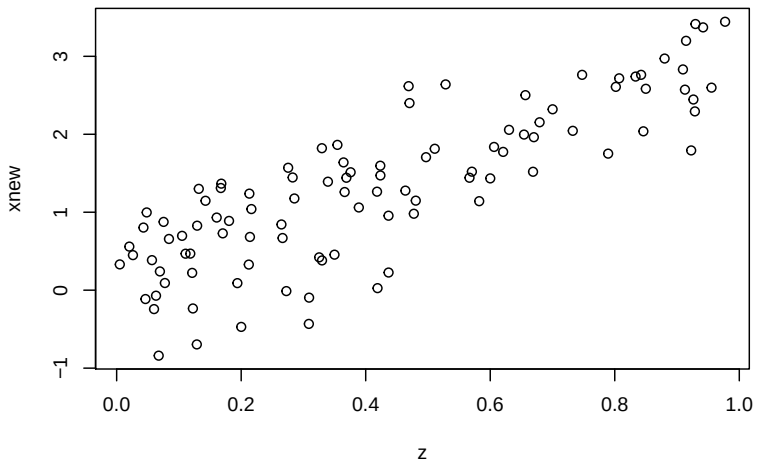
$$\mathbb{P}(\pi_m = \pi | \mathbf{X}_{()}, \mathbf{Y}, \mathbf{Z}) = \frac{q^n(\mathbf{X}_{(\pi)}|\mathbf{Z})}{\sum_{\pi'} q^n(\mathbf{X}_{(\pi')}|\mathbf{Z})}$$

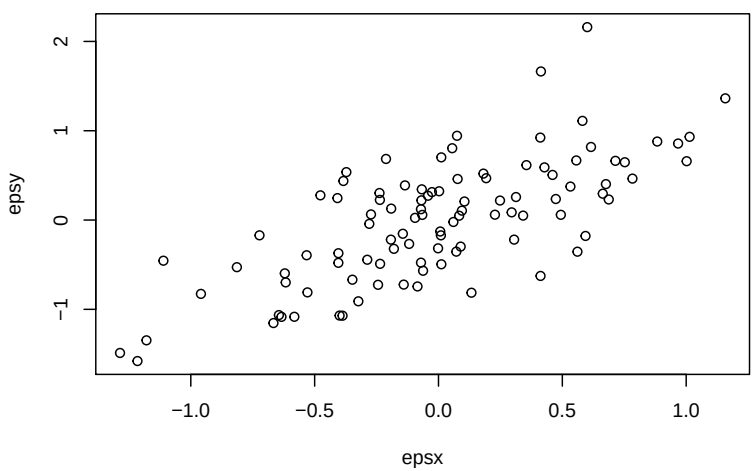
and set $X_i^{(m)} = X_{(\pi_m(i))}$. Conditional on $\mathbf{X}_{()}, \mathbf{Y}, \mathbf{Z}$ and under H_0, then $\Pi, \pi_1, \ldots, \pi_M$ are i.i.d.

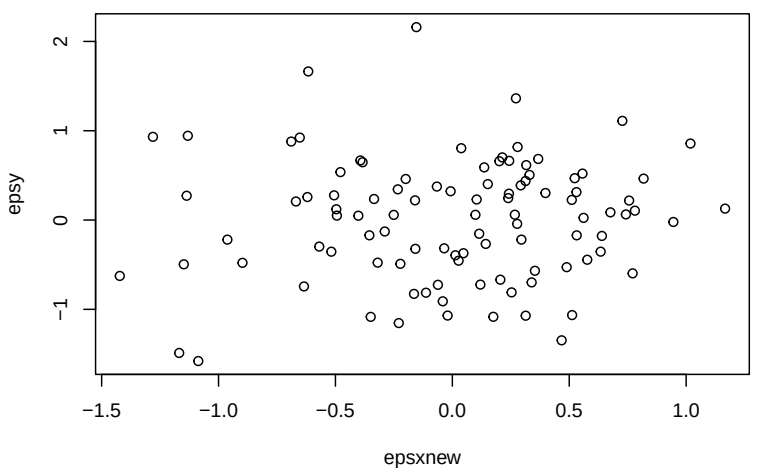
The conditional permutation test

Consider $X|Z \sim \mathcal{N}(\beta Z, \sigma^2)$ and $Y|X, Z \sim \mathcal{N}(\beta Z + \gamma X, \sigma^2)$. We can compare the residuals $\hat{\epsilon}_X$ and $\hat{\epsilon}_Y$ after regressing X and Y on Z.









The conditional permutation test

Under H_0 and assuming that $Q(\cdot|z)$ is the true conditional distribution of $X|Z = z$, the sequence

$$(\mathbf{X}, \mathbf{Y}, \mathbf{Z}), (\mathbf{X}^{(1)}, \mathbf{Y}, \mathbf{Z}), \ldots, (\mathbf{X}^{(M)}, \mathbf{Y}, \mathbf{Z})$$

is exchangeable.

Under these conditions, given any test statistic $T : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n \to \mathbb{R}$, the random variable

$$P = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\{T(\mathbf{X}^{(m)}, \mathbf{Y}, \mathbf{Z}) \geq T(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\}}{1 + M}$$

satisfies $\mathbb{P}(P \leq \alpha) \leq \alpha$.

The conditional randomization test

The *conditional randomization test* (CRT) of Candès et al. (2018) is an approach in the same framework that draws

$$\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)} | \mathbf{X}, \mathbf{Y}, \mathbf{Z} \overset{i.i.d.}{\sim} Q^n(\cdot|\mathbf{Z}) = Q(\cdot|Z_1) \times \ldots \times Q(\cdot|Z_n)$$

without the restriction that the $\mathbf{X}^{(m)}$ be reorderings of $\mathbf{X}$.

Our conditional permutation test (CPT) is similar to this, and can be thought of as drawing $\mathbf{X}^{(m)}$ from the distribution $Q^n(\cdot|\mathbf{Z})$ conditional on the event that $\mathbf{X}^{(m)}$ has the same order statistics as $\mathbf{X}$.

Comparing the CPT and CRT

The CPT forces the $\mathbf{X}^{(m)}$ to be more similar to $\mathbf{X}$ than with the CRT.

If we only know $q(\cdot|z)$ up to base measure and normalizing constant, i.e. the truth is $q^*(x|z) = q(x|z)h(x)c(z)$, then the permutation distribution

$$\frac{\prod_{i=1}^n q(X_{(\pi(i))}|Z_i)}{\sum_{\pi'} \prod_{i=1}^n q(X_{(\pi'(i))}|Z_i)} = \frac{\prod_{i=1}^n q^*(X_{(\pi(i))}|Z_i)}{\sum_{\pi'} \prod_{i=1}^n q^*(X_{(\pi'(i))}|Z_i)}$$

is correct. In this way the CPT is more robust than the CRT.

If, e.g., we have a semiparametric model in which

$$q_*(x|z) = \exp(x \cdot z^T\theta - f(x) - g(z))$$

then it will be significantly easier to estimate q_* up to base measure than to estimate all of q_*.

Robustness

The input conditional distribution $Q(\cdot|z)$ will generally be different to the true conditional distribution $Q_*(\cdot|z)$, and the Type I error control may not hold exactly.

Theorem

Under H_0, for any test statistic T and significance level $\alpha \in [0,1]$ we have for both the CPT and CRT that

$$\mathbb{P}(P \leq \alpha|\mathbf{Y},\mathbf{Z}) \leq \alpha + d_{\mathrm{TV}}\big(Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\big),$$

where $d_{\mathrm{TV}}(Q_1,Q_2) := \sup_A |Q_1(A) - Q_2(A)|$ is the total variation distance.

If we can choose Q with $d_{\mathrm{TV}}\big(Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\big) = o(1)$ then we have an approximately valid test.

Sketch of proof for CRT

Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be our data, and let $\check{\mathbf{X}}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ be drawn i.i.d. from $Q^n(\cdot|\mathbf{Z})$ independent of $\mathbf{X}, \mathbf{Y}$. Let $A_\alpha \subseteq (\mathcal{X}^n)^{M+1}$ be given by

$$A_\alpha = \left\{ (\mathbf{x}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}) : \frac{1 + \sum_{m=1}^{M} \mathbb{1}\{T(\mathbf{x}^{(m)}, \mathbf{Y}, \mathbf{Z}) \geq T(\mathbf{x}, \mathbf{Y}, \mathbf{Z})\}}{1 + M} \leq \alpha \right\},$$

the rejection region of the CRT.

Sketch of proof for CRT

Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be our data, and let $\check{\mathbf{X}}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ be drawn i.i.d. from $Q^n(\cdot|\mathbf{Z})$ independent of $\mathbf{X}, \mathbf{Y}$. Let $A_\alpha \subseteq (\mathcal{X}^n)^{M+1}$ be given by

$$A_\alpha = \left\{(\mathbf{x}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}) : \frac{1 + \sum_{m=1}^{M} \mathbb{1}\{T(\mathbf{x}^{(m)}, \mathbf{Y}, \mathbf{Z}) \geq T(\mathbf{x}, \mathbf{Y}, \mathbf{Z})\}}{1 + M} \leq \alpha \right\},$$

the rejection region of the CRT. We then have

$$\begin{aligned}
\mathbb{P}(P \leq \alpha|\mathbf{Y}, \mathbf{Z}) &= \mathbb{P}\big((\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}) \in A_\alpha|\mathbf{Y}, \mathbf{Z}\big) \\
&\leq \mathbb{P}\big((\check{\mathbf{X}}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}) \in A_\alpha|\mathbf{Y}, \mathbf{Z}\big) \\
&\qquad + d_{\mathrm{TV}}\Big(\big((\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)})|\mathbf{Y}, \mathbf{Z}\big), \big((\check{\mathbf{X}}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)})|\mathbf{Y}, \mathbf{Z}\big)\Big) \\
&= \mathbb{P}\big((\check{\mathbf{X}}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}) \in A_\alpha|\mathbf{Y}, \mathbf{Z}\big) + d_{\mathrm{TV}}\big(Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\big) \\
&\leq \alpha + d_{\mathrm{TV}}\big(Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\big),
\end{aligned}$$

where the final bound follows by exchangeability.

Robustness

If Q_* belongs to a k parameter family and we have access to an unlabeled sample of size N then we will typically have a valid test when $N \gg nk$. Suppose

$$X|Z = z \sim \mathcal{N}(z^T \beta_*, \sigma_*^2)$$

for some unknown β_* and σ_* that we estimate by the maximum likelihood estimators $\hat{\beta}$ and $\hat{\sigma}^2$. Then

$$d_{\mathrm{TV}}^2\big(Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\big) \leq \frac{1}{2}\sum_{i=1}^n d_{\mathrm{KL}}\big(Q_*(\cdot|Z_i), Q(\cdot|Z_i)\big)$$

$$= \frac{n}{2}\left(\log\frac{\hat{\sigma}^2}{\sigma_*^2} + \frac{\sigma_*^2}{\hat{\sigma}^2} - 1\right) + \sum_{i=1}^n \frac{(Z_i^T\hat{\beta} - Z_i^T\beta_*)^2}{2\hat{\sigma}^2} = O_p\left(\frac{n(1+\mathbb{E}\|Z\|^2)}{N}\right).$$

Robustness

As a second example, suppose $\mathcal{X} = \{0, 1\}$ and we estimate the regression function $p_*(z) := \mathbb{P}(X = 1|Z = z)$. Then, under appropriate smoothness conditions we will be able to achieve

$$d_{\mathrm{TV}}^2\big(Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z})\big) \leq \frac{1}{2}\sum_{i=1}^{n} d_{\mathrm{KL}}\big(Q_*(\cdot|Z_i), Q(\cdot|Z_i)\big)$$
$$\asymp \sum_{i=1}^{n}\big\{\hat{p}(Z_i) - p_*(Z_i)\big\}^2 \lesssim nN^{-a_k},$$

where N^{-a_k} is the (minimax) rate of convergence depending on the ambient dimension k and the smoothness. When N is sufficiently large, our test will be approximately valid.

Lower bound

For the CRT, our upper bound is tight when M is large.

Theorem

Under H_0, there exists a test statistic T such that, for the CRT,

$$\sup_{\alpha \in [0,1]} \left\{ \mathbb{P}(P \leq \alpha | \mathbf{Y}, \mathbf{Z}) - \alpha \right\} \geq d_{\mathrm{TV}}\left(Q_*^n(\cdot|\mathbf{Z}), Q^n(\cdot|\mathbf{Z}) \right) - \frac{1 + o(1)}{2} \sqrt{\frac{\log(M)}{M}}$$

as $M \to \infty$.

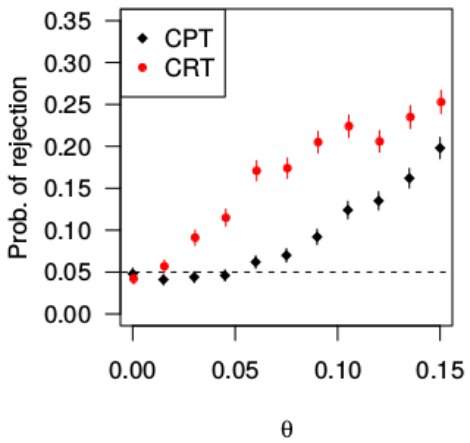
Combining this result with the previous upper bound, for the worst case test statistic the CPT is at least as robust as the CRT. We see in practice that the CPT is often much more robust.

Gaussian simulations under the null

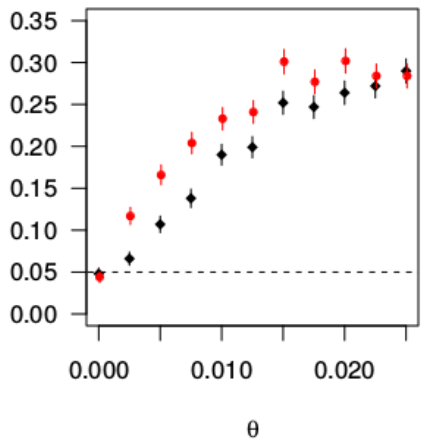
With $n = 50, p = 20$ and $a, b \sim \mathcal{N}_p(0, I_p)$ suppose that

$$Z \sim \mathcal{N}_p(0, I_p), \quad X|Z \sim \mathcal{N}(\mu(b^T Z), 1) \quad Y|X, Z \sim \mathcal{N}(p^{-1} a^T Z, 1)$$

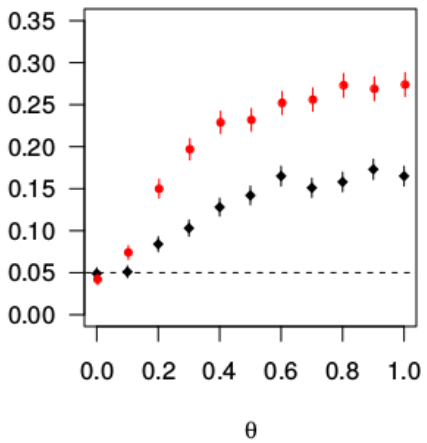
for some function μ, so that H_0 holds. However, suppose we take $Q(\cdot|z) = \mathcal{N}(b^T z, 1)$, so that our model is misspecified unless $\mu(\eta) = \eta$, and we use the test statistic $T(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = |\mathrm{Corr}(\mathbf{X}, \mathbf{Y})|$.



$\mu(\eta) = \eta + \theta\eta^2$



$\mu(\eta) = \eta - \theta\eta^3$



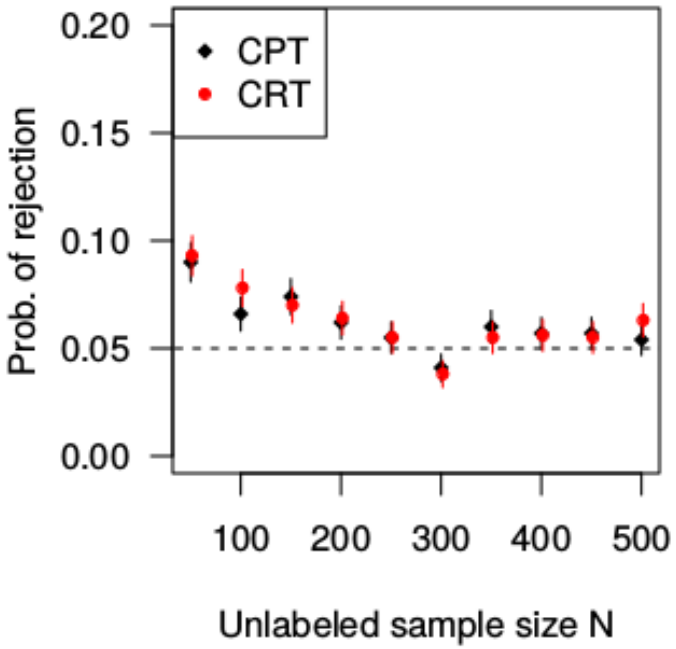
$\mu(\eta) = \tanh(\theta\eta)/\theta$

Estimation error simulations under the null

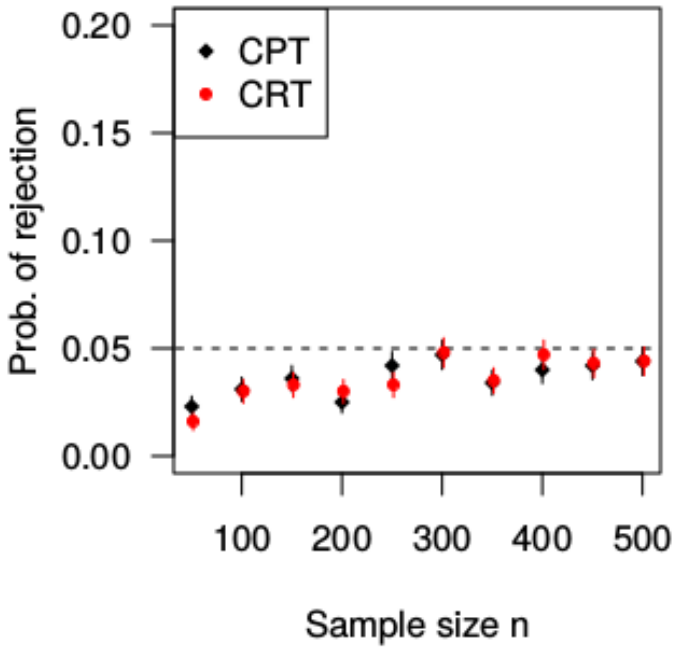
Suppose now that

$$Z \sim \mathcal{N}_p(0, I_p), \quad X|Z \sim \mathcal{N}(b^T Z, 1) \quad Y|X, Z \sim \mathcal{N}(p^{-1} a^T Z, 1)$$

but that b is unknown. We will use the value $\hat{b}$, where this estimator is either calculated using an unlabled sample $(X_i^{\text{unlab}}, Y_i^{\text{unlab}}), i = 1, \ldots, N$ or by reusing the test data.



(a) Trained on unlabeled data



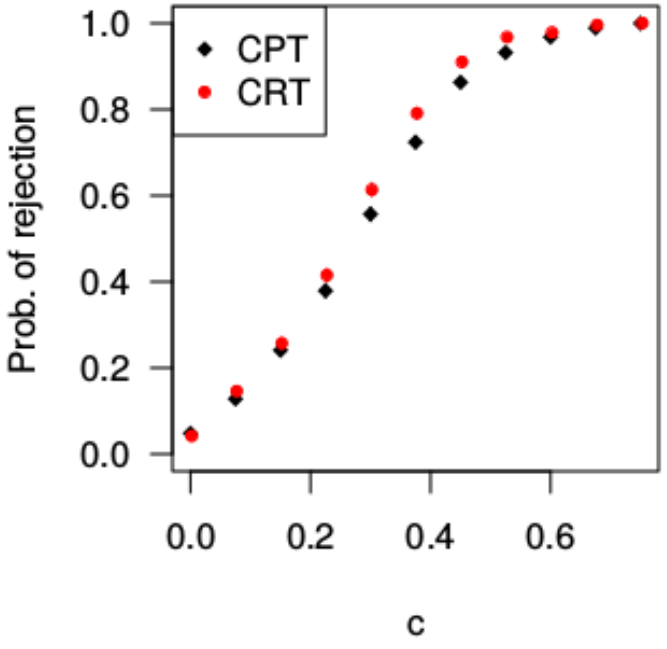
(b) Trained by reusing data

Power simulations

Now suppose that

$$Z \sim \mathcal{N}_p(0, I_p), \quad X|Z \sim \mathcal{N}(b^T Z, 1) \quad Y|X, Z \sim \mathcal{N}(a^T Z + cX, 1)$$

so that H_0 does not hold, and c controls the strength of the dependence.



The cost of the extra robustness is a slight decrease in power.

Capital bikeshare data set

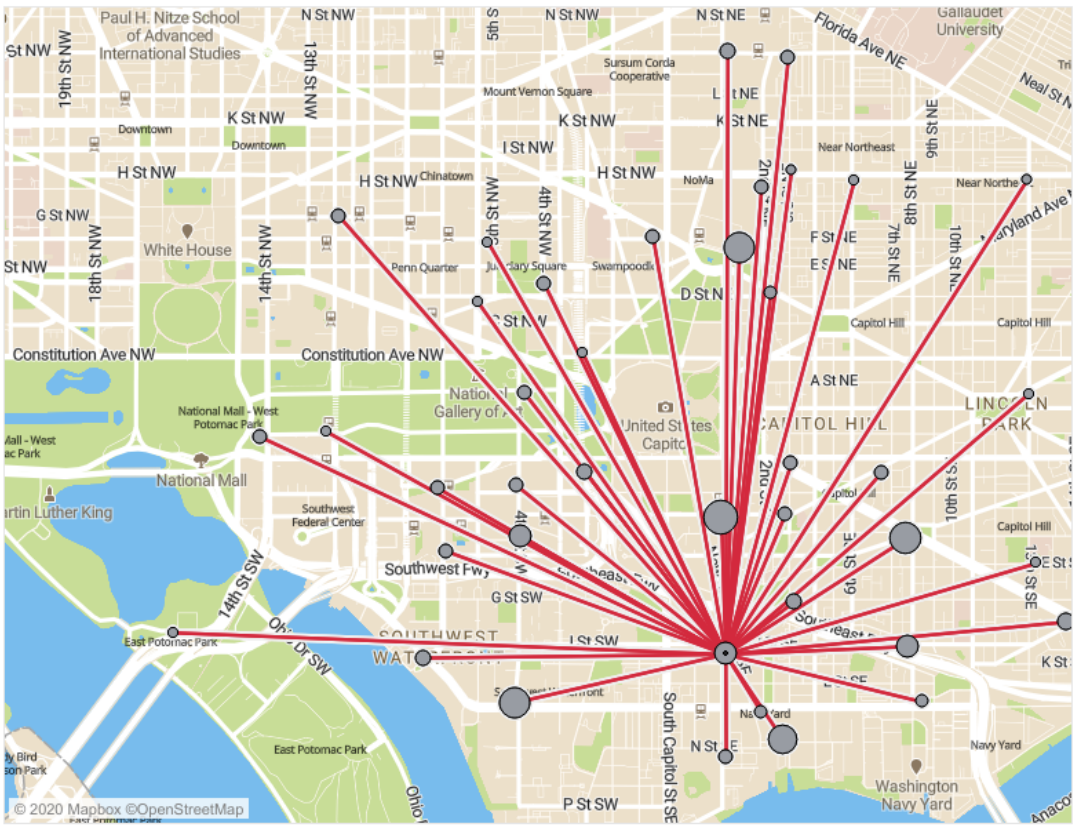
We implement the CPT and CRT on the Capital bikeshare data set. This contains each ride ever taken, recording the start time and location, end time and location, and a user type that can be 'Member' or 'Casual'.

capital bikeshare TOP DESTINATIONS BY STATION

Select Starting Station: 1st & K St SE
Weekend/Weekday: All
Show Top Destinations: 40

1st & K St SE

Top 40 Destinations : All Days



Trip Counts from 1st & K St SE

Top 40 Destinations: All Days

Mouse over to view route on map. All times in minutes.

Destination	Total Trips	Avg. Ride Time	Fastest Time
3rd & D St SE	268	5.9	3.6
4th & C St SW	264	13.6	6.1
2nd & G St NE	262	13.2	7.7
4th & D St NW / Judiciary Square	260	13.8	7.8
Eastern Market / 7th & North Caroli..	259	9.9	4.9
North Capitol St & F St NW	257	13.6	8.2
Metro Center / 12th & G St NW	255	20.9	11.4
1st & N St SE	250	14.5	1.5
4th St & Madison Dr NW	245	13.2	6.6
Jefferson Dr & 14th St SW	237	28.7	9.8
L'Enfant Plaza / 7th & C St SW	237	10.7	6.2
7th & E St SW	228	9.9	5.8
3rd & M St NE	223	16.6	10.6
3rd St & Pennsylvania Ave SE	220	10.5	4.4
2nd St & Massachusetts Ave NE	210	13.9	7.3
M St & New Jersey Ave SE	208	18.7	1.4
Potomac Ave & 8th St SE	205	6.7	3.6
Lincoln Park / 13th & East Capitol St ..	165	14.4	7.9
Constitution Ave & 2nd St NW/DOL	163	13.3	7.7

All 2017 Capital Bikeshare trips. Data via Capital Bikeshare (https://s3.amazonaws.com/capitalbikeshare-data/index.html)

Capital bikeshare data set

We use the following data

- Test data: all rides taken on weekdays in Oct 2011. Sample size n=7,346 rides (after screening).
- Training data: all rides taken on weekdays in Sep and Nov 2011. Sample size N=149,912 rides.

Capital bikeshare data set

We use the following data

- Test data: all rides taken on weekdays in Oct 2011. Sample size n=7,346 rides (after screening).
- Training data: all rides taken on weekdays in Sep and Nov 2011. Sample size N=149,912 rides.

We take X to be the duration of the ride, Y the user type, date or day of the week, and Z the start and end locations and time of day. For our conditional distribution we use

$$Q(\cdot|z) = \mathcal{N}(\hat{\mu}(z), \hat{\sigma}^2(z)),$$

where $\hat{\mu}(z)$ and $\hat{\sigma}^2(z)$ are calculated using the training data for each combination of start and end location with a kernel weighting times of day.

Capital bikeshare data set

Writing $R_i = X_i - \hat{\mu}(Z_i)$ we take T to be the correlation between $\mathbf{R}$ and $\mathbf{Y}$ when Y is the user type. When Y is the day of the week we take T to be

$$\max_{y \in \{\text{Mon},\ldots,\text{Fri}\}} \Big| \mathrm{Corr}\big(\mathbf{R}, (\mathbb{1}\{Y_1 = y\}, \ldots, \mathbb{1}\{Y_n = y\})\big) \Big|.$$

With $M = 100$ we obtain the following average p-values over ten trials of the experiment:

Variable Y	CPT	CRT
User type	0.0010 (0.0000)	0.0010 (0.0000)
Date	0.1146 (0.0032)	0.1293 (0.0032)
Day of week	0.1980 (0.0037)	0.2063 (0.0032)

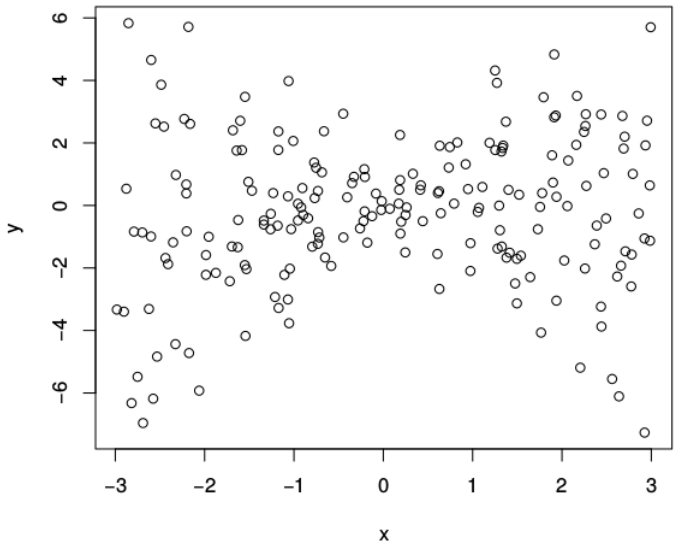
The CPT and CRT perform similarly here.

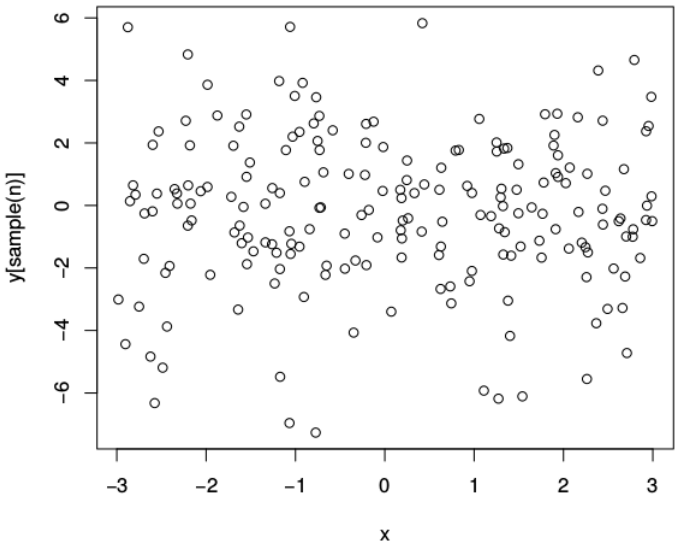
Computation

To implement the CPT we must be able to sample from the permutation distribution

$$\mathbb{P}(\pi_m = \pi | \mathbf{X}_{()}, \mathbf{Y}, \mathbf{Z}) = \frac{q^n(\mathbf{X}_{(\pi)}|\mathbf{Z})}{\sum_{\pi'} q^n(\mathbf{X}_{(\pi')}|\mathbf{Z})}$$

This distribution is highly non-uniform, and non-trivial to sample from.





Computation

If we ran a Metropolis–Hastings algorithm then the acceptance odds ratio

$$\frac{\prod_{i=1}^{n} q(X_{(\pi'(i))}|Z_i)}{\prod_{i=1}^{n} q(X_{(\pi(i))}|Z_i)} = \frac{q^n(\mathbf{X}_{(\pi')}|\mathbf{Z})}{q^n(\mathbf{X}_{(\pi)}|\mathbf{Z})}$$

would be extremely small for nearly all proposals π'. Uniform proposals would result in extremely slow mixing.

A more efficient algorithm is given by the Metropolis–Hastings algorithm with proposals of the form $\pi' = \pi \circ \sigma_{ij}$, where π is the current state and σ_{ij} is the transposition of i and j. The acceptance odds ratio is then

$$\frac{q(X_{(\pi(j))}|Z_i) \cdot q(X_{(\pi(i))}|Z_j)}{q(X_{(\pi(i))}|Z_i) \cdot q(X_{(\pi(j))}|Z_j)}.$$

This can be parallelized by proposing $\lfloor n/2 \rfloor$ transpositions at the same time.

Parallelized algorithm

Algorithm 1 Parallelized pairwise sampler for the CPT

Input: Initial permutation $\Pi^{[0]}$, integer $S \geq 1$.

for $s = 1, 2, \ldots, S$ **do**

Sample uniformly without replacement from $\{1, \ldots, n\}$ to obtain disjoint pairs

$$(i_{s,1}, j_{s,1}), \ldots, (i_{s,\lfloor n/2 \rfloor}, j_{s,\lfloor n/2 \rfloor}).$$

Draw independent Bernoulli variables $B_{s,1}, \ldots, B_{s,\lfloor n/2 \rfloor}$ with odds ratios

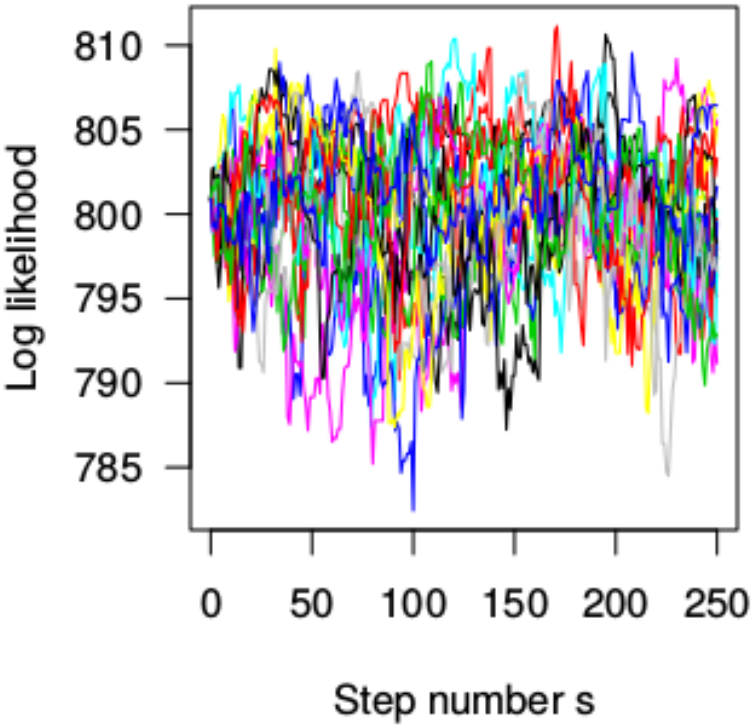
$$\frac{\mathbb{P}(B_{s,k} = 1)}{\mathbb{P}(B_{s,k} = 0)} = \frac{q(X_{(\Pi^{[s-1]}(j_{s,k}))} | Z_{i_{s,k}}) \cdot q(X_{(\Pi^{[s-1]}(i_{s,k}))} | Z_{j_{s,k}})}{q(X_{(\Pi^{[s-1]}(i_{s,k}))} | Z_{i_{s,k}}) \cdot q(X_{(\Pi^{[s-1]}(j_{s,k}))} | Z_{j_{s,k}})}.$$

Define $\Pi^{[s]}$ by swapping $\Pi^{[s-1]}(i_{s,k})$ and $\Pi^{[s-1]}(j_{s,k})$ for each k with $B_{s,k} = 1$.

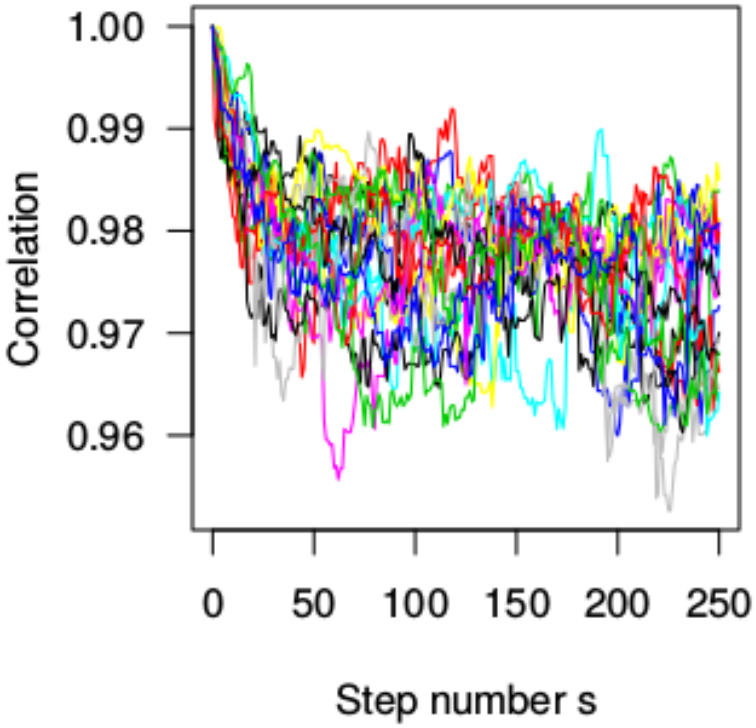
end for

Mixing time

The algorithm mixes quickly, and $S = 50$ seems to be a sufficient number of iterations.



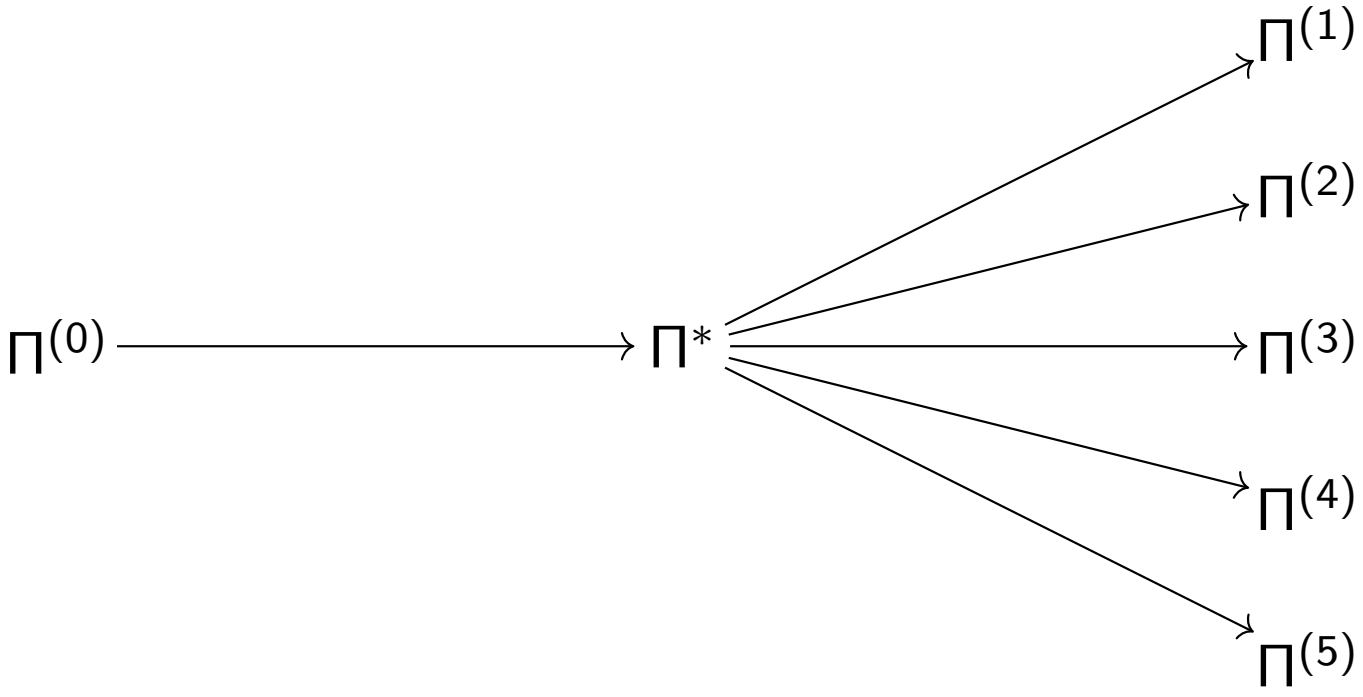
(a) Log-likelihood of $\mathbf{X}^{[s]}$



(b) $\text{Corr}(\mathbf{X}, \mathbf{X}^{[s]})$

Star-shaped sampler

Because of the weak dependence between $\Pi^{[0]}, \Pi^{[S]}, \Pi^{[2S]}, \ldots, \Pi^{[MS]}$, the data $\mathbf{X}, \mathbf{X}^{[S]}, \ldots, \mathbf{X}^{[MS]}$ are not exactly exchangeable under H_0. This can be corrected with a slightly different sampler.



This produces exchangeable data with any value of S, under H_0 and assuming $Q(\cdot|z)$ is correct.

Conclusion

- We introduce the conditional permutation test, which modifies the standard permutation test so that it may be used in a conditional setting if the relationship between X and Z is (approximately) known.
- We compare the CPT and the CRT in terms of their robustness to model misspecification, and find theoretical and numerical evidence to suggest that the CPT is more robust.
- We provide efficient implementations of the CPT.

Thank you!

B., Wang, Y., Barber, R. F. and Samworth, R. J. (2019) The conditional permutation test for independence while controlling for confounders. *J. Roy. Statist. Soc., Ser B,* **82**(1), 175–197.

References

Bach, F. R. and Jordan, M. I. (2002) Kernel independent component analysis. *J. Mach. Learn. Res.*, **3**, 1–48.

Barber, R. F. and Candès, E. (2018) A knockoff filter for high-dimensional selective inference. *Ann. Statist.*, **47**(5), 2504–2537.

B., and Samworth, R. J. (2019) Nonparametric independence testing via mutual information. *Biometrika*, **106**(3), 547–566.

B., Kontoyiannis, I. and Samworth, R. J. (2020) Optimal rates for independence testing via *U*-statistic permutation tests. `arXiv:2001.05513`.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014) Inference on treatment effects after selection among high-dimensional controls. `The Review of Economic Studies`, **81**(2), 608–650.

Candès, E., Fan, Y., Janson, L. and Lv, J. (2018) Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Statist. Soc. B*, **80**, 551–577.

Deb, N. and Sen, B. (2019). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Available at* `arXiv:1909.08733`.

Fisher, R. A. (1935) *The Design of Experiments* (1st Ed.). Oliver and Boyd, Edinburgh.

Gretton A., Bousquet O., Smola A. and Schölkopf B. (2005) Measuring Statistical Dependence with Hilbert-Schmidt Norms. *Algorithmic Learning Theory*, 63–77.

References

Gretton, A. and Györfi, L. (2010). Consistent nonparametric tests of independence. *J. Mach. Learn. Res.*, **11**, 1391–423.

Kendall, M. G. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–93.

Kojadinovic, I. and Holmes, M. (2009) Tests of independence among continuous random vectors based on Cramér–von Mises functionals of the empirical copula process. *J. Multivariate Anal.*, **100**, 1137–54.

Heller, R., Heller, Y., Kaufman, S., Brill, B. and Gorfine, M. (2016) Consistent distribution-free K-sample and independence tests for univariate random variables. *J. Mach. Learn. Res.*, **17**, 1–54.

Hoeffding, W. (1948) A non-parametric test of independence. *Ann. Math. Statist.*, **19**, 546–57.

Lehmann, E. L. and Romano, J. P. (2005) *Testing Statistical Hypotheses* (3rd Ed.). Springer, New York.

Meynaoui, A., Albert, M., Laurent, B. and Marrel, A. (2019) Adaptive test of independence based on HSIC measures. *Available at* `arXiv:1902.06441`.

Pearson, K. (1920) Notes on the history of correlation. *Biometrika*, **13**, 25–45.

Pfister, N., Bühlmann, P., Schölkopf, B. and Peters, J. (2018) Kernel-based tests for joint independence. *J. R. Statist. Soc. B.*, **80**, 5–31.

Pitman, E. J. G. (1938) Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, **29**, 322–335.

References

Romano, Y., Sesia, M. and Candès, E. (2019) Deep knockoffs. *J. Amer. Statist. Assoc., to appear.*

Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013) Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.*, **41**, 2263–2291.

Shah, R. D. and Peters, J. (2019) The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist., to appear*.

Shi, H., Drton, M., and Han, F. (2019) Distribution-free consistent independence tests via Hallin's multivariate rank. *Available at* `arXiv:1909.10024`.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. *Ann. Statist.*, **35**, 2769–2794.

Székely, G. J. and Rizzo, M. L. (2013) The distance correlation t-test of independence in high dimension. *J. Multivariate Anal.*, **117**, 193–213.

Tansey, W., Veitch, V., Zhang, H., Rabadan, R. and Blei, D. M. (2018) The holdout randomization test: Principled and easy black box feature selection. `arXiv:1811.00645`.

Weihs, L., Drton, M. and Meinshausen, N. (2017) SymRC: Estimating symmetric rank covariances. `https://github.com/Lucaweihs/SymRC`.