Testing conditional independence under isotonicity

Rohan Hore[1], Jake A. Soloff[1], Rina Foygel Barber[1] and Richard J. Samworth[2]

[1] *Department of Statistics, University of Chicago*

[2] *Statistical Laboratory, University of Cambridge*

October 18, 2024

Abstract

We propose a powerful, nonparametric test of conditional independence $X \perp\!\!\!\perp Y \mid Z$ assuming only that X is stochastically increasing in Z. In particular, unlike recent work on conditional independence testing, our test does not require knowledge of the conditional distribution $X \mid Z$ beyond a shape constraint. Our method is a constrained form of permutation testing, affording the analyst a great deal of flexibility in designing a powerful test statistic.

1 Introduction

Consider the problem of testing the conditional independence (CI) hypothesis

$$H_0^{\mathrm{CI}} : X \perp\!\!\!\perp Y \mid Z,$$

where X and Y are variables of interest (such as a treatment X and an outcome Y), while Z represents a (potentially high-dimensional) confounder. Our available data consist of a sample $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n) \overset{\mathrm{iid}}{\sim} P$, where P is an unknown distribution on $(X, Y, Z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Throughout, we write $P_{X|Z}$ and $P_{Y|Z}$ to denote the conditional distributions of X given Z and Y given Z respectively.

In the case where the distribution of Z is nonatomic, i.e. $\mathbb{P}\{Z = z\} = 0$ for all $z \in \mathcal{Z}$, Shah and Peters (2020) established that, without further assumptions, there is no universally valid test of H_0^{CI} that achieves non-trivial power for *any* alternative distribution $P \notin H_0^{\mathrm{CI}}$; see also Neykov, Balakrishnan and Wasserman (2021) and Kim et al. (2022). Existing approaches to testing conditional independence have therefore sought to guarantee validity (Type I error control) over restricted classes of null distributions that impose one of the following additional structures:

(a) a parametric model, such as joint Gaussianity of (X, Y, Z), or a Gaussian linear model for $Y \mid (X, Z)$ (Kalisch and Bühlmann, 2007);

(b) a known or well-estimated conditional distribution $P_{X|Z}$ (Candès et al., 2018; Barber, Candès and Samworth, 2020; Berrett et al., 2020; Niu et al., 2024); or

(c) smoothness of the conditional distribution $P_{X|Z}$ (Shah and Peters, 2020; Lundborg, Shah and Peters, 2022; Kim et al., 2022; Lundborg et al., 2024+).

1.1 Our contributions

In this work, we introduce a new, nonparametric structure under which we can test conditional independence: we assume a shape constraint—specifically, a form of stochastic monotonicity—for the conditional distribution of $X \mid Z$. Such a constraint is motivated by several applications, particularly in biomedicine, where for instance incidence of diabetes becomes more prevalent with age (Yan et al., 2023), and left ventricular wall thickness is a known risk factor for hypertrophic cardiomyopathy (O'Mahony et al., 2014). Our specific constraint is as follows:

Assumption 1 (Monotonicity of the conditional distribution $P_{X|Z}$)**.** *Assume that $\mathcal{X} \subseteq \mathbb{R}$ and that $\preceq$ is a partial order on $\mathcal{Z}$. We assume X is stochastically increasing in Z, meaning that*

$$\text{if } z \preceq z' \text{ then } P\{X \geq x|z\} \leq P\{X \geq x|z'\} \text{ for all } x.$$

This assumption does not fall into any of the categories (a)–(c) above. We will often consider the case where the control variable Z is univariate, $\mathcal{Z} \subseteq \mathbb{R}$, under the usual total order $\leq$, though our framework also allows for multivariate $Z \in \mathbb{R}^d$. In this case, the most common partial order is the coordinatewise order.

Our main contribution is to introduce a broad strategy for testing the isotonic conditional independence (ICI) null hypothesis

$$H_0^{\text{ICI}} : X \perp\!\!\!\perp Y \mid Z, \text{ and } P_{X|Z} \text{ satisfies Assumption 1.} \tag{1}$$

Naturally, this test should only be applied in settings where the monotonicity condition of Assumption 1 is well-motivated, so that a rejection of H_0^{ICI} can reasonably be interpreted as evidence that $X \not\perp\!\!\!\perp Y \mid Z$. However, we emphasize again that some additional assumption beyond H_0^{CI} is essential for any valid test to have non-trivial power at any alternative.

1.2 Background: testing independence

To set the stage for some of the notation and ideas underlying our methodology, we briefly review the simpler framework of permutation testing of the null hypothesis of marginal independence, $X \perp\!\!\!\perp Y$.

Given a joint distribution P on $\mathcal{X} \times \mathcal{Y}$, let $(X_i, Y_i)_{i\in[n]} \overset{\text{iid}}{\sim} P$, and write $\mathbf{X} = (X_i)_{i=1}^n$ and $\mathbf{Y} = (Y_i)_{i=1}^n$. We can reframe the problem of testing marginal independence as testing whether the entries of $\mathbf{X}$ are i.i.d. given $\mathbf{Y}$. Specifically, and according to de Finetti's theorem (de Finetti, 1929), permutation tests look for violations of exchangeability of $\mathbf{X}$ given $\mathbf{Y}$. The general approach proceeds as follows: based on $\mathbf{Y}$, the analyst chooses any statistic $T : \mathcal{X}^n \to \mathbb{R}$, with larger values of $T(\mathbf{X})$ indicating evidence against the independence null[1]. Writing $\mathcal{S}_n$ for the set of permutations of $[n]$, and for $\sigma \in \mathcal{S}_n$, let $T_\sigma = T(\mathbf{X}^\sigma)$ denote

[1] We emphasize that $T(\mathbf{X})$ is allowed to depend on both $\mathbf{X}$ and $\mathbf{Y}$ —for instance we may define the function as $T(\mathbf{x}) = |\sum_{i=1}^n x_i Y_i|$, though we suppress the dependence on $\mathbf{Y}$ in our notation.

the value of the statistic when the entries of $\mathbf{X}$ are permuted according to σ—that is, $\mathbf{X}^\sigma = (X_{\sigma(1)}, \ldots, X_{\sigma(n)})$. Finally, define a p-value

$$p = \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \mathbb{1}\left\{T_\sigma \geq T\right\}.$$

This construction produces a valid p-value for *any* choice of test statistic T, and the statistic T can be tailored to have power against certain specific alternatives. Indeed, this strategy has been successfully employed to construct independence tests via nearest neighbor distances and mutual information (Berrett and Samworth, 2019), moment methods (Kim, Balakrishnan and Wasserman, 2022), kernels (Pfister et al., 2018), the Hilbert–Schmidt independence criterion (Albert et al., 2022) and U-statistics (Berrett and Samworth, 2021; Berrett, Kontoyiannis and Samworth, 2021),

In our work, since we are interested in conditional (rather than marginal) independence, we will follow a similar strategy, but will use a restricted class of functions T and a (data-dependent) subgroup of permutations $\sigma \in \mathcal{S}_n$, both of which respect the stochastic monotonicity Assumption 1. Our framework still affords the analyst a great deal of flexibility in designing their test, while controlling Type I error across the more challenging null class H_0^{ICI}.

2 Methodology

In this section we give a general procedure, called the `PairSwap-ICI` test, for testing the isotonic conditional independence null H_0^{ICI}. Intuitively, it is plausible that we should be able to construct powerful tests against some alternatives. For example, if $Z_i \preceq Z_j$, then the shape constraint ensures that $X_i \leq X_j$ should hold *at least* half of the time; if we instead observe $X_i \gg X_j$, then this may be due to the influence of Y. Our test builds on and formalizes this intuition, allowing the analyst to specify, based on $\mathbf{Y}$ and $\mathbf{Z}$, pairs (i, j) such that $Z_i \preceq Z_j$ and then to use, for example, large differences $X_i - X_j$ as evidence against the null.

More formally, based on $\mathbf{Y}$ and $\mathbf{Z}$, and without access to $\mathbf{X}$, the analyst chooses:

(i) A sequence of ordered pairs

$$(i_1, j_1), \ldots, (i_L, j_L)$$

of indices in $[n] = \{1, \ldots, n\}$, where all $2L$ entries are distinct. We require the pairs to be ordered in the sense that

$$Z_{i_\ell} \preceq Z_{j_\ell} \tag{2}$$

for each $\ell \in [L] = \{1, \ldots, L\}$.

(ii) A sequence of functions $\psi_1, \ldots, \psi_L$, where each $\psi_\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfies the *anti-monotonicity* property

$$\psi_\ell(x + \Delta, x' - \Delta') - \psi_\ell(x' - \Delta', x + \Delta) \geq \psi_\ell(x, x') - \psi_\ell(x', x), \tag{3}$$

for all $\Delta, \Delta' \geq 0$.

With these choices in place, our test statistic $T : \mathcal{X}^n \to \mathbb{R}$ is defined as[2]

$$T(\mathbf{x}) = \sum_{\ell=1}^{L} \psi_\ell(x_{i_\ell}, x_{j_\ell}). \tag{4}$$

In order to calibrate the test, the analyst compares the observed test statistic $T = T(\mathbf{X})$ with versions of T where indices within pairs (i_ℓ, j_ℓ) are randomly swapped. Specifically, for $\mathbf{s} \in \{\pm 1\}^L$, define $T_{\mathbf{s}} = T(\mathbf{X}^{\mathbf{s}})$, where $\mathbf{X}^{\mathbf{s}}$ is a swapped version of the data vector $\mathbf{X}$, with entries

$$\begin{cases} (X^{\mathbf{s}}_{i_\ell}, X^{\mathbf{s}}_{j_\ell}) = (X_{i_\ell}, X_{j_\ell}) & s_\ell = +1, \\ (X^{\mathbf{s}}_{i_\ell}, X^{\mathbf{s}}_{j_\ell}) = (X_{j_\ell}, X_{i_\ell}) & s_\ell = -1, \end{cases}$$

so that $s_\ell = -1$ indicates that the random variables X_{i_ℓ} and X_{j_ℓ} are swapped, while $s_\ell = +1$ indicates no swap. Informally, the two constraints (2) and (3) ensure that, under the null, each $\psi_\ell(X_{i_\ell}, X_{j_\ell})$ is likely to be no larger than its swapped version, $\psi_\ell(X_{j_\ell}, X_{i_\ell})$—and thus, the statistic $T = T(\mathbf{X})$ is likely to be no larger than its swapped copies, $T_{\mathbf{s}}$. If instead $T > T_{\mathbf{s}}$ for many swaps $\mathbf{s}$, this indicates evidence against the null. To formalize this intuition, we define p-value for `PairSwap-ICI` test as

$$p := \frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1}\left\{T_{\mathbf{s}} \geq T\right\}. \tag{5}$$

In words, we are comparing the observed value T of the statistic, against all possible permuted statistic values $T_{\mathbf{s}}$ that we would obtain by swapping indices within matched pairs of our observed data $\mathbf{X}$.

Our method is quite flexible, in that the analyst may select any pairs (i_ℓ, j_ℓ) subject to monotonicity (2) and functions (ψ_ℓ) satisfying (3) to construct a valid test. In particular, they can decide on these aspects of the test *after* exploring the data $\mathbf{Y}, \mathbf{Z}$, choosing to include a pair (i_ℓ, j_ℓ) if the data observed in $\mathbf{Y}$ indicates that $X_{i_\ell} > X_{j_\ell}$ would be likely under the alternative. However, the quality of the matches (i_ℓ, j_ℓ) and functions (ψ_ℓ) affects the power of our test; we discuss effective strategies for designing a statistic T in Section 3.

Example: a linear test statistic. Before proceeding, we give a simple example of a test statistic that we might choose to use: consider a linear test statistic,

$$T(\mathbf{x}) = \sum_{i=1}^{n} \beta_i x_i$$

for some coefficients $\beta_i \in \mathbb{R}$. This function can be used as the test statistic for the `PairSwap-ICI` test, as long as the coefficients satisfy

$$\beta_{i_\ell} \geq \beta_{j_\ell} \text{ for each } \ell \in [L]. \tag{6}$$

[2] Again, as for marginal permutation tests in Section 1.2, here we suppress dependence on $\mathbf{Y}$ and $\mathbf{Z}$ in the notation $T(\mathbf{x})$, even though this statistic does depend on $\mathbf{Y}$ and $\mathbf{Z}$ through the choices of the matched pairs (i_ℓ, j_ℓ) and functions ψ_ℓ.

To see why, first note that without loss of generality we can take $\beta_i = 0$ for all $i \in [n] \setminus \{i_1, j_1, \ldots, i_L, j_L\}$, i.e., all data points not belonging to any of the L pairs. This is because the indicator $\mathbb{1}\{T_{\mathbf{s}} \geq T\}$, appearing in the computation of the p-value, is invariant to these terms. Next, define

$$\psi_\ell(x, x') = \beta_{i_\ell} x + \beta_{j_\ell} x',$$

which satisfies (3) because for any $\Delta, \Delta' \geq 0$ with $x, x', x + \Delta, x' - \Delta' \in \mathcal{X}$, we have

$$\psi_\ell(x + \Delta, x' - \Delta') - \psi_\ell(x' - \Delta', x + \Delta) - \psi_\ell(x, x') + \psi_\ell(x', x) = (\beta_{i_\ell} - \beta_{j_\ell})(\Delta + \Delta') \geq 0$$

by our assumption that $\beta_{i_\ell} \geq \beta_{j_\ell}$. We then have $T(\mathbf{x})$ equal to the test statistic defined in (4).

We remark that choosing such a test statistic is by no means implying an assumption that the dependence between X and Y follows a linear model —it may be the case that $T(\mathbf{x}) = \beta^\top \mathbf{x}$ has good power for distinguishing the null from the alternative even if a linear model is only an approximation to the true model.

2.1 Validity

Our first main result is that our method yields a valid test of H_0^{ICI}.

Theorem 1. *Under H_0^{ICI}, the conditional Type I error of the* `PairSwap-ICI` *test satisfies $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \leq \alpha$ for all $\alpha \in [0, 1]$. In particular, the test enjoys marginal error control: $\mathbb{P}\{p \leq \alpha\} \leq \alpha$ for all α.*

Our proof of Theorem 1 formalizes our intuition at the start of this section, making use of the fact that, under the null, $\psi_\ell(X_{i_\ell}, X_{j_\ell})$ tends to be smaller than its swapped version $\psi_\ell(X_{j_\ell}, X_{i_\ell})$.

Proof of Theorem 1. Our proof is split into three steps. First we derive some deterministic properties of the p-value p. Next, we compare to the *sharp null*, where the pair X_{i_ℓ}, X_{j_ℓ} are identically distributed (rather than stochastically ordered). Finally, we examine the validity of the test under the sharp null.

Step 1: some deterministic properties of the p-value. First, fix $\alpha \in [0, 1]$ and define a function $p : \mathbb{R}^n \to [0, 1]$ as

$$p(\mathbf{x}) = \frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1}\{T(\mathbf{x}^{\mathbf{s}}) \geq T(\mathbf{x})\},$$

so that the p-value p in (5) can be written as $p = p(\mathbf{X})$. Now find a bijection $\sigma : [2^L] \to \{-1, 1\}^L$ such that $T(\mathbf{x}^{\sigma(1)}) \geq \ldots \geq T(\mathbf{x}^{\sigma(2^L)})$, and let $r \in \{0, 1, \ldots, 2^L\}$ be such that $\alpha \in \left[r/2^L, (r+1)/2^L\right)$. Then, since $\sum_{k=1}^{2^L} \mathbb{1}\left\{T(\mathbf{x}^{\sigma(k)}) \geq T(\mathbf{x}^{\sigma(j)})\right\} \in [2^L]$ for each $j \in [2^L]$, we have deterministically that

$$\begin{aligned}\frac{1}{2^L}\sum_{\mathbf{s}\in\{\pm1\}^L}\mathbb{1}\left\{p(\mathbf{x}^{\mathbf{s}})\leq\alpha\right\} &= \frac{1}{2^L}\sum_{j=1}^{2^L}\mathbb{1}\left\{\frac{1}{2^L}\sum_{k=1}^{2^L}\mathbb{1}\left\{T(\mathbf{x}^{\sigma(k)})\geq T(\mathbf{x}^{\sigma(j)})\right\}\leq\alpha\right\}\\ &= \frac{1}{2^L}\sum_{j=1}^{2^L}\mathbb{1}\left\{\sum_{k=1}^{2^L}\mathbb{1}\left\{T(\mathbf{x}^{\sigma(k)})\geq T(\mathbf{x}^{\sigma(j)})\right\}\leq r\right\}\\ &\leq \frac{1}{2^L}\sum_{j=1}^{2^L}\mathbb{1}\left\{\sum_{k=1}^{2^L}\mathbb{1}\left\{k\leq j\right\}\leq r\right\} = \frac{1}{2^L}\sum_{j=1}^{2^L}\mathbb{1}\left\{j\leq r\right\} = \frac{r}{2^L}\leq\alpha.\end{aligned} \tag{7}$$

In addition, we claim that $p(\mathbf{x})$ is monotone in its coordinates, namely, $p(\mathbf{x})$ is monotone nonincreasing in each x_{i_ℓ}, and monotone nondecreasing in each x_{j_ℓ}. To see why, for each $\mathbf{s}\in\{\pm1\}^L$ we can calculate

$$\begin{aligned}\mathbb{1}\left\{T(\mathbf{x}^{\mathbf{s}})\geq T(\mathbf{x})\right\} &= \mathbb{1}\left\{\sum_{\ell:s_\ell=+1}\psi_\ell(x_{i_\ell},x_{j_\ell})+\sum_{\ell:s_\ell=-1}\psi_\ell(x_{j_\ell},x_{i_\ell})\geq\sum_{\ell=1}^{L}\psi_\ell(x_{i_\ell},x_{j_\ell})\right\}\\ &= \mathbb{1}\left\{\sum_{\ell:s_\ell=-1}\Big(\psi_\ell(x_{i_\ell},x_{j_\ell})-\psi_\ell(x_{j_\ell},x_{i_\ell})\Big)\leq 0\right\}.\end{aligned}$$

By the anti-monotonicity condition (3) on ψ_ℓ, this function is monotone nonincreasing in each x_{i_ℓ}, and monotone nondecreasing in each x_{j_ℓ}, and therefore the same is true for $p(\mathbf{x})$ as well.

Step 2: compare to the sharp null. For each $i\in[n]$, let $P_i=P_{X|Z}(\cdot\mid Z_i)$ denote the distribution of X_i (after conditioning on $\mathbf{Y},\mathbf{Z}$). By Assumption 1, we know that $P_{i_\ell}\preceq_{\mathrm{st}}P_{j_\ell}$ for each pair $\ell\in[L]$, where $\preceq_{\mathrm{st}}$ denotes the stochastic ordering on distributions. Next, we also define distributions $\bar{P}_\ell$ for each $\ell\in[L]$, given by the mixture

$$\bar{P}_\ell=\frac{1}{2}P_{i_\ell}+\frac{1}{2}P_{j_\ell}.$$

In particular, this implies that

$$P_{i_\ell}\preceq_{\mathrm{st}}\bar{P}_\ell\preceq_{\mathrm{st}}P_{j_\ell},\ \ \ell\in[L]. \tag{8}$$

We will now compare the observed data values, whose distribution (conditional on $\mathbf{Y},\mathbf{Z}$) is given by

$$\mathbf{X}=(X_1,\dots,X_n)\ \sim\ P_1\times\cdots\times P_n,$$

against a different distribution,

$$\mathbf{X}_\sharp=((X_\sharp)_1,\dots,(X_\sharp)_n)\ \sim\ (P_\sharp)_1\times\cdots\times(P_\sharp)_n,$$

where the distributions $(P_\sharp)_i$ are defined by setting

$$(P_\sharp)_{i_\ell}=(P_\sharp)_{j_\ell}=\bar{P}_\ell$$

for each $\ell \in [L]$ (and, for any index $i \in [n] \backslash \{i_1, j_1, \dots, i_L, j_L\}$ that does not belong to any of the L matched pairs, we simply take $(P_\sharp)_i = P_i$). We can think of this alternative vector of observations as being drawn from a *sharp null*, because for each pair ℓ, the random variables $(X_\sharp)_{i_\ell}, (X_\sharp)_{j_\ell}$ are identically distributed (rather than stochastically ordered, as for X_{i_ℓ}, X_{j_ℓ}). In particular, this implies that

$$(\mathbf{X}_\sharp)^{\mathbf{s}} \overset{\mathrm{d}}{=} \mathbf{X}_\sharp \tag{9}$$

for any $\mathbf{s} \in \{\pm 1\}^L$ (after conditioning on $\mathbf{Y}, \mathbf{Z}$).

In Step 1, we verified that the function $p(\mathbf{x})$ is monotone nonincreasing in each x_{i_ℓ}, and monotone nondecreasing in each x_{j_ℓ}. In particular, combined with the stochastic ordering (8), this means that $p(\mathbf{X}_\sharp) \preceq_{\mathrm{st}} p(\mathbf{X})$ (conditional on $\mathbf{Y}, \mathbf{Z}$). We therefore have

$$\mathbb{P}\left\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} = \mathbb{P}\left\{p(\mathbf{X}) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} \leq \mathbb{P}\left\{p(\mathbf{X}_\sharp) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\}.$$

From this point on, then, we only need to verify validity of the p-value $p(\mathbf{X}_\sharp)$ computed under the sharp null.

Step 3: validity under the sharp null. For data $\mathbf{X}_\sharp$ drawn under a sharp null, we have

$$\begin{aligned}\mathbb{P}\left\{p(\mathbf{X}_\sharp) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} &= \frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{P}\left\{p((\mathbf{X}_\sharp)^{\mathbf{s}}) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} \\ &= \mathbb{E}\left[\frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1}\left\{p((\mathbf{X}_\sharp)^{\mathbf{s}}) \leq \alpha\right\} \,\middle|\, \mathbf{Y}, \mathbf{Z}\right] \leq \alpha,\end{aligned}$$

where the first step holds by (9), while the last step holds by the deterministic calculation (7) from Step 1. □

The p-value constructed in (5) requires computing $T_{\mathbf{s}} = T(X^{\mathbf{s}})$ for all 2^L values of $\mathbf{s} \in \{-1, 1\}^L$, which may be computationally prohibitive for moderate or large L. In practice, it is common to use a Monte Carlo approximation to the p-value: we sample $\mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)} \overset{\text{iid}}{\sim} \mathrm{Unif}(\{\pm 1\}^L)$, and then compute

$$\hat{p}_M = \frac{1 + \sum_{m=1}^M \mathbb{1}\left\{T_{\mathbf{s}^{(m)}} \geq T\right\}}{1 + M}.$$

The extra '1+' term appearing in the numerator and denominator is necessary to ensure error control for this Monte Carlo version of our test (Davison and Hinkley, 1997; Phipson and Smyth, 2010);in particular, this correction ensures we cannot have $\hat{p}_M = 0$. The following theorem verifies that this version of the test also controls the Type I error.

Theorem 2. *Fix any $M \geq 1$. Under H_0^{ICI}, it holds that $\mathbb{P}\{\hat{p}_M \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \leq \alpha$ for all $\alpha \in [0, 1]$, and consequently, $\mathbb{P}\{\hat{p}_M \leq \alpha\} \leq \alpha$.*

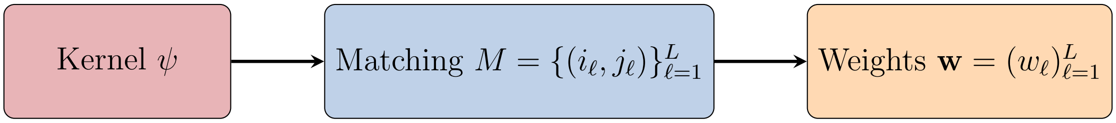
3 Designing a powerful `PairSwap-ICI` test

In this section, we construct a principled, powerful implementation of our test, by designing concrete choices for the pairs (i_ℓ, j_ℓ) and the functions ψ_ℓ introduced in Section 2. Throughout, we will restrict our attention to test statistics $T(\mathbf{x})$ of the form

$$T(\mathbf{x}) = \sum_{\ell=1}^{L} w_\ell \, \psi(x_{i_\ell}, x_{j_\ell}). \tag{10}$$

That is, in the original definition of the test statistic (4), we take $\psi_\ell(\cdot) = w_\ell \psi(\cdot)$ for some sequence of non-negative weights $\mathbf{w} = (w_\ell)_{\ell=1}^L$ and some *fixed* kernel ψ (which is required to satisfy the anti-monotonicity condition (3)).

With this simplification, designing a test statistic now requires specifying the kernel ψ, deciding which pairs (i_ℓ, j_ℓ) get matched, and finally, how much weight w_ℓ to assign to each pair, as depicted in this flowchart:


Kernel ψ
Matching $M = \{(i_\ell, j_\ell)\}_{\ell=1}^L$
Weights $\mathbf{w} = (w_\ell)_{\ell=1}^L$

As guaranteed by Theorem 1, our test controls the Type I error for any choice of ψ, M and $\mathbf{w}$, subject to the conditions (2) and (3) outlined at the start of Section 2. However, for the test to be effective, we need to tailor these choices to the specific application of interest. Of course, all of these choices interact with each other: what constitutes a good matching depends on how we choose the weights, and vice versa.

3.1 Specifying the kernel ψ

We begin by considering several simple options for the kernel ψ. As a first example, consider $\psi(x, x') = x - x'$. This choice of ψ means that $\psi(X_{i_\ell}, X_{j_\ell})$ is likely to be ≤ 0 under the null (since $Z_{i_\ell} \preceq Z_{j_\ell}$), but under the alternative, may be likely to be large (if the pair (i_ℓ, j_ℓ) is chosen wisely). Of course, we also allow for nonlinear test statistics to handle a broader range of settings. If X has heavy tails, then the distribution of a linear statistic T can be very sensitive to extreme values. We can ameliorate this sensitivity by using $\psi(x, x') = \text{sign}(x - x')$, or $\psi(x, x') = (-K) \vee (x - x') \wedge K$ (the truncation of $x - x'$ to some bounded interval $[-K, K]$) for some appropriate $K > 0$.

Example: a linear test statistic, revisited. To give more motivation for these simple choices, we will now see that a `PairSwap-ICI` test run with any *linear* test statistic $T_{\text{lin}}(\mathbf{x}) = \sum_{i=1}^n \beta_i x_i$, can always be expressed in the form (10) with the linear kernel $\psi(x, x') = x - x'$. To see why, define

$$w_\ell = \beta_{i_\ell} - \beta_{j_\ell},$$

(and note that we must have $w_\ell \geq 0$ due to (6), i.e., this is a valid choice of weight). Then we can write

$$T_{\text{lin}}(\mathbf{x}) = T_\psi(\mathbf{x}) + T_{\text{sym}}(\mathbf{x}),$$

where T is defined as in (10), and where the term

$$T_{\text{sym}}(\mathbf{x}) = \sum_{\ell=1}^{L} \frac{\beta_{i_\ell} + \beta_{j_\ell}}{2}(x_{i_\ell} + x_{j_\ell}) + \sum_{i \in [n] \setminus \{i_1, j_1, \ldots, i_L, j_L\}} \beta_i x_i$$

is symmetric in the pair (x_{i_ℓ}, x_{j_ℓ}) for each ℓ—and thus, $T_{\text{sym}}(\mathbf{x}^{\mathbf{s}}) = T_{\text{sym}}(\mathbf{x})$ for any $\mathbf{x}$ and any $\mathbf{s} \in \{\pm 1\}^L$. Therefore,

$$\mathbb{1}\left\{T_{\mathbf{s}} \geq T\right\} = \mathbb{1}\left\{(T_\psi)_{\mathbf{s}} \geq T_\psi\right\}$$

for every $\mathbf{s}$—that is, the p-value p defined in (5) is *identical* if we use the test statistic T of the form (10) in place of the original linear test statistic T_{lin}.

3.2 Oracle strategies for choosing the matching and weights

In this section, we build intuition for how to choose the matching M and weights $\mathbf{w}$ effectively by sketching the asymptotics of our test, assuming some oracle knowledge (or estimates) of the data distribution. Let us consider any statistic T of the form (10). Throughout this section, the kernel ψ is a fixed function, and we wish to choose the weights $\mathbf{w} = (w_\ell)_{\ell \in [L]}$ and matching $M = \{(i_\ell, j_\ell)\}_{\ell \in [L]}$ to maximize the power of our test. Given the data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, the reference statistic $T_{\mathbf{s}}$ is a sum of L independent random variables. Under some regularity conditions on the weights $\mathbf{w}$ and the function ψ, a CLT approximation gives, for large L, that

$$p \approx \bar{\Phi}(\hat{T}) \qquad \text{where} \qquad \hat{T} = \hat{T}(\mathbf{w}, M) := \frac{\sum_{\ell=1}^L w_\ell \psi(X_{i_\ell}, X_{j_\ell})}{\sqrt{\sum_{\ell=1}^L w_\ell^2 \psi(X_{i_\ell}, X_{j_\ell})^2}},$$

and where $\bar{\Phi}$ denotes the standard Gaussian survival function, i.e., $\bar{\Phi}(t) = 1 - \Phi(t)$, where Φ is the standard Gaussian cumulative distribution function. The above approximation holds for fixed $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ and relies only on the CLT approximation for a weighted sum of L independent signs $s_1, \ldots, s_L \in \{\pm 1\}$.

In particular, the above calculation tells us that we should aim to choose weights that *maximize* the approximate probability of rejection, $\mathbb{P}\left\{\bar{\Phi}(\hat{T}) \leq \alpha\right\}$, in order to achieve the best possible power. As we will see more formally in Theorem B.9, under some conditions this power can be further approximated as

$$\approx \Phi\left(\frac{\sum_\ell w_\ell \mathbb{E}\left[\psi_\ell(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right]}{\sqrt{\sum_\ell w_\ell^2 \text{Var}\left(\psi_\ell(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right)}} - \bar{\Phi}^{-1}(\alpha)\right). \tag{11}$$

Treating the matching M as fixed, maximizing this approximation leads to the choice of weights

$$w_\ell^* = \frac{\max\left\{\mathbb{E}\left[\psi_\ell(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right], 0\right\}}{\text{Var}\left(\psi_\ell(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right)}. \tag{12}$$

Using a plugin estimate for the moments. In practice, of course, computing $T^*(\mathbf{w}, M)$ assumes that we are able to compute the conditional expected values, $\mathbb{E}\left[\psi(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right]$ and $\mathbb{E}\left[\psi(X_{i_\ell}, X_{j_\ell})^2 \mid \mathbf{Y}, \mathbf{Z}\right]$. In practice, we will instead replace these with plugin estimates: let

$$\hat{E}_{ij} \approx \mathbb{E}\left[\psi(X_i, X_j) \mid \mathbf{Y}, \mathbf{Z}\right], \quad \hat{V}_{ij} \approx \text{Var}\left(\psi(X_i, X_j) \mid \mathbf{Y}, \mathbf{Z}\right)$$

denote estimates, which are required to be independent of $\mathbf{X}$. For example, we might compute an estimated model for the conditional distribution of $X \mid Y, Z$ using a separate data set, and then use this estimated model to produce the estimates $\hat{E}_{ij}, \hat{V}_{ij}$. To make this concrete, for linear statistics $\psi(x, x') = x - x'$, producing these estimates $\hat{E}_{ij}, \hat{V}_{ij}$ requires us to estimate the first two moments of $X_i \mid Y_i, Z_i$ for each i. Alternatively, if $\psi(x, x') = \mathrm{sign}(x - x')$, then we would need to estimate $\mathbb{P}\{X_i > X_j \mid \mathbf{Y}, \mathbf{Z}\}$.

With these estimates in place, the power is then approximately maximized by choosing weights

$$\hat{w}^*_\ell = \frac{\max\left\{\hat{E}_{i_\ell j_\ell}, 0\right\}}{\hat{V}_{i_\ell j_\ell}}.$$

Choosing the matching. In the oracle setting, once we fix the choice of weights to (12), the estimator (11) of the test's power is maximized by solving the following maximum-weight matching problem:

$$M^* \in \underset{\text{Matchings } M \text{ on } [n]}{\mathrm{argmax}} \sum_{(i,j)\in M} W^*_{ij}\, \mathbf{1}\{Z_i \preceq Z_j\} \text{ where } W^*_{ij} = \frac{\max\left\{\mathbb{E}\left[\psi_\ell(X_i, X_j) \mid \mathbf{Y}, \mathbf{Z}\right], 0\right\}}{\mathrm{Var}\left(\psi_\ell(X_i, X_j) \mid \mathbf{Y}, \mathbf{Z}\right)}, \tag{13}$$

where the maximum is taken over all valid matchings—that is, all collections of disjoint pairs $(i_1, j_1), \ldots, (i_L, j_L)$ from the indices $[n]$ (note that we may choose any $L \leq n/2$). We then run `PairSwap-ICI` with this oracle matching M^*, and with weights $w_\ell = W^*_{i_\ell j_\ell}$ for each pair (i_ℓ, j_ℓ) in the oracle matching M^*.

Similarly, using the plugin estimates, the power is approixmately maximized by solving the matching problem

$$\hat{M}^* \in \underset{\text{Matchings } M \text{ on } [n]}{\mathrm{argmax}} \sum_{(i,j)\in M} \hat{W}^*_{ij}\, \mathbf{1}\{Z_i \preceq Z_j\} \text{ where } \hat{W}^*_{ij} = \frac{\max\left\{\hat{E}_{ij}, 0\right\}}{\hat{V}_{ij}}. \tag{14}$$

To run `PairSwap-ICI`, we then take weights $w_\ell = \hat{W}^*_{i_\ell j_\ell}$ for each pair (i_ℓ, j_ℓ) in the plugin oracle matching $\hat{M}^*$.

The plugin matching $\hat{M}^*$ can be computed in polynomial time (Edmonds, 1965; Duan and Pettie, 2014). Specifically, if $m = \#\{(i,j) : Z_i \preceq Z_j\}$, can be computed in time $O(mn + n^2 \log n)$ using an algorithm of Gabow (1985). For our experiments in Sections 5 and 6, we use the Python package `networkx` (Hagberg, Swart and S Chult, 2008), which uses the Blossom algorithm (Edmonds, 1965) and runs in time $O(n^3)$.

3.3 Heuristic strategies for choosing the matching and weights

We now discuss simple alternative approaches for the one-dimensional setting ($\mathcal{Z} \subseteq \mathbb{R}$) that do not require accurate estimation of the first two moments of $\psi(X_i, X_j)$. Our heuristic methods apply to the linear case $\psi(x, x') = x - x'$. As we will see in our experiments, these heuristic approaches may be preferable if we do not have a quality working model of the conditional distribution $X \mid Y, Z$ or if maximum-weight matching is computationally prohibitive.

3.3.1 A simple weighting scheme

To motivate this approach, suppose that matched pairs (i_ℓ, j_ℓ) are constrained such that $Z_{i_\ell} \approx Z_{j_\ell}$ for all ℓ. Furthermore, for simplicity, we assume that the conditional mean function $\mu^*(y,z) := \mathbb{E}[X|Y=y, Z=z]$ is *linear* in its first argument

$$\mu^*(y,z) = \beta^* y + \mu_Z^*(z), \tag{15}$$

and that the conditional variance σ^{*2} is constant. In this case, the ideal weights w_ℓ^* in (12) are approximately $\frac{\beta^*}{\sigma^{*2}}(Y_{i_\ell} - Y_{j_\ell})$. Crucially, our test is invariant to rescaling the weights—that is, we do not need to know $\frac{\beta^*}{\sigma^{*2}}$—so we instead set

$$\hat{w}_\ell := Y_{i_\ell} - Y_{j_\ell}.$$

In this case, the constraints (2) and (3) mean that the pair of observations (Y_{i_ℓ}, Z_{i_ℓ}) and (Y_{j_ℓ}, Z_{j_ℓ}) must be discordant.

3.3.2 Neighbor matching

Recall that we require (1) $Z_{i_\ell} \approx Z_{j_\ell}$, (2) $Z_{i_\ell} \leq Z_{j_\ell}$ and (3) $Y_{i_\ell} \geq Y_{j_\ell}$. A naïve matching strategy, then, is to sort (Y_i, Z_i) according to the Z-values in ascending order $Z_{(1)} \leq \cdots \leq Z_{(n)}$, matching $(2j, 2j+1)$ if $Y_{(2j)} \geq Y_{(2j+1)}$ for $j \in \{1, \ldots, \lfloor n/2 \rfloor\}$, and otherwise leaving $2j$ and $2j+1$ unmatched.

An obvious limitation of this naïve matching strategy is that many pairs can fail to have $Y_{(2j)} \geq Y_{(2j+1)}$ just by chance. For instance, even if $Z_{(2j)}$ and $Z_{(2j+1)}$ are extremely close, we might expect $Y_{(2j)}$ and $Y_{(2j+1)}$ to be roughly iid (given $\mathbf{Z}$), so $\text{sign}(Y_{(2j)} - Y_{(2j+1)})$ is roughly a Rademacher variable. In particular, we are throwing out many observations that might be matched with other nearby observations.

3.3.3 Cross-bin matching

In order to increase to total number of matches L, we propose *cross-bin matching*. Unlike neighbor matching, where immediate neighbors in Z are matched, we bin the observations according to Z and match observations in adjacent bins. Figure 1 illustrates the approach.

Specifically, we sort the observations by Z values in increasing order and partition them into K bins such that approximately n/K observations fall in each bin. Formally, $A_1, \ldots, A_K$ form a partition of the $[n]$, where

$$A_1 = \{1, \ldots, \lfloor n/K \rfloor\}, \ldots, A_k = \{(k-1)\lfloor n/K \rfloor, \ldots, k\lfloor n/K \rfloor\}, \ldots$$
$$, A_K = \{(K-1)\lfloor n/K \rfloor, \ldots, n\}.$$

Now given the binning, for any $k \in [K]$, we define the set of positive and negative samples as

$$J_k^+ = \{i \in A_k,\ Y_i \geq \text{Median}(\{Y_i : i \in A_k\})\},$$
$$J_k^- = \{i \in A_k,\ Y_i < \text{Median}(\{Y_i : i \in A_k\})\}$$

respectively. Now, we order the observations in both the positive and negative set by increasing and decreasing order in Y. Formally, $i_{k,1}, i_{k,2}, \ldots$ are indices in J_k^+ such that

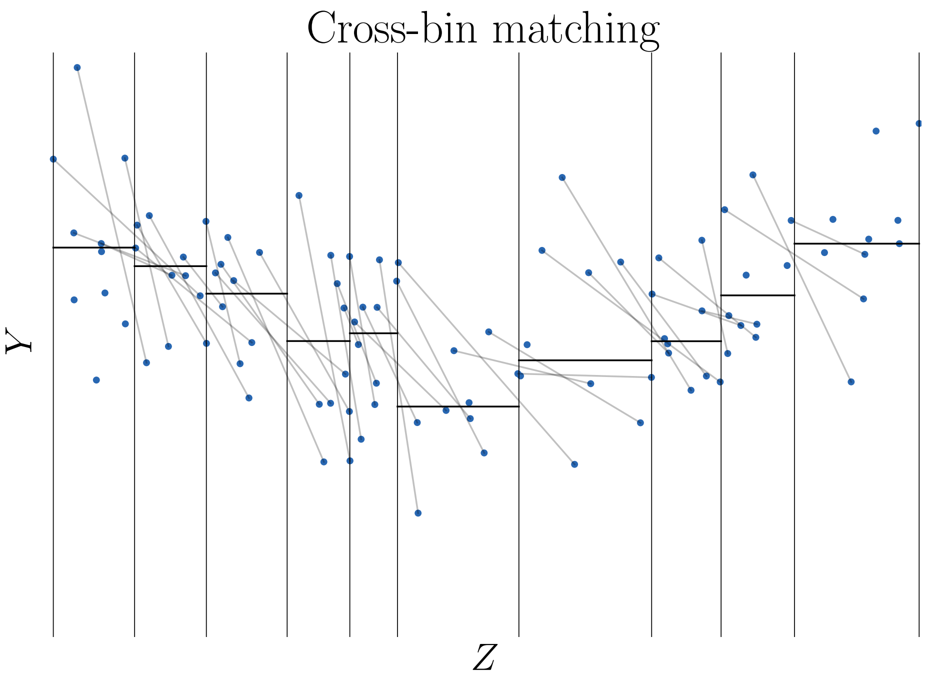


Figure 1: Demonstration of the cross-bin matching scheme described in Section 3.3.3.

$Y_{i_{k,1}} \geq Y_{i_{k,2}} \geq \dots$ and similarly, $j_{k,1}, j_{k,2}, \dots$ are indices in J_k^- such that $Y_{j_{k,1}} \leq Y_{j_{k,2}} \leq \dots$ Finally, we take a greedy approach to maximize $\sum_{(i,j)\in M}(Y_i - Y_j)^2$ and for $k \in [K-1]$, we diagonally match the positive samples from left bin with the negative samples from each bin i.e. for any given k we match $(i_{k,1}, j_{k+1,1}), (i_{k,2}, j_{k+1,2}), \dots$ until either one of the positive or negative samples are matched completely or for some ℓ, $Y_{i_{k,\ell}} < Y_{j_{k+1,\ell}}$. Observe, by the choice of binning, for matched samples it always holds that

$$Z_{i_{k,\ell}} \leq Z_{j_{k+1,\ell}} \text{ and } Y_{i_{k,\ell}} \geq Y_{j_{k+1,\ell}}$$

i.e. we satisfy the anti-monotonicity constraint if we choose our weights for matched samples to be simply their respective Y value.

4 Power analysis

In this section, we study the power of our testing procedure under the following general model. We assume that we are given triples $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = (X_i, Y_i, Z_i)_{i\in[n]} \overset{\text{iid}}{\sim} P$, drawn according to the model

$$\mathbf{X} = \mu(\mathbf{Y}, \mathbf{Z}) + \boldsymbol{\zeta}, \tag{16}$$

where $\mu : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ (applied componentwise) is a measurable function, and with $(Y_i, Z_i)_{i\in[n]} \overset{\text{iid}}{\sim} P_{Y,Z}$ drawn independently from $(\zeta_i)_{i\in[n]} \overset{\text{iid}}{\sim} P_\zeta$. We suppose that P_ζ has mean 0 and unknown variance $\sigma^2 > 0$. Throughout this section, we further assume that the functions (ψ_ℓ) employed in our testing procedure are linear in both coordinates. In order to state the power guarantees under this signal plus noise model, we first introduce some notation.

By a *test function*, we mean a measurable function $\phi : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to [0,1]$, and say that ϕ is a *valid* test or *controls the Type I error* over H_0^{ICI} if

$$\sup_{P \in H_0^{\mathrm{ICI}}} \mathbb{E}_P \left[\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\right] \leq \alpha. \tag{17}$$

We denote by $\mathrm{Mon}(\mathcal{Z})$ the set of all measurable functions that are monotonic on the partially ordered set $(\mathcal{Z}, \preceq)$. Further, we define

$$\mathrm{ISS}_n = \inf_{g \in \mathrm{Mon}(\mathcal{Z})} \mathbb{E}_{P^n_{Y,Z}} \left[\|\mu(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2\right], \quad \widehat{\mathrm{ISS}}_n = \inf_{g \in \mathrm{Mon}(\mathcal{Z})} \|\mu(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2, \tag{18}$$

referring to them as the oracle and empirical *isotonic signal strength* respectively. We denote the corresponding oracle and empirical L_2 projections as μ_{ISO} and $\widehat{\mu}_{\mathrm{ISO}}$.

We will shortly demonstrate that the ability of our test procedure to distinguish the alternatives from the class of null models $\mathcal{H}_0^{\mathrm{ICI}}$ is governed by the empirical isotonic signal strength. In fact, in Section 4.2, we show that this connection is not exclusive to our testing procedure: the oracle ISS_n serves as a fundamental measure for characterizing the hardness of any valid test procedure. With this hardness result in mind, ISS_n can be interpreted as the *distance* of the model P from null class $\mathcal{H}_0^{\mathrm{ICI}}$. If ISS_n is too small, then the model P becomes essentially indistinguishable from $\mathcal{H}_0^{\mathrm{ICI}}$, implying that we can barely outperform the trivial testing procedure that ignores the data and rejects the null hypothesis randomly with probability α. Conversely, once $\widehat{\mathrm{ISS}}_n$ exceeds a threshold that we determine, the power of the test can approach 1, especially with appropriately designed matching schemes.

The organization of the rest of the section is as follows: in Section 4.1, we state asymptotic upper and lower bounds on the power of our test, optimized over all matching schemes, and conditioned on $(\mathbf{Y}, \mathbf{Z})$ under the general class of alternatives given in (16). While these power guarantees correspond to the max-weight matching, which relies on the oracle knowledge of μ, we show in Section 4.1.1 that even with an consistent estimate $\hat{\mu}$, we can derive very similar power guarantees. Next, in Section 4.2, we demonstrate how under suitable model assumptions, the oracle ISS_n characterizes the hardness of this test, and this connection is not limited to our proposed test procedure. Finally, in Section 4.3, we specialize our power guarantees to the special case of partially linear Gaussian models, and show that even without knowledge of the oracle μ, we can achieve near-optimal power guarantees with some of the natural choices of matching schemes proposed in Section 3.3.

4.1 ISS dictates the power of our testing procedure

To analyze the best-case performance of our testing procedure, we study the conditional power $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}$ of max-weight matching from Section 3.2, while the power analysis for general matching schemes is deferred to Theorem B.9 in the appendix. We assume that the distributions $P_{Y,Z}$ and P_ζ in (16) do not depend on sample size n, while the regression function $\mu(\cdot, \cdot)$ does, but we suppress the dependence of n in our notation for simplicity. A more detailed finite-sample power analysis is presented in the Appendix B.2, where the exact dependence of the power on ISS_n and other related terms is discussed.

Theorem 3. *Suppose that* $\left(\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty \vee \|\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4\right)/\widehat{\mathrm{ISS}}_n = \mathrm{o}_P(1)$*. Then, under the model* (16)*, the conditional power of oracle matching satisfies*

$$\Phi\left(\frac{\widehat{\mathrm{ISS}}_n}{\sqrt{2}\sigma} - \bar{\Phi}^{-1}(\alpha)\right) - \mathrm{o}_P(1) \leq \mathbb{P}\{p \leq \alpha \mid \mathbf{Y},\mathbf{Z}\} \leq \Phi\left(\frac{\widehat{\mathrm{ISS}}_n}{\sigma} - \bar{\Phi}^{-1}(\alpha)\right) + \mathrm{o}_P(1). \tag{19}$$

In both the upper and lower bounds, the signal-to-noise ratio $\widehat{\mathrm{ISS}}_n/\sigma$ dictates the dominant term, and hence the power of our test procedure. These bounds match up to a factor of $\sqrt{2}$ in the signal-to-noise ratio. The reason for this asymmetry is explained in Section 4.3.

It is important to note that the power result requires that $\left(\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty \vee \|\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4\right)/\widehat{\mathrm{ISS}}_n$ is $\mathrm{o}_P(1)$, which is rather necessary and unavoidable. The above term quantifies the concentration of the signal $\mu(Y,Z)$ across samples, and the assumption ensures that the signal is not accumulated within a small subset of sample points. Otherwise, the power of our test will only depend on swaps of a small subset of matched pairs, leading to low power in practice. The exact dependence of this term on the upper and lower bounds is stated in Corollary 3.

4.1.1 Optimal power guarantees with an estimate of μ

We further note that the power guarantees for max-weight matching in Theorem 3 requires the oracle knowledge of μ, which is not accessible in practice. A natural solution for bypassing this issue is data splitting — i.e., we start with learning an estimate $\hat{\mu}$ from one random split of the data, and then implement the max-weight matching with this estimated $\hat{\mu}$ on the remaining split. It is important to note that learning $\hat{\mu}$ on a different split makes sure that the calculated weights are independent of $\mathbf{X}$, which is crucial for validity of our test. While the above outlined data splitting is more accurate, for the sake of simplicity, while stating the following result we assume that we have a prior data where we can learn $\hat{\mu}$. Finally, under suitable consistency assumptions (details are given below) on $\hat{\mu}$, we can recover the similar power guarantees as in Theorem 3.

Theorem 4. *Consider the setting from Theorem 3. Suppose we use plugin oracle matching with an estimate $\hat{\mu}$ learnt on a prior data, where $\hat{\mu}$ satisfies*

$$\|\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \mu(\mathbf{Y},\mathbf{Z})\|_2/\widehat{\mathrm{ISS}}_n, \quad \|\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \mu(\mathbf{Y},\mathbf{Z})\|_\infty^4/\widehat{\mathrm{ISS}}_n \ \text{ is } \ \mathrm{o}_P(1).$$

Then, the conditional power satisfies (19)*.*

To summarize, with or without the oracle knowledge of μ, the optimal power of our test is governed by $\widehat{\mathrm{ISS}}_n$. As briefly mentioned in the introduction of this section, the connection of ISS with power is not limited to the particular testing procedure, we propose. In fact, under additional model assumptions, with any valid testing procedure, it is impossible to attain non-trivial power if the oracle ISS_n is too small. We state and prove this in the following subsection.

4.2 ISS characterizes hardness of testing the null H_0^{ICI}

In this subsection, we aim to characterize the hardness of testing the null H_0^{ICI} using the notion of ISS. In particular, via the oracle ISS_n under certain model assumptions, we identify the alternative models that are indistinguishable from the class of null models using any valid testing procedure i.e. any test Φ that satisfies (17). Towards this goal, we start with a simple total-variation calculation that gives a naive upper bound on power function, uniformly valid for any test procedure.

Lemma 5. *Fix $\alpha \in (0,1)$. For any test ϕ that satisfies (17), and for any distribution $P_{X,Y,Z}$,*

$$\mathbb{E}_{P_{X,Y,Z}}[\phi(\mathbf{X},\mathbf{Y},\mathbf{Z})] \leq \alpha + \inf_{Q_{X,Y,Z}\in H_0^{\mathrm{ICI}}} \mathrm{d}_{\mathrm{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right).$$

The offset total variation term can be read as the distance of $P_{X,Y,Z}$ from null class H_0^{ICI}, meaning the closer P is to null models, the harder it will be to get non-trivial power against P. While in general it is hard to derive exact expressions for the total variation term, under some model restrictions on P, we can come up with interpretable upper bounds for the same. Few such examples are listed below.

Gaussian setting: In the first example, we consider a special class of Gaussian alternatives, which is given by (16) where we assume that $P_\zeta = \mathcal{N}(0,\sigma^2)$ for some $\sigma > 0$. In this special case, ISS gives a meaningful upper bound on the offset total-variation term, up to a constant.

Corollary 1. *Suppose, $P_\zeta = \mathcal{N}(0,\sigma^2)$ and $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ satisfy (16). Then, for any test ϕ that satisfies (17), we have*

$$\mathbb{E}_P[\phi(\mathbf{X},\mathbf{Y},\mathbf{Z})] \leq \alpha + \frac{\mathrm{ISS}_n}{2\sigma}.$$

Proof of Corollary 1. By Lemma 5, it is enough to argue that

$$\inf_{Q_{X,Y,Z}\in H_0^{\mathrm{ICI}}} \mathrm{d}_{\mathrm{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right) \leq \frac{\mathrm{ISS}_n}{2\sigma}.$$

Consider $\mu_{\mathrm{ISO}} \in \mathrm{Mon}(\mathcal{Z})$. Similar to (16), define a model $Q_{X,Y,Z}$ as

$$\mathbf{X} = \mu_{\mathrm{ISO}}(\mathbf{Z}) + \boldsymbol{\zeta}, \quad (Y_1,Z_1),\cdots,(Y_n,Z_n) \overset{\mathrm{iid}}{\sim} P_{Y,Z}, \quad \zeta_1,\cdots,\zeta_n \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,\sigma^2).$$

By definition, $Q_{X,Y,Z} \in H_0^{\mathrm{ICI}}$ and further,

$$\begin{aligned}\mathrm{d}_{\mathrm{TV}}(P_{X,Y,Z}, Q_{X,Y,Z}) &\leq E_{P_{Y,Z}}\left[\mathrm{d}_{\mathrm{TV}}\left(\mathcal{N}\left(\mu(\mathbf{Y},\mathbf{Z}),\sigma^2 I_n\right), \mathcal{N}\left(\mu_{\mathrm{ISO}}(\mathbf{Z}),\sigma^2 I_n\right) \mid \mathbf{Y},\mathbf{Z}\right)\right] \\ &= \frac{\mathbb{E}_{P_{Y,Z}}\|\mu(\mathbf{Y},\mathbf{Z}) - \mu_{\mathrm{ISO}}(\mathbf{Z})\|_2}{2\sigma} = \frac{\mathrm{ISS}_n}{2\sigma}.\end{aligned}$$

This proves the result. □

As an immediate consequence, we note that it is impossible to achieve non-trivial power with any valid testing procedure when P_ζ is gaussian, and the ISS for alternative model is negligible. Next, we consider an example where $P_{X|Y,Z}$ is binary and we prove a similar characterization of power using ISS.

Binary setting: Suppose $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are generated from the model $P_{X,Y,Z}$ given by

$$X_i \sim \text{Ber}(\mu(Y_i, Z_i)), \quad (Y_1, Z_1), \cdots, (Y_n, Z_n) \overset{\text{iid}}{\sim} P_{Y,Z}. \tag{20}$$

We will show that ISS_n leads to a very interpretable upper bound on the power of our test, as long as $\mu(Y, Z)$ is away from the extremities i.e., 0 or 1. Below we state and prove the result.

Corollary 2. *Suppose* $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ *satisfy* (20) *where* $\mu(Y, Z) \in (\epsilon, 1-\epsilon)$ *almost surely for some* ϵ. *Then, for any test* ϕ *that satisfies* (17), *we have*

$$\mathbb{E}_P[\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})] \leq \alpha + \left(\frac{1}{\epsilon(1-\epsilon)}\right)^{1/2} \text{ISS}_n.$$

Proof of Corollary 2. Similar to the proof of Corollary 1, it is enough to argue that

$$\inf_{Q_{X,Y,Z} \in H_0^{\text{ICI}}} \text{d}_{\text{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right) \leq \left(\frac{1}{\epsilon(1-\epsilon)}\right)^{1/2} \text{ISS}_n.$$

Now, consider the model $Q_{X,Y,Z} \in H_0^{\text{ICI}}$ given by

$$X_i \sim \text{Ber}(\mu_{\text{ISO}}(Z_i)), \quad (Y_1, Z_1), \cdots, (Y_n, Z_n) \overset{\text{iid}}{\sim} P_{Y,Z}.$$

Firstly, we note that $\text{d}_{\text{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right)$ can be computed as

$$\mathbb{E}_{\mathbf{Y},\mathbf{Z}}\left[\text{d}_{\text{TV}}\left(\text{Ber}(\mu(Y_1, Z_1)) \times \cdots \times \text{Ber}(\mu(Y_n, Z_n)), \text{Ber}(\mu_{\text{ISO}}(Z_1)) \times \cdots \times \text{Ber}(\mu_{\text{ISO}}(Z_n))\right)\right]$$

Then, the inner total variation term can be upper bounded by the corresponding Hellinger distance as follows.

$$\begin{aligned}
&\text{d}_{\text{TV}}\left(\text{Ber}(\mu(Y_1, Z_1)) \times \cdots \times \text{Ber}(\mu(Y_n, Z_n)), \text{Ber}(\mu_{\text{ISO}}(Z_1)) \times \cdots \times \text{Ber}(\mu_{\text{ISO}}(Z_n))\right) \\
&\quad \leq \sqrt{2} \cdot H\left(\text{Ber}(\mu(Y_1, Z_1)) \times \cdots \times \text{Ber}(\mu(Y_n, Z_n)), \text{Ber}(\mu_{\text{ISO}}(Z_1)) \times \cdots \times \text{Ber}(\mu_{\text{ISO}}(Z_n))\right)
\end{aligned}$$

Now, for squared Hellinger distance, $H^2(P_1 \times \cdots \times P_k, Q_1 \times \cdots Q_k)$ is bounded by $\sum_{i=1}^k H^2(P_i, Q_i)$ and then, using Lemma D.18, the squared Hellinger distance can be further upper bounded by

$$\sum_{i=1}^n H^2\left(\text{Ber}(\mu(Y_i, Z_i)), \text{Ber}(\mu_{\text{ISO}}(Z_i))\right) \leq \sum_{i=1}^n \frac{(\mu(Y_i, Z_i) - \mu_{\text{ISO}}(Z_i))^2}{2\mu(Y_i, Z_i)(1 - \mu(Y_i, Z_i))}$$

Finally, since $\mu(Y, Z) \in (\epsilon, 1-\epsilon)$ almost surely, the final bound from above can be further upper bounded by $\frac{1}{\epsilon(1-\epsilon)} \cdot \|\mu(\mathbf{Y}, \mathbf{Z}) - \mu_{\text{ISO}}(\mathbf{Z})\|_2^2$. This concludes the proof. □

While for both examples, the oracle ISS_n is used to bound power, in (19) the conditional power of our test is controlled by the empirical variant $\widehat{\text{ISS}}_n$. While there is always a gap, under 'nice' settings both the oracle and empirical ISS terms will lie very close to each other. This justifies the nomenclature of 'Isotonic signal strength', at least for the gaussian and bernoulli examples. Beyond these settings, we can possibly define some other variants of oracle ISS_n (e.g., as we see in the analysis in Appendix C, we can consider $\|\mu(\mathbf{Y}, \mathbf{Z}) - \widetilde{\mu}_{\text{ISO}}(\mathbf{Z})\|_2$ where $\widetilde{\mu}_{\text{ISO}}(\mathbf{Z})$ is the isotonic L_1 projection.) to achieve similar upper bounds on power.

4.3 Near-optimal power guarantees without oracle knowledge of μ, and without sample-splitting

In earlier sections, we have established maximal power guarantees of our test procedure via the max-weight matching scheme; these are valid even without the oracle knowledge of μ with the cost of some additional consistency assumptions. In this subsection, we show that we can achieve similar non-trivial power guarantees with more natural choices of matching schemes. To prove such results, we specialize to the following model class, and show that some of these natural choices from Section 3.3 lead to near-optimal power guarantees.

Consider the class of partially linear gaussian models given by

$$(X, Y, Z) \text{ satisfy (16) with } P_\zeta = \mathcal{N}(0, \sigma^2),\; \mu(Y, Z) = \mu_0(Z) + \beta\, Y \tag{21}$$

for some $\mu_0 \in \mathrm{Mon}(\mathcal{Z})$ and some $\beta \in \mathbb{R}_{\geq 0}$. Without loss of generality, we further assume $\mathbb{E}\, Y = 0$. Observe that $\mathcal{H}_0^{\mathrm{st}}$ is satisfied if and only if $\beta = 0$. Thus, after parametrizing this model class with a single parameter β, we can consider a much simpler testing problem: $\mathcal{H}_0 : \beta = 0$ against $\mathcal{H}_1 : \beta > 0$. Under this specific model class, $\mathbf{Y}^T(\mathbf{X} - \mu(\mathbf{Z}))$ is a sufficient statistic for this parametric family indexed by β with

$$\mathbf{Y}^T(\mathbf{X} - \mu_0(\mathbf{Z})) \mid \mathbf{Y}, \mathbf{Z} \sim \mathcal{N}\big(\beta\|\mathbf{Y}\|_2^2, \sigma^2\|\mathbf{Y}\|_2^2\big).$$

Hence, an oracle test for $\mathcal{H}_0 : \beta = 0$ can be built by thresholding this sufficient statistic at $\sigma \cdot \|\mathbf{Y}\|_2 \cdot \Phi^{-1}(1-\alpha)$, where Φ denotes the standard normal distribution function, and the power of this oracle test is given by

$$\mathbb{P}\big\{\mathcal{N}(\beta\|\mathbf{Y}\|_2^2, \|\mathbf{Y}\|_2^2) > \sigma \cdot \|\mathbf{Y}\|_2 \cdot \bar{\Phi}^{-1}(\alpha)\big\} = \Phi\left((\beta/\sigma) \cdot \|\mathbf{Y}\|_2 - \bar{\Phi}^{-1}(\alpha)\right).$$

Observe by weak law of large numbers, $\big|\|\mathbf{Y}\|_2 - \big(\mathrm{Var}(Y)\big)^{1/2}\big| = o_P(1)$, and hence for large sample sizes, the oracle power can be computed as $\Phi\left((\beta/\sigma) \cdot \big(\mathrm{Var}(Y)\big)^{1/2} - \bar{\Phi}^{-1}(\alpha)\right) \pm o_P(1)$. We would shortly present that matching schemes like immediate-neighbor and cross-bin matching depicts a very similar behavior in power.

In order to state these results, we go back to the asymptotic framework from Section 4.1 where now, the dependence of n on $\mu(Y, Z)$ is only through β and the function μ_0 remains fixed (to make this explicit, we write β_n now onwards). Under this framework and the model (21), we first note the following bounds on oracle ISS_n.

Lemma 6. *Under the model class (21), it holds that*

$$\sqrt{n}\beta_n\, \big(\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]\big)^{1/2} \lessapprox\; \mathrm{ISS}_n \leq \sqrt{n}\beta_n\big(\mathrm{Var}(Y)\big)^{1/2}$$

Proof.

$$\begin{aligned}\mathrm{ISS} = \mathbb{E}_{P_{Y,Z}}\left[\|\mu(\mathbf{Y}, \mathbf{Z}) - \mu_{\mathrm{ISO}}(\mathbf{Z})\|_2\right] &\leq \mathbb{E}_{P_{Y,Z}}\left[\|\mu(\mathbf{Y}, \mathbf{Z}) - \mu_0(\mathbf{Z})\|_2\right] \\ &= \mathbb{E}_{P_{Y,Z}}\left[\|\beta_n \mathbf{Y}\|_2\right] \leq \beta_n\Big(\mathbb{E}_{P_{Y,Z}} \sum_i Y_i^2\Big)^{1/2} = \sqrt{n}\beta_n\big(\mathrm{Var}(Y)\big)^{1/2}.\end{aligned}$$

$$\begin{aligned}\mathrm{ISS} = \mathbb{E}_{P_{Y,Z}}\left[\|\mu(\mathbf{Y},\mathbf{Z}) - \mu_{\mathrm{ISO}}(\mathbf{Z})\|_2\right] &\geq \mathbb{E}_{P_{Y,Z}}\left[\|\mu(\mathbf{Y},\mathbf{Z}) - \mathbb{E}\big[\mu(\mathbf{Y},\mathbf{Z}) \mid \mathbf{Z}\big]\|_2\right] \\ &= \mathbb{E}_{P_{Y,Z}}\left[\|\beta_n\ \big(\mathbf{Y} - \mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]\big)\|_2\right] \gtrapprox \sqrt{n}\beta_n\ \big(\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]\big)^{1/2}.\end{aligned}$$

□

A direct implication of this lemma is that the threshold $1/\sqrt{n}$ determines a phase-transition of the power analysis. More specifically, if $\beta_n \ll 1/\sqrt{n}$, then by Corollary 1, every valid testing procedure is powerless. Hence, to ensure that we at least have non-trivial power, we will assume β_n scales as $\omega(1/\sqrt{n})$. Further, after scaling with $\sqrt{n}\beta_n$, the difference in between the upper and lower bound is driven by $\mathrm{Var}\big(\mathbb{E}[Y \mid Z]\big)$ which implies that if the conditional mean function is smooth, then upper and lower bounds are in fact very close.

Further Lemma 6 implies that under the model class (21) and mild assumptions stated in Theorem 3, the conditional power for max-weight matching satisfies

$$\begin{aligned}\Phi\left(\sqrt{n}\beta_n\left\{\frac{\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]}{2\sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right) - o_P(1)\ &\leq \mathbb{P}\{p \leq \alpha \mid \mathbf{Y},\mathbf{Z}\} \\ &\leq \Phi\left(\sqrt{n}\beta_n\left\{\frac{\mathrm{Var}(Y)}{2\sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right) + o_P(1). \quad (22)\end{aligned}$$

Below we show that other heuristic matching schemes from Section 3.3 match these power guarantees, taking advantage of the linearity in this model class. More precisely, if for our matched pairs $Z_{i_\ell} \approx Z_{j_\ell}$, then we have

$$\mu(Y_{i_\ell}, Z_{i_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell}) \approx \beta_n\big(Y_{i_\ell} - Y_{j_\ell}\big) \propto \big(Y_{i_\ell} - Y_{j_\ell}\big).$$

Since our test procedure depends on weights only through their 'in-pair differences', we can simply choose our weights to be the corresponding Y values without compromising much in power. In particular, the asymptotic power for neighbor matching matches the lower bound in (22) up to a constant. The result is formally stated below.

Theorem 7. *Let β_n be $\omega(1/\sqrt{n})$. Suppose Y is bounded, and $\mu_0(Z)$ is a sub-Gaussian random variable. Then, the power for immediate neighbor matching satisfies*

$$\left|\mathbb{P}\{p \leq \alpha \mid \mathbf{Y},\mathbf{Z}\} - \Phi\left(\sqrt{n}\beta_n\left\{\frac{\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]}{4\sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right)\right| = o_P(1).$$

While this is a good starting point, immediate neighbor matching is inefficient since half of the sample points remain unmatched in this strategy. Cross-bin matching from Section 3.3 on the other hand improves in number of matched samples, and thus that is also reflected in power guarantees.

Theorem 8. *Consider the setting as in Theorem 7 and suppose, the number of bins i.e., K is chosen to satisfy $K = \omega_P(\sqrt{n})$. Then, under suitable smoothness assumptions (stated formally in Theorem B.11), the power for cross-bin matching satisfies*

$$\mathbb{P}\{p \leq \alpha \mid \mathbf{Y},\mathbf{Z}\} \geq \Phi\left(\sqrt{n}\beta_n\left\{\frac{\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]}{2\sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right) - o_P(1).$$

Further, if $P_{Y|Z}$ is symmetric almost surely, then the power of cross-bin matching further satisfies

$$\left|\mathbb{P}\{p \le \alpha \mid \mathbf{Y}, \mathbf{Z}\} - \Phi\left(\sqrt{n}\beta_n\left\{\frac{\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]}{\sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right)\right| = \mathrm{o}_P(1).$$

The first part of the result states that asymptotically, under no further model assumptions, the power of cross bin matching is no smaller than the lower bound from (22). However, with the additional assumption of symmetry of $P_{Y|Z}$, cross-bin matching can recover power as good as the upper bound in (22), when the conditional variance of Y given Z is constant across $\mathcal{Z}$-space or more particularly, when $\mathbb{E}[\mathrm{Var}(Y \mid Z)]$ is same as $\mathrm{Var}(Y)$.

Explaining the gap between the lower and upper bounds on power We see that under this specialized model, $\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]/\sigma^2$ plays the role of signal-to-noise ratio in the power results stated above. Further, we rewrite the asymptotic power from Theorem 7 as

$$\Phi\left(\sqrt{n}\beta_n\left\{\frac{\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]}{2 \cdot 2 \cdot \sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right),$$

and stress that the source of these two factors of 2 are different. Below, we explain the sources for each of the two in details.

- *Insufficient matched pairs:* For neighbor matching, a matched pair of consecutive observations contribute only if the anti-monotonicity holds. Since asymptotically, neighbours are almost distinguishable, the anti-monotonicity holds with probability $\approx 1/2$. As a result, half of the samples are discarded in the process of matching. This contributes to one of the factors of 2. Cross-bin matching, and max-weight matching however leaves only $\mathrm{o}(n)$ many samples unmatched, and that helps in recovering the gap.
- *Asymmetry in matching:* The other factor of 2 is rather unavoidable. We observe that for general matching schemes, in our general power calculations (Theorem B.9) the dominating term is governed by the term $\sum_\ell (\mu_{i_\ell} - \mu_{j_\ell})(w_{i_\ell} - w_{j_\ell})$ which is maximized when for matched pairs,

  $$w_{i_\ell} \approx \mu_{i_\ell}, w_{j_\ell} \approx \mu_{j_\ell}, \quad \text{and} \quad \mu_{i_\ell} \approx -\mu_{j_\ell}.$$

  Without further model assumptions, it is specially hard to guarantee $\mu_{i_\ell} \approx -\mu_{j_\ell}$, which contributes to the other factor of 2 in Theorem 7. This also explains the $\sqrt{2}$ in Theorem 3, and the factor of 2 in Theorem 8. Under the gaussian partial model, $\mu_{i_\ell} - \mu_{j_\ell} \approx Y_{i_\ell} - Y_{j_\ell}$ if the matched observations lie in vicinity in the $\mathcal{Z}$-space. Thus, the assumption of $P_{Y|Z}$ being symmetric recovers this factor of 2, as we note in Theorem 8.

5 Simulations

In this section, we evaluate the performance of our method on simulated data, and compare different matching strategies. For simplicity, we focus on the univariate case $\mathcal{Z} = \mathbb{R}$.

5.1 Conservativeness under the null H_0^{ICI}

Theorem 1 establishes valid, finite-sample Type I error control for our method. The purpose of this section is to evaluate how conservative the Type I error is under various null distributions. Because our inference relies on the fact that matched pairs (X_{i_ℓ}, X_{j_ℓ}) are stochastically ordered under the null, intuitively the conservativeness test depends on the strength of monotonicity in the conditional distribution.

To see how the dependence between X and Z affects the rejection probability, we sample X from an additive noise model

$$X \mid Y, Z \sim \mathcal{N}(\mu(\gamma Z), 1),$$

where Y, Z are independent standard normal random variables. As long as μ is nondecreasing and $\gamma \geq 0$, this joint distribution belongs to the null H_0^{ICI}. The scalar γ controls the strength of the monotonicity of $X \mid Z$. In particular, as $\gamma \downarrow 0$ we expect the Type I error $\mathbb{P}\{p \leq \alpha\}$ to approach α.

In our simulations, we consider two functions μ, the identity $\mu(z) = z$ and the standard Gaussian cdf $\mu(z) = \Phi(z)$. Figure 2 shows the Type I error as a function of γ for three levels of α. We see find similar results for each α, where the test typically becomes more conservative as γ increases. Under the null, our test is more conservative for cross-bin matching than for neighbor matching, since the Z values are further apart in cross-bin matching.

5.2 Power under alternatives

In Section 4.3 we show that our heuristic methods—neighbor matching and cross-bin matching—achieve asymptotic power one in the partial linear model (21) provided the signal β_n exceeds the detection threshold $n^{-1/2}$. In this section, we simulate from the Gaussian linear model

$$X \mid Y, Z \sim \mathcal{N}(\beta_n Y + \gamma Z, 1),$$

with $\beta_n = n^{-1/3}$. The pair (Y, Z) is drawn from a bivariate Gaussian

$$\mathcal{N}\left(0, \begin{bmatrix} 1 & \rho_{YZ} \\ \rho_{YZ} & 1 \end{bmatrix}\right).$$

Figure 3 shows the power as a function of the sample size n for various choices of γ and ρ_{YZ}. In Setting 1, we set $\gamma = 1$, and cross-bin matching uniformly dominates neighbor matching because it allows us to make many more matches of similar quality. On the other hand, in Setting 2 we set $\gamma = 10$, so the strong dependence of X on Z means the quality of a match (i_ℓ, j_ℓ) degrades much more quickly as the gap $Z_{j_\ell} - Z_{i_\ell}$ increases. However, cross-bin matching still overtakes neighbor matching in sufficiently large samples. This is because the bin width decreases as n increases, so the quality of the cross-bin matches rivals that of the neighbor matches (with many more matches). The dependence ρ_{YZ} between Y and Z does not have a major impact on the power of these two methods.

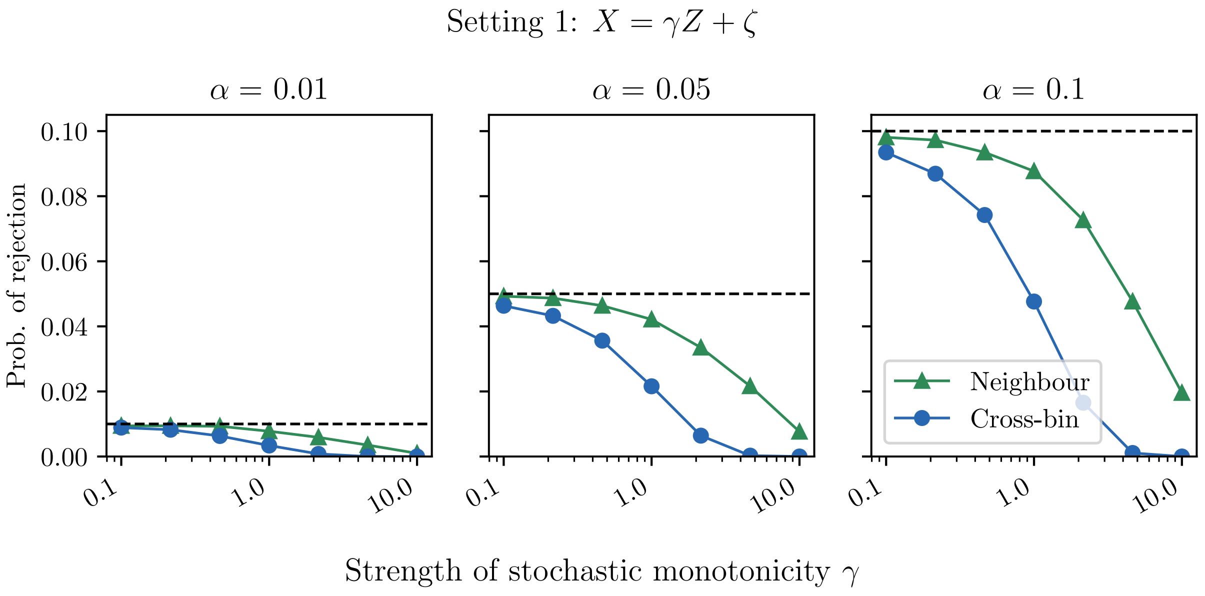
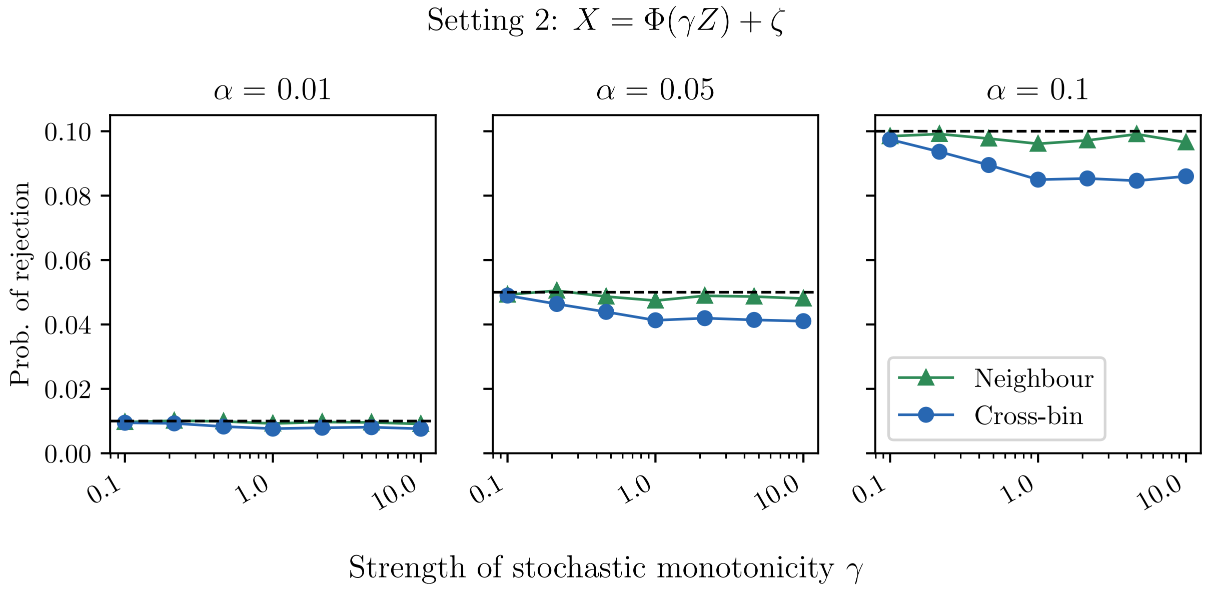

Setting 1: $X = \gamma Z + \zeta$

Figure 2: Simulation results illustrating Type I error control under the null H_0^{ICI} for two forms of the conditional mean $\mathbb{E}[X \mid Z]$. Each subplot shows the rejection probability, averaged over 10^6 simulation trials, as a function of the strength of stochastic monotonicity γ.


Setting 2: $X = \Phi(\gamma Z) + \zeta$

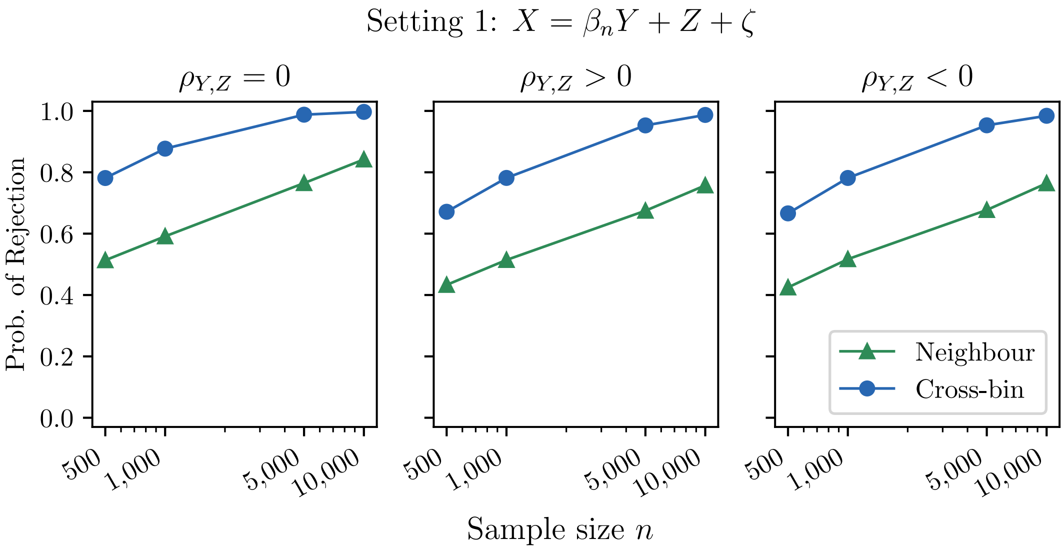
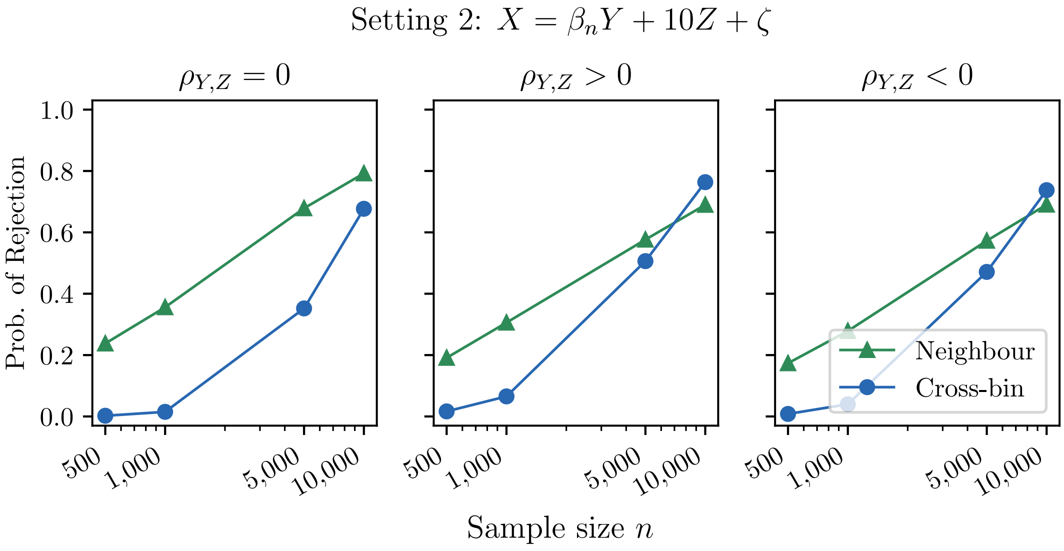


Figure 3: Simulation results demonstrating power for two alternatives at level $\alpha = 0.1$. Each subplot shows the rejection probability, averaged over 10^4 simulation trials, as a function of the sample size n. Columns correspond to different relationships between Y and Z. In each setting, X follows a Gaussian linear model with mean $\beta_n Y + \gamma Z$, where $\beta_n = n^{-1/3}$ and $\gamma = 1$ (above) or $\gamma = 10$ (below).



6 Experiment on real data: Risk factors for Diabetes

In this section, we evaluate the performance of our proposed testing procedure on a real dataset using various matching schemes, and compare it against other well-established conditional independence testing methods in the literature. We use a dataset[3] on the incidence of diabetes among the Pima Indian population near Phoenix, Arizona, originally collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. The dataset contains 768 observations, and it includes information on whether each of the patient has been diagnosed with diabetes according to World Health Organization standards. Additional variables provide data on the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin levels, body mass index (BMI), diabetes pedigree function and age.

It is well-known that the likelihood of developing diabetes increases with age (e.g., CDC [4] lists advanced age as one of the risk factors for type 1 and type 2 diabetes). Therefore, if we choose X to represent the incidence of diabetes and Z as the age of the patient, then we would expect X to exhibit stochastic monotonicity with respect to Z. This relationship is also verified empirically in the left-most panel of Figure 5. Most of the other variables, such as `BloodPressure`, `BMI`, `Glucose`, and `Pregnancies`, are also considered potential risk factors for diabetes. In this experiment, we aim to determine whether these variables remain significant risk factors for diabetes, even after controlling for age.

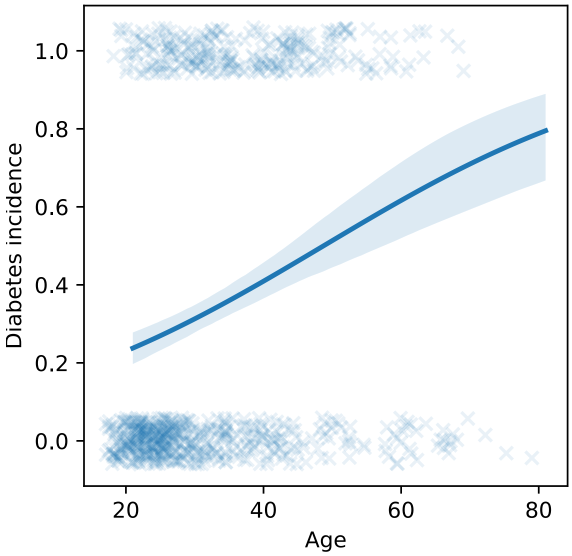


Figure 4: A scatter-plot (jittered for better visibility) of `Age` and `Diabetes Incidence` along with the fitted logistic regression model to demonstrate the stochastic monotonicity between them.

[3] The data for this experiment were obtained from `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`. Additional data descriptions can be found in Smith et al. (1988).

[4] For more details, refer to the list of diabetes risk factors from U.S. Centers for Disease Control and Prevention.

Experiment 1: marginal independence testing: In our first experiment, we consider six variables: `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, and `BMI`, and aim to assess whether each of them is an individual risk factor for diabetes incidence. Specifically, we test the hypothesis $H_0 : X \perp\!\!\!\perp Y$, where Y represents one of the six variables listed above. For this purpose, we will be using the permutation test for independence with $T(\mathbf{X}, \mathbf{Y}) = \mathbf{X}^T\mathbf{Y}$, as outlined in Section 1.2.

Experiment 2: conditional independence testing, after controlling for age: Next, for the same set of six choices for Y, we test the hypothesis $H_0^{\mathrm{CI}} : X \perp\!\!\!\perp Y \mid Z$, where Z denotes `age`. This allows us to identify risk factors for diabetes after controlling for age. As noted earlier, we expect the distribution of $X \mid Z = z$ to be stochastically monotone in z, which supports the application of the `PairSwap-ICI` testing procedure developed in this paper for this purpose.

Experiment 3: conditional independence testing, with synthetic control $\tilde{X}$: Finally, we consider a semi-synthetic experiment where X is replaced by ghost observations $\tilde{X}$, generated from an estimated model for $P_{X|Z}$ that satisfies stochastic monotonicity. We then test the hypothesis $\tilde{\mathcal{H}}_0 : \tilde{X} \perp\!\!\!\perp Y \mid Z$ for the same choices of Y from Experiment 1. Since $\tilde{X}$ is generated solely based on Z, the null hypothesis of conditional independence holds trivially, and we therefore expect our `PairSwap-ICI` testing procedure to yield significantly larger p-values compared to the previous experiment. Since X is binary, it suffices to fit an isotonic regression to estimate the conditional mean $\mathbb{E}[X \mid Z]$ and then sample $\tilde{X}$ from the Bernoulli distribution with this fitted conditional mean. Following the theory of Henzi, Ziegel and Gneiting (2021, Theorem 1), this is the best approximation for $P_{X|Z}$ as per *continuous ranked probability score (CRPS)*, while respecting the monotonocity constraint.

		Pregnancies	Glucose	Blood pressure	Skin thickness	Insulin	BMI
Permutation test (testing marginal indep.)		**0.001**(0.00)	**0.001**(0.00)	0.151(0.003)	0.12(0.002)	**0.024**(0.001)	**0.001**(0.00)
PairSwap-ICI	neighbor matching	0.438(0.005)	**0.002**(0.00)	0.495(0.005)	0.234(0.004)	0.157(0.003)	**0.031**(0.001)
	cross-bin matching	0.415(0.005)	**0.001**(0.00)	0.49(0.005)	0.146(0.003)	0.123(0.003)	**0.001**(0.00)
PairSwap-ICI (with synthetic control)	neighbor matching	0.507(0.005)	0.504(0.005)	0.504(0.005)	0.504(0.005)	0.507(0.005)	0.508(0.005)
	cross-bin matching	0.509(0.005)	0.522(0.005)	0.516(0.005)	0.509(0.005)	0.509(0.005)	0.521(0.005)

Table 1: p-values, averaged over 3000 random sub-samples along with the estimated standard errors (within brackets) for the different tests from different experiments, as outlined in Section 6. The p-values significant at the 0.05 level are marked in bold.

Results: For each experiment, we generate 3,000 random sub-samples of the data, each consisting of half the size of the full dataset. We then compute p-values using the permutation test for marginal independence and the `PairSwap-ICI` test for conditional independence. For experiments involving synthetic control, which require estimating $P_{X|Z}$, the sub-sampled data

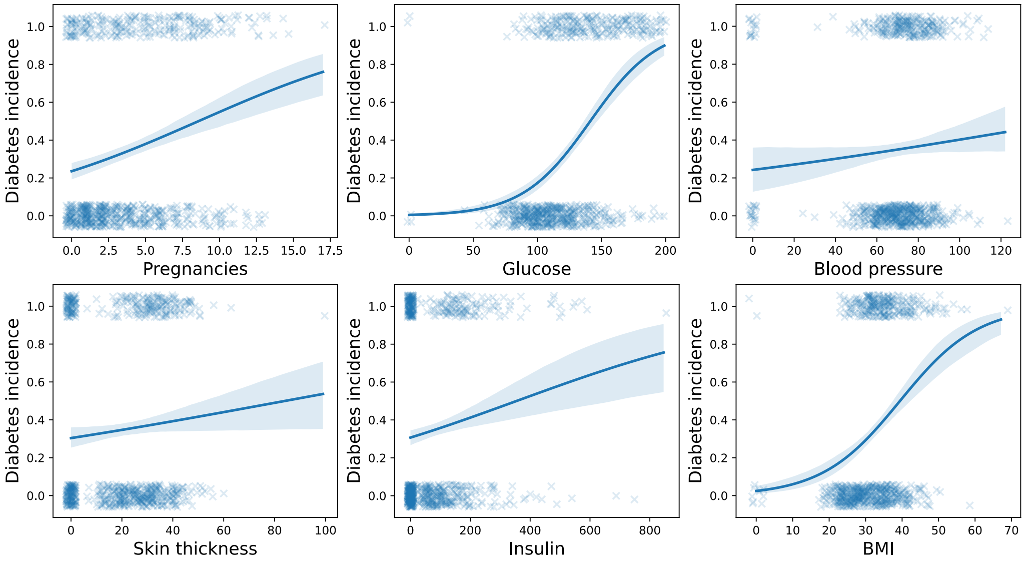


Figure 5: Scatter plots of X (jittered for better visibility) and other feature variables along with the fitted logistic regression models to demonstrate the dependence among these variables and `Diabetes Incidence`.

is further divided into training and test sets, with $\hat{P}_{X|Z}$ being computed in training set. For all the experiments, p-values are computed in the test set. Finally, we report the average p-values from the $3,000$ sub-samples, along with the corresponding choice of the Y variable in Table 1.

We observe that according to the marginal independence test, most variables show significant dependence with the incidence of diabetes at the 0.05 level of significance with the exception of `SkinThickness` and `BloodPressure`. Among the results from our conditional independence tests conditioned on `Age` in Table 1, the p-values for `Pregnancies` and `Insulin` are insignificant. This suggests that, after controlling for age, the data does not provide sufficient evidence to support them as risk factors for diabetes. However, we observe that both variables show marginal dependence with diabetes incidence, as visually demonstrated in Figure 5 and confirmed by Experiment 1.

On the other hand, `Glucose` and `BMI` both are identified as potential risk factors at the 0.05 level of significance for diabetes by the marginal independence test and the `PairSwap-ICI` test, even after controlling for age. We also note that all the averaged p-value from Experiment 3 with synthetic control $\tilde{X}$ is concentrated around 0.5. Since $(\tilde{X}, Y, Z)$ satisfy H_0^{ICI} the p-values from Experiment 3 should be roughly uniform, and thus this behaviour is expected as per the result we have established in Theorem 1.

7 Discussion

In this paper, we have developed a nonparametric test of conditional independence assuming only stochastic monotonicity of the conditional distribution $P_{X|Z}$. This nonparametric constraint is natural in many applications and allows us to circumvent the impossibility of assumption-free conditional independence testing (Shah and Peters, 2020). We introduced a variety of approaches to constructing a valid test statistic. Our test controls the type I error in finite samples and has power against an array of alternatives. We close our discussion with some interesting avenues for future work.

- *Optimal power in general settings.* Theorem 8 shows that the power for cross-bin matching can rival that of a parametric oracle with knowledge of the conditional mean μ, provided that the conditional distribution $P_{Y|Z}$ is symmetric. As we discussed, this condition appears to be essentially unavoidable. How can we achieve the oracle power against non-symmetric alternatives?

- *Avoiding data splitting.* The max-weight matching test derived in Section 3 requires modeling the conditional mean and conditional variance of the kernel $\psi(X_i, X_j)$ as a function of Y_i, Y_j, Z_i, Z_j. We proposed to estimate these moments on a hold-out data set. Can we instead perform cross-fitting and retain finite-sample error control?

- *Alternative methods.* A notable benefit of our stochastic monotonicity assumption is that we can consistently estimate the conditional distribution $P_{X|Z}$ using isotonic distributional regression (Henzi, Ziegel and Gneiting, 2021). Hence, an alternative approach to testing the restricted null H_0^{ICI} is to first estimate this conditional distribution on one split of the data, and then run a conditional independence test which assumes knowledge of $P_{X|Z}$ (Berrett et al., 2020; Candès et al., 2018). Since we are plugging in the estimated conditional distribution, such tests will only be valid asymptotically. Is there any way to modify such tests to be valid in finite samples?

- *Alternative shape constraints.* We view stochastic monotonicity as one form of positive dependence for the joint distribution (X, Z). Are there natural approaches to test conditional independence under other models of dependence, such as likelihood-ratio ordering or total positivity (Shaked and Shanthikumar, 2007)?

References

Albert, M., Laurent, B., Marrel, A. and Meynaoui, A. (2022) Adaptive test of independence based on HSIC measures. *The Annals of Statistics*, **50**, 858–879.

Azadkia, M. and Chatterjee, S. (2021) A simple measure of conditional dependence. *The Annals of Statistics*, **49**, 3070–3102.

Barber, R. F., Candès, E. J. and Samworth, R. J. (2020) Robust inference with knockoffs. *The Annals of Statistics*, **48**, 1409–1431.

Berrett, T. B., Kontoyiannis, I. and Samworth, R. J. (2021) Optimal rates for independence testing via U-statistic permutation tests. *The Annals of Statistics*, **49**, 2457–2490.

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

Berrett, T. B. and Samworth, R. J. (2021) USP: an independence test that improves on Pearson's chi-squared and the G-test. *Proceedings of the Royal Society A*, **477**, 20210549.

Berrett, T. B., Wang, Y., Barber, R. F. and Samworth, R. J. (2020) The conditional permutation test for independence while controlling for confounders. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **82**, 175–197.

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

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and their Application*, vol. 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

de Finetti, B. (1929) Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*, 179–190.

Duan, R. and Pettie, S. (2014) Linear-time approximation for maximum weight matching. *Journal of the ACM (JACM)*, **61**, 1–23.

Edmonds, J. (1965) Paths, trees, and flowers. *Canadian Journal of mathematics*, **17**, 449–467.

Gabow, H. N. (1985) A scaling algorithm for weighted matching on general graphs. In *26th Annual Symposium on Foundations of Computer Science (sfcs 1985)*, 90–100, IEEE.

Hagberg, A., Swart, P. and S Chult, D. (2008) Exploring network structure, dynamics, and function using NetworkX. Tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Henzi, A., Ziegel, J. F. and Gneiting, T. (2021) Isotonic distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **83**, 963–993.

Kalisch, M. and Bühlmann, P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, **8**.

Kim, I., Balakrishnan, S. and Wasserman, L. (2022) Minimax optimality of permutation tests. *The Annals of Statistics*, **50**, 225–251.

Kim, I., Neykov, M., Balakrishnan, S. and Wasserman, L. (2022) Local permutation tests for conditional independence. *Ann. Statist.*, **50**, 3388–3414.

Lundborg, A. R., Kim, I., Shah, R. D. and Samworth, R. J. (2024+) The Projected Covariance Measure for assumption-lean variable significance testing. *The Annals of Statistics, to appear. arXiv preprint arXiv:2211.02039.*

Lundborg, A. R., Shah, R. D. and Peters, J. (2022) Conditional independence testing in Hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **84**, 1821–1850.

Neykov, M., Balakrishnan, S. and Wasserman, L. (2021) Minimax optimal conditional independence testing. *Ann. Statist.*, **49**, 2151–2177.

Niu, Z., Chakraborty, A., Dukes, O. and Katsevich, E. (2024) Reconciling model-X and doubly robust approaches to conditional independence testing. *The Annals of Statistics*, **52**, 895–921.

O'Mahony, C., Jichi, F., Pavlou, M., Monserrat, L., Anastasakis, A., Rapezzi, C., Biagini, E., Gimeno, J. R., Limongelli, G., McKenna, W. J. et al. (2014) A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (HCM risk-SCD). *European Heart Journal*, **35**, 2010–2020.

Pfister, N., Bühlmann, P., Schölkopf, B. and Peters, J. (2018) Kernel-based tests for joint independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **80**, 5–31.

Phipson, B. and Smyth, G. K. (2010) Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.*, **9**, Art. 39, 14.

Shah, R. D. and Peters, J. (2020) The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.*, **48**, 1514–1538.

Shaked, M. and Shanthikumar, J. G. (2007) *Stochastic orders*. Springer Series in Statistics, Springer, New York.

Shevtsova, I. G. (2010) An improvement of convergence rate estimates in the Lyapunov theorem. *Doklady Mathematics*, **82**, 862–864.

Smith, J. W., Everhart, J. E., Dickson, W., Knowler, W. C. and Johannes, R. S. (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, 261, American Medical Informatics Association.

Yan, Z., Cai, M., Han, X., Chen, Q. and Lu, H. (2023) The interaction between age and risk factors for diabetes and prediabetes: a community-based cross-sectional study. *Diabetes, Metabolic Syndrome and Obesity*, 85–93.

A Proof of Theorem 2

The proof of this result follows the same structure as the proof of Theorem 1.

Step 1: some deterministic properties of the p-value. Define a function $\hat{p}_M : \mathbb{R}^n \times (\{\pm 1\}^L)^M \rightarrow [0,1]$ as

$$\hat{p}_M(\mathbf{x}; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}) = \frac{1 + \sum_{m=1}^M \mathbb{1}\left\{T(\mathbf{x}^{\mathbf{s}^{(m)}}) \geq T(\mathbf{x})\right\}}{1+M}.$$

As in the proof of Theorem 1, this function is monotone nonincreasing in each x_{i_ℓ}, and monotone nondecreasing in each x_{j_ℓ}.

Step 2: compare to the sharp null. Define $\mathbf{X}_\sharp$ as in the proof of Theorem 1. Following identical arguments as in that proof, we can verify that, for any fixed $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}$, it holds that

$$\hat{p}_M\big(\mathbf{X}_\sharp; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big) \preceq_{\mathrm{st}} \hat{p}_M\big(\mathbf{X}; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big)$$

conditional on $\mathbf{Y}, \mathbf{Z}$. Since $\hat{p}_M = \hat{p}_M(\mathbf{X}; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)})$ by construction, we therefore have

$$\mathbb{P}\left\{\hat{p}_M \leq \alpha \;\middle|\; \mathbf{Y}, \mathbf{Z}, \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\right\} \leq \mathbb{P}\left\{\hat{p}_M\big(\mathbf{X}_\sharp; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big) \leq \alpha \;\middle|\; \mathbf{Y}, \mathbf{Z}, \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\right\}.$$

Marginalizing over the random draw of the swaps, $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)} \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(\{\pm 1\}^L)$, we therefore have

$$\mathbb{P}\left\{\hat{p}_M \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} \leq \mathbb{P}\left\{\hat{p}_M\big(\mathbf{X}_\sharp; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big) \leq \alpha \;\middle|\; \mathbf{Y}, \mathbf{Z}\right\}.$$

Step 3: validity under the sharp null. We now need to verify the validity of the Monte Carlo p-value, under the sharp null. Unlike the first two steps, for this step the arguments are somewhat different than in the proof of Theorem 1.

First, let $\mathbf{s}^{(0)}$ be an additional draw from $\mathrm{Unif}(\{\pm 1\}^L)$, sampled independently from all other random variables. Then it holds that

$$(\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}) \overset{\mathrm{d}}{=} \big(\mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\big),$$

where $\circ$ denotes the elementwise product, and so

$$\hat{p}_M\big(\mathbf{X}_\sharp; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big) \overset{\mathrm{d}}{=} \hat{p}_M\Big(\mathbf{X}_\sharp; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\Big)$$

conditional on $\mathbf{Y}, \mathbf{Z}$. Moreover, by construction of the sharp null data $\mathbf{X}_\sharp$,

$$\mathbf{X}_\sharp \overset{\mathrm{d}}{=} (\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}$$

holds conditional on $\mathbf{Y}, \mathbf{Z}, \mathbf{s}^{(0)}, \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}$, and therefore

$$\hat{p}_M\Big(\mathbf{X}_\sharp; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\Big) \overset{\mathrm{d}}{=} \hat{p}_M\Big((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\Big)$$

holds conditional on $\mathbf{Y}, \mathbf{Z}, \mathbf{s}^{(0)}, \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}$. Combining all these calculations so far, then, we have

$$\hat{p}_M\big(\mathbf{X}_\sharp; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big) \overset{\mathrm{d}}{=} \hat{p}_M\Big((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\Big), \tag{23}$$

conditional on $\mathbf{Y}, \mathbf{Z}$.

Next we calculate this last p-value: by definition,

$$\begin{aligned}\hat{p}_M\Big((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}};\mathbf{s}^{(0)}\circ\mathbf{s}^{(1)},\dots,\mathbf{s}^{(0)}\circ\mathbf{s}^{(M)}\Big) &= \frac{1+\sum_{m=1}^M \mathbb{1}\left\{T\left((\mathbf{X}_\sharp)^{\mathbf{s}^{(m)}}\right)\geq T\left((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}\right)\right\}}{1+M}\\ &= \frac{\sum_{m=0}^M \mathbb{1}\left\{T\left((\mathbf{X}_\sharp)^{\mathbf{s}^{(m)}}\right)\geq T\left((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}\right)\right\}}{1+M},\end{aligned}$$

where the first step holds since, for each $m=1,\dots,M$,

$$\Big[(\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}\Big]^{\mathbf{s}^{(0)}\circ\mathbf{s}^{(m)}} = (\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}\circ\mathbf{s}^{(0)}\circ\mathbf{s}^{(m)}} = (\mathbf{X}_\sharp)^{\mathbf{s}^{(m)}}$$

by definition of the swap operation. In other words, the p-value $\hat{p}_M\Big((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}};\mathbf{s}^{(0)}\circ\mathbf{s}^{(1)},\dots,\mathbf{s}^{(0)}\circ\mathbf{s}^{(M)}\Big)$ is simply comparing the value of the statistic $T((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}})$ against the list of $M+1$ values $T((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}),\dots,T((\mathbf{X}_\sharp)^{\mathbf{s}^{(M)}})$. We therefore have

$$\mathbb{P}\left\{\hat{p}_M\Big((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}};\mathbf{s}^{(0)}\circ\mathbf{s}^{(1)},\dots,\mathbf{s}^{(0)}\circ\mathbf{s}^{(M)}\Big)\leq\alpha \;\Big|\; \mathbf{X}_\sharp,\mathbf{Y},\mathbf{Z}\right\}\leq\alpha,$$

since, conditional on $\mathbf{X}_\sharp,\mathbf{Y},\mathbf{Z}$, the sign vectors $\mathbf{s}^{(0)},\dots,\mathbf{s}^{(M)}$ are i.i.d., and therefore the rank of the statistic $T((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}})$ among the list $T((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}),\dots,T((\mathbf{X}_\sharp)^{\mathbf{s}^{(M)}})$ is uniformly distributed. Marginalizing over $\mathbf{X}_\sharp$, therefore,

$$\mathbb{P}\left\{\hat{p}_M\Big((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}};\mathbf{s}^{(0)}\circ\mathbf{s}^{(1)},\dots,\mathbf{s}^{(0)}\circ\mathbf{s}^{(M)}\Big)\leq\alpha \;\Big|\; \mathbf{Y},\mathbf{Z}\right\}\leq\alpha.$$

Finally, combining this with our earlier calculation (23), we have

$$\mathbb{P}\left\{\hat{p}_M\big(\mathbf{X}_\sharp;\mathbf{s}^{(1)},\dots,\mathbf{s}^{(M)}\big)\leq\alpha \;\big|\; \mathbf{Y},\mathbf{Z}\right\}\leq\alpha,$$

which completes the proof.