A New Test for Conditional Independence with High-dimensional Dependent Data

SONGYEN CHEN

June 9, 2025

Introduction

- The conditional independence (CI) assumption is a crucial condition for many studies with broad applications. It is of importance to investigate whether the independence between the covariate X for the response Y given a set of covariates W in modeling.
- Extensive efforts have been made to test CI between X and Y given $W = w$, such as:
  Su and White (2008) employ the weighted Hellinger distance to test for CI. Huang (2010) and Cheng and Huang (2012) consider the weighted average of maximal nonlinear correlation. Székely and Rizzo (2014) develop the partial distance correlation. Wang et al. (2015) develop a conditional distance correlation. Wang and Hong (2018) consider a characteristic function based testing. Zhang et al. (2011) propose a kernel-based conditional independence test in the reproducing kernel Hilbert spaces (RKHS). Runge (2018) and Ai et al. (2024) construct the CI testing based on the conditional mutual information. Zhu et al. (2020) construct the Blum–Kiefer–Rosenblatt (BKR)-based CI testing by transforming the CI problem into the UI problem.

Examples

- (Causal Inference: Treatment Effect Model)

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i \mid X_i,$$

where Y_i is the potential outcome, D_i is the binary treatment variable, and X_i is the control variable.

- (Instrumental Variable Model)

$$e_i \perp\!\!\!\perp Z_i \mid X_i,$$

where Z_i is the IV, e_i is the structural error, and X_i is the endogenous variable.

- (Directed acyclic graphs, DAG)

$$d - sep(X, Y \mid W),$$

X is d-separated from Y by W if all the paths between sets X and Y are blocked by elements of W.

In this work

- Rare works on CI problem in high-dimensional time series except for a novel study of Zhou et al. (2022), where they project the high-dimensional data onto low-dimensional subspaces.
- The limiting distribution of test statistic in Zhou et al. (2022) is nonstandard, and depends on the numerical re-construction of critical values, such as the bootstrap procedure.
- Highlight of our work: we develop a new CI test for high-dimensional time series, which has a pivotal, standard limiting distribution regardless of dimension of covariates.

Conditional Independence Test

Testing Framework

- Let $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, $W \in \mathbb{R}^d$, $d > n$, $d \to \infty$, $p, q < d$ can be fixed in low-dimension, or $pq \to \infty$.
- Our target:

$$H_o : X \perp\!\!\!\perp Y \mid W \quad (CI) \tag{1}$$

- Consider the projections of X and Y on W:

$$X = \mathbb{E}[X \mid W] + \xi,$$

$$Y = \mathbb{E}[Y \mid W] + \eta$$

- Under the strong exogeneity condition: $(\xi, \eta) \perp\!\!\!\perp W$, (1) is equivalent to:

$$H_o : \xi \perp\!\!\!\perp \eta \quad (UI) \tag{2}$$

$\xi \in \mathbb{R}^p$ and $\eta \in \mathbb{R}^q$.

- Many unconditional test based on dependency measures can be applied on (ξ, η), including:
  distance correlation: Székely et al. (2007), Székely and Rizzo (2013, 2014), Yao et al. (2018); projection correlation: Zhu et al. (2017); Blum–Kiefer–Rosenblatt (BKR) correlation: Zhu et al. (2018); Mutual information: Berrett and Samworth (2019), Ai et al. (2024); Hilbert-Schmidt-independence-criterion (HSIC): Gretton et al. (2007); ranked distance correlation: Heller et al. (2012)...etc.
- Drawbacks: curse of dimensionality in high dimension, nonstandard limiting distribution
- No well-developed work for high-dimensional dependent data (serially correlated time series).
- The distance correlation is $O(n^2)$, the lowest computation complexity above, so that we employ it in this work.

Close Work: Zhu et al. (2020)

- (Idea) Under the assumption: $\{F_1(X \mid W), F_2(Y \mid W)\} \perp\!\!\!\perp W$, (1) $\iff$ $F_1(X|W) \perp\!\!\!\perp F_2(Y|W)$ $\iff$ $\rho^{BKR} = 0$, where

$$\rho^{BKR} := \int \int \{F_{V,W}(v,w) - F_V(v)F_W(w)\}^2 dF_V(v) dF_W(w). \quad (3)$$

- Sample counterparts:

$$\widehat{\rho}^{BKR} = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \{F_n(\hat{V}_i, \hat{W}_j) - F_n(\hat{V}_i) F_n(\hat{W}_j)\}^2,$$

where $F_n(.,.)$ and $F_n(.)$ are the empirical distribution functions, $\hat{V}_j = \hat{F}_1(X_i \mid W_i) = n^{-1} \sum_{i=1}^{n} K_h(w_i - w) \boldsymbol{I}(X_i \leq x) / n^{-1} \sum_{i=1}^{n} K_h(w_i - w)$, $\hat{W}_j = \hat{F}_2(Y_j \mid W_j) = n^{-1} \sum_{j=1}^{n} K_h(w_j - w) \boldsymbol{I}(Y_j \leq y) / n^{-1} \sum_{j=1}^{n} K_h(w_j - w)$, and $K_h(w) = K(w/h)/h^d$ is the product of d univariate kernel functions.

- $n\widehat{\rho}^{BKR} \xrightarrow{\text{d}} \sum_{i,j}^{\infty} \chi^2_{ij}(1)/\pi^4 i^2 j^2$ under H_o.
- Disadvantage: when d grows with the sample size n, or $d \to \infty$, the testing is hard to work.

Bias-reduced Distance Correlation (Székely and Rizzo, 2013)

- (Distance covariance)

$$V(X,Y) = \int_{\mathbb{R}^{p+q}} |\varphi_{X,Y}(s,t) - \varphi_X(s)\varphi_Y(t)|^2 w(s,t) dsdt,$$

  where $w(s,t) = (c_p c_q ||s||^{p+1} ||t||^{q+1})$ with $c_p = \pi^{(p+1)/2}/\Gamma((p+1)/2)$.
- (Distance correlation)

$$DC(X,Y) = V(X,Y)/[V(X)V(Y)]^{1/2}, \tag{4}$$

  if $V(X)V(Y) > 0$; 0, if otherwise.
- (1) $\iff$ (2) $\iff$ $DC(\xi,\eta) = 0$.

- Bias-corrected estimator:

$$V_n^*(X,Y) = \frac{1}{n(n-3)} \sum_{k \neq \ell} A_{k,\ell}^* B_{k,\ell}^*,$$

where
$A_{k,\ell} = a_{k,\ell} - (n-2)^{-1} \sum_i a_{i,\ell} - (n-2)^{-1} \sum_j a_{k,j} + [(n-1)(n-2)]^{-1} \sum_{i,j} a_{i,j}$
with $a_{k,\ell} = ||X_k - X_\ell||$.

- $$R_n^*(X,Y) = \frac{V_n^*(X,Y)}{[V_n^*(X) V_n^*(Y)]^{1/2}}$$

if $V_n^*(X) V_n^*(Y) > 0$.

- Test statistic:

$$T_n = \sqrt{\frac{n(n-1)}{2}} R_n^*(\xi, \eta) \xrightarrow{\mathsf{d}} \mathcal{N}(0,1)$$

under H_o.

Time-series blocks smoothing

- Similar to the technique of blocks bootstrap (Hall,1985; Carlstein,1986; Künsch,1989), we block the original data to preserve the dependence among the underlying data.

A1. (exponential decaying rate) $Z_i := \{(X_i', Y_i', W_i')'\}$ is the weakly stationary and geometrically β-mixing time series with the mixing coefficient

$$\beta_i^Z(j) = \sup_t \left\{ \sup_{B \in \mathcal{F}_{i,t+j}^{\infty}} \left| \mathbb{P}(B \mid \mathcal{F}_{i,-\infty}^{t}) - \mathbb{P}(B) \right| \right\} = O(e^{-cj}), \forall c, j > 0$$

- Let Z_i^* be the independent copy of Z_i from a distribution F_Z. Let M and L be two integers denoting the block size and separation between adjacent blocks, respectively. Then, the total number of blocks is, $B = [(n - M)/L + 1]$, where $[.]$ is the integer truncation operator. For each $k. = 1, \ldots, B$, the kth data block

$$Z_k^* = (Z_{(k-1)L+1} : Z_{(k-1)L+M}).$$

Lemma

Assume that $M \to \infty$ as $n \to \infty$, $L = O(M)$, $M = O(n^{1/2-\epsilon})$, $L \leq M$, $M/L \to c \geq 1$, and $B = O(n^{1/2+\epsilon})$, $\epsilon \in (0, 1/2]$, Then the new sample $(b_1, \ldots, b_B)$ is asymptotically independent.

- $$X \perp\!\!\!\perp Y \mid W \iff X^* \perp\!\!\!\perp Y^* \mid W^* \iff \xi^* \perp\!\!\!\perp \eta^*$$

Sample-splitting

- What if $(\xi, \eta) \not\perp\!\!\!\perp W$?
- (mimic the strong exogeneity) To form a 2-fold random partitions of $\{1, 2, \ldots, B\}$ blocks by dividing them into two groups S_j and $S_j^c = [B] \setminus S_j$ at each j, with equal size $m = [B/2]$, $j = 1, 2, \ldots, J$, where $[.]$ is the integer symbol; for each set S_j, S_j^c is the set of all observation indices from B blocks that do not belong to S_j.
- Let $\hat{h}_j^x$ and $\hat{h}_j^y$ be the cross-fitting estimators by the machine learning approaches based on the training group S_j^c, and the testing based on the testing group S_j.
- $$(\hat{\xi}^*, \hat{\eta}^*)(S_j) \perp\!\!\!\perp \hat{h}(W^*)(S_j^c).$$
- Assumption **[A4]** ensures that

$$||(\hat{\xi}^*, \hat{\eta}^*) - (\xi^*, \eta^*)||_\infty = o(1).$$

Theoretical Properties

A2. $M = O(n^{1/(2+\tau)})$, for $\tau \in (0, 1]$, and $pq = o(n^{\tau/(2+\tau)})$.

A3. $\max_{1\le i\le p} \mathbb{E}\,|\xi_i|^2 < \Delta_1$, $\max_{1\le j\le q} \mathbb{E}\,|\eta_j|^2 < \Delta_2$,
$\max_{1\le i\le p} \mathbb{E}\,|\xi_i|^{4+4\tau} + \max_{1\le j\le q} \mathbb{E}\,|\eta_j|^{4+4\tau} < \infty$, and
$\max\left\{p^{-1}\sum_{i=1}^{p} \mathbb{E}\,|\xi_i|^{4+4\tau}, q^{-1}\sum_{i=1}^{q} \mathbb{E}\,|\eta_i|^{4+4\tau}\right\} < \Delta < \infty$.

A4. Let $h_o^x(W) = \mathbb{E}[X \mid W]$ and $h_o^y(W) = \mathbb{E}[Y \mid W]$.

$$||\hat{h}_j^x - h_o^x||_2 = o(n^{-1/4}),$$

$$||\hat{h}_j^y - h_o^y||_2 = o(n^{-1/4}),$$

Theorem

Under Assumptions **[A1]** to **[A4]**,

$$T_n^* = \sqrt{\frac{m(m-1)}{2}} R_n^*(\xi^*, \eta^*) \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1).$$

- With a predetermined significance level α, we reject the null hypothesis when

$$T_n^* > \Phi^{-1}(1 - \alpha/2).$$

Simulation

- DGP1: $Y_t = 0.5W_{t-1} + e_{1t}$ and $X_t = 0.5W_{t-1} + e_{2t}$.
- DGP2: $Y_t = \sqrt{h_{1t}}e_{1t}, h_{1t} = 0.01 + 0.1h_{1t-1} + 0.4W_{t-1}^2$,
  $X_t = \sqrt{h_{2t}}e_{2t}, h_{2t} = 0.01 + 0.1h_{2t-1} + 0.15W_{t-1}^2$.
- $(e_{1t}, e_{2t}) \sim \mathcal{N}(0, \Sigma)$, $\Sigma_{jk} = \rho = \{0, 0.2\}, j \neq k$ and $\Sigma_{jj} = 1$. $W \sim \mathcal{N}(0, \boldsymbol{I}_d)$.
- Replication 1000, significance levels $\alpha = 0.05$. $n = \{200, 400\}$.
  $p = 5, q = 10, d = 1000$
- $H_o : Y_t \perp\!\!\!\perp X_t \mid W_{t-1}$.
- LASSO for $h^x(.), h^y(.)$.

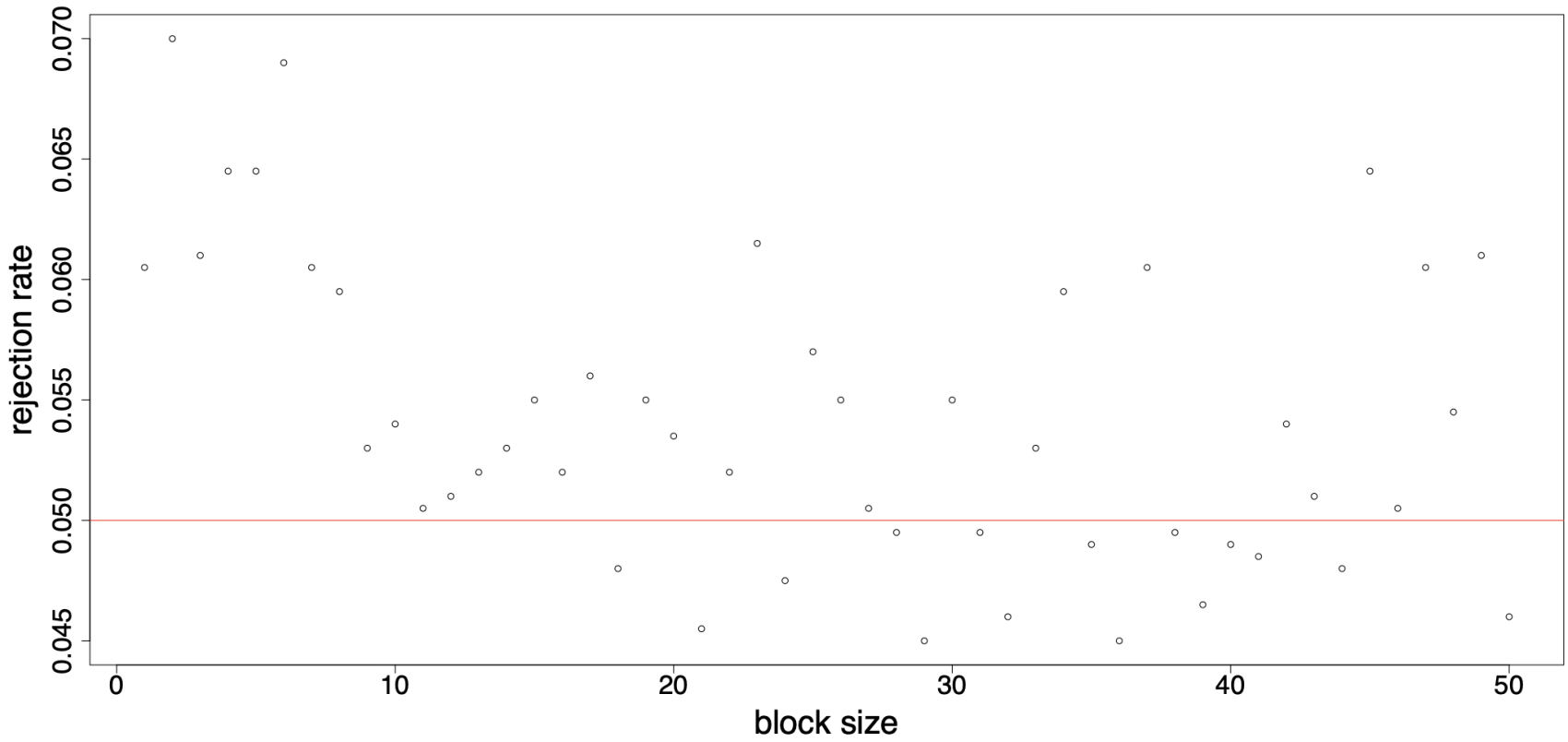

Figure: DGP1: Rejection rate relative to various block sizes M. $n = 1500$.

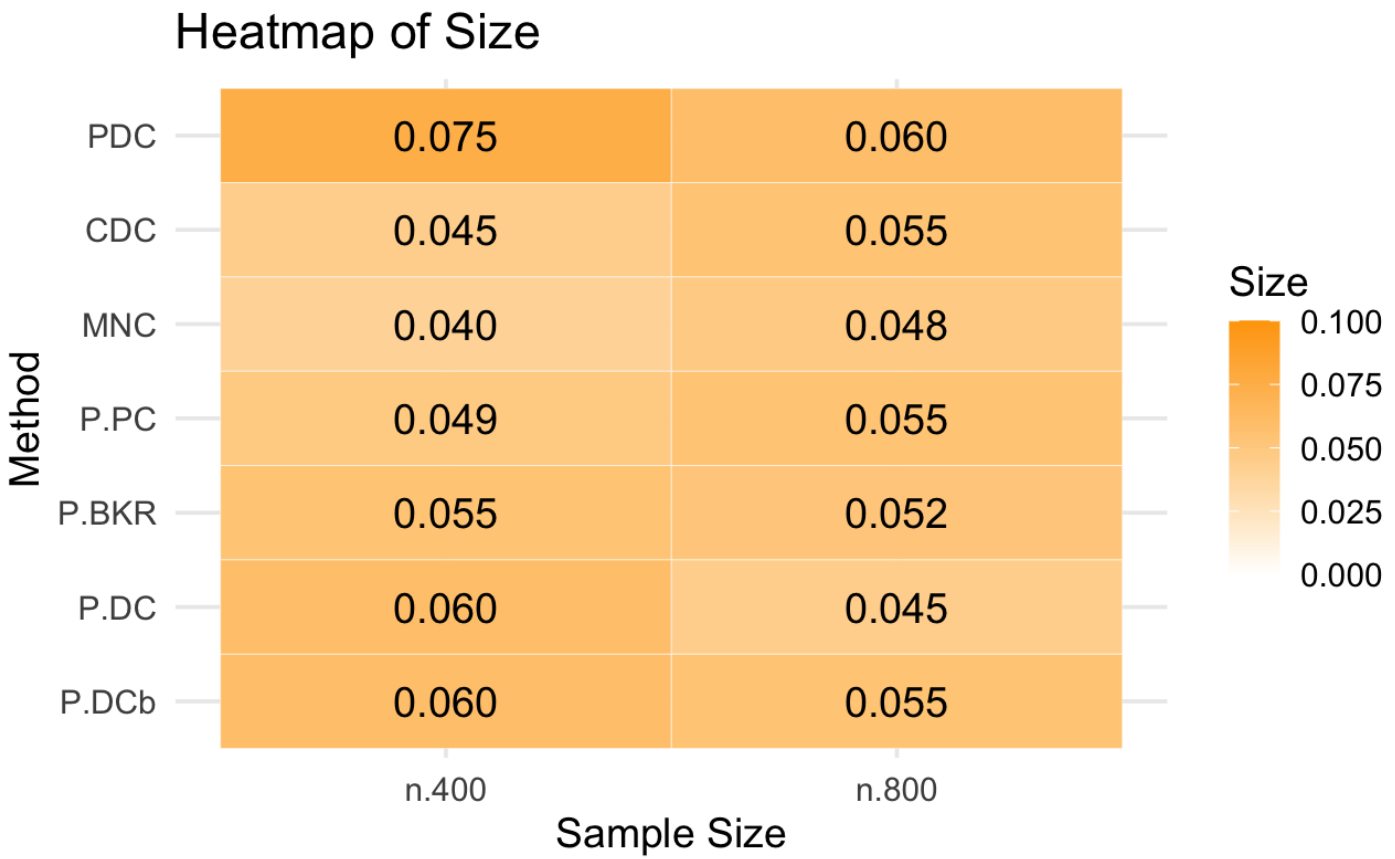

Figure: DGP1: Size of Tests for $\rho = 0$. 'PDC': partial distance correlation; 'CDC': conditional distance correlation; 'MNC': maximal nonlinear correlation; 'P.PC': projection+projection correlation; 'P.BKR': projection+BKR; 'P.DC': projection+distance correlation; 'P.DCb': projection+bias-corrected distance correlation.

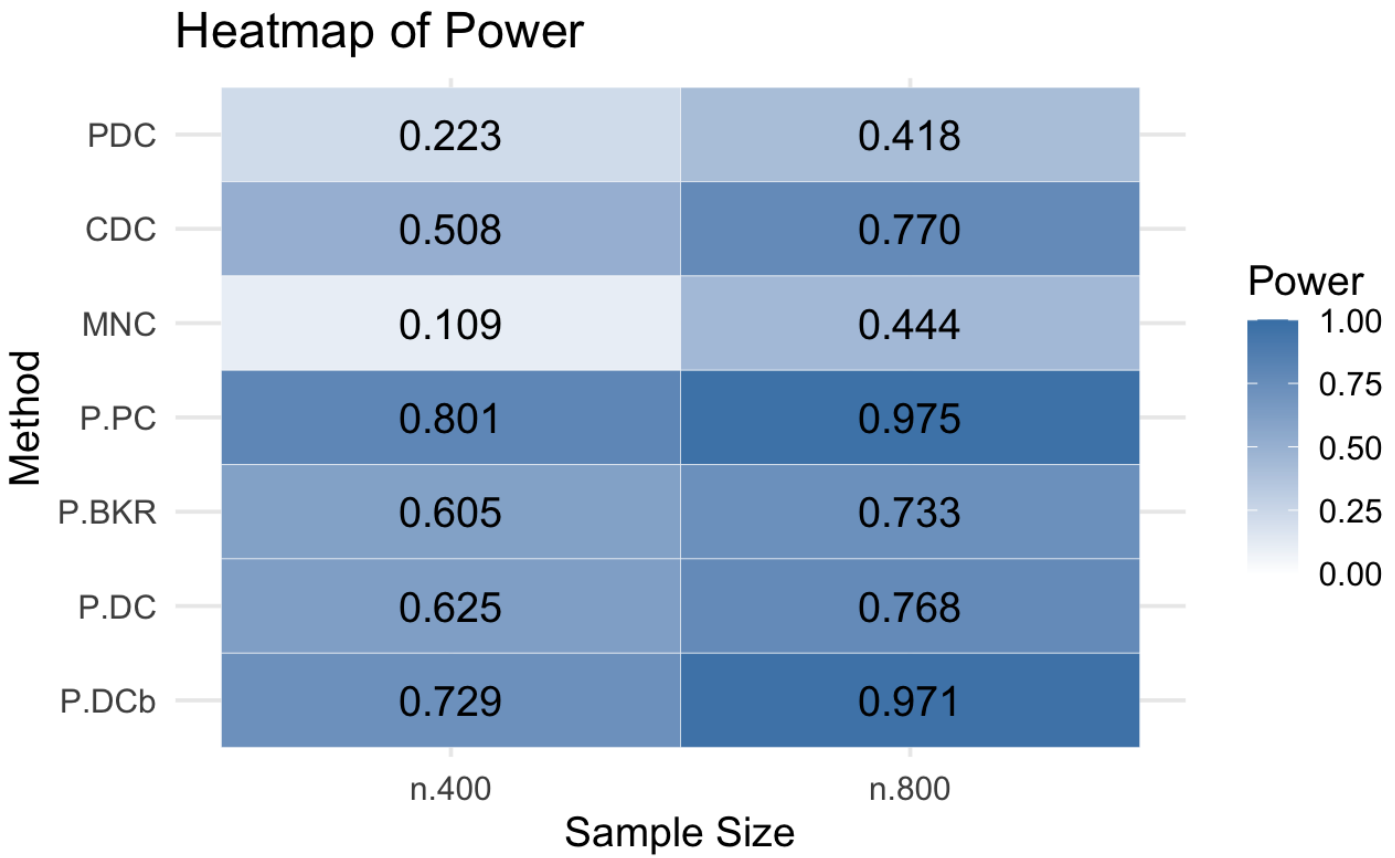

Figure: DGP2: Power of Tests. 'PDC': partial distance correlation; 'CDC': conditional distance correlation; 'MNC': maximal nonlinear correlation; 'P.PC': projection+projection correlation; 'P.BKR': projection+BKR; 'P.DC': projection+distance correlation; 'P.DCb': projection+bias-corrected distance correlation.

Application: Predictor Screening

- Let x_t denote the 100 Fama-French (FF) portfolios at time t, and Y_{t+1} denote the portfolio return at time $t+1$.
- $$\widehat{M}_n = \{j \in \{1, 2, \cdots, k\} : \left|R_n^*(\xi_{Y_t}, \eta_{X_{t-1}})\right| \geq c_n\},$$

  c_n is chosen among the top $[p/10]$ largest predictors of all.

- Benchmark: $\hat{Y}_{t+1} = \hat{a} + \hat{b}Y_t$.
- Working model: $\hat{Y}_{t+1} = \hat{a} + \hat{b}X_t + \hat{\beta}W_t$, $Z_t := \{Y_t, Y_{t-1}, x_t, x_{t-1}, \cdots, x_{t-4}\}$ with $Z_t \in \mathbb{R}^k, k = p + d = 502$, $q = 1$.
- $X_t := Z_{j,t} = \{Y_t, Y_{t-1}, x_{t-j}\} \in \mathbb{R}^{102}$, $W_t := Z_{-j,t} \in \mathbb{R}^{400}$.

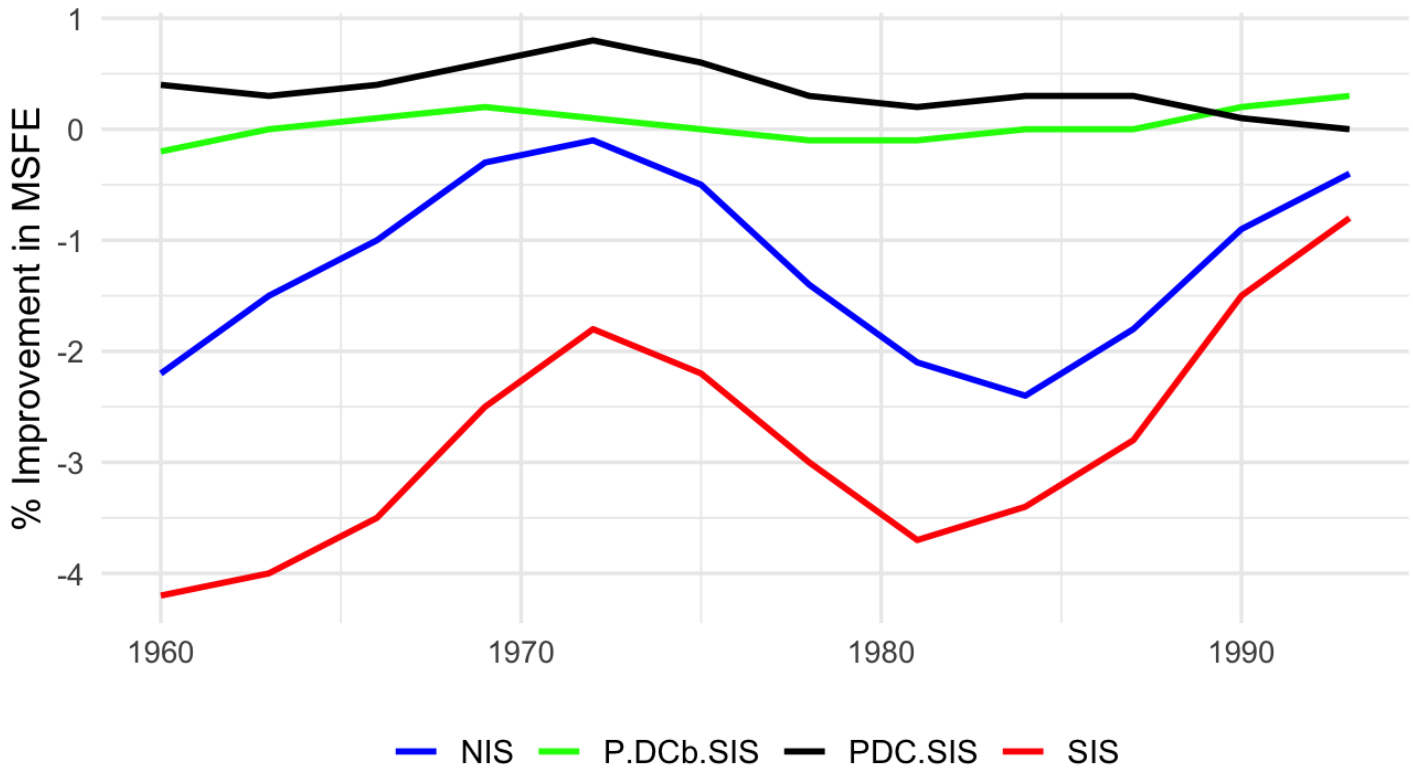

Figure: ΔMSFE (%) over the benchmark AR (1) model. 'SIS': sure independence screening; 'NIS': nonparametric independence screening; 'PDC-SIS': partial distance correlation + SIS; 'P.DCb-SIS': projection+bias-corrected distance correlation + SIS.

Conclusion

- We develop a new conditional independence test with high-dimensional time series using the projection-based bias-corrected distance correlation.
- The test has a standard limiting distribution, with the reasonable size control and satisfactory power.
- Limitation: The number of sample size cannot be too small, in order to preserve the sufficient amounts of blocks for training set.