A kNN-based Non-Parametric Conditional Independence Test for Mixed Data and Application in Causal Discovery (Appendix)

Johannes Huegle (✉), Christopher Hagedorn, and Rainer Schlosser

University of Potsdam, Hasso Plattner Institute, Potsdam, Germany
{Johannes.Huegle,Christopher.Hagedorn,Rainer.Schlosser}@hpi.de

Appendix Overview

The following appendix complements the information provided in the paper *A kNN-based Non-Parametric Conditional Independence Test for Mixed Data and Application in Causal Discovery*.

In Appendix A, we recap the required assumptions and their implication for application. Further, we provide detailed proofs of the theoretical results on the robustness of `mCMIkNN`. In Appendix C, we provide details on constraint-based causal discovery and a detailed proof of the asymptotic consistency of `mCMIkNN`-based causal discovery. Moreover, we give a more detailed synthetic evaluation in Appendix D and further information on the real-world use case in Appendix E.

A `mCMIkNN`: Assumptions and Computational Complexity

In this section, we provide more information on the assumptions introduced in Sec. 3 *A Non-parametric Conditional Independence Test*, the computational complexity, and its implications for application.

First, recap all assumptions on P_{XYZ} and parameters k_{CMI} and k_{perm}.

Assumptions 1 *Let $(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}, \mathcal{B}, P_{XYZ})$ be a probability space defined on the metric space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ with dimensionality $d_x + d_y + d_z$, equipped with the Borel σ-algebra $\mathcal{B}$, and a regular joint probability measure P_{XYZ}. Throughout this work, we assume:*

- *(A1)* $P_{XY|Z}$ *is non-singular such that* $f \equiv \frac{dP_{XY|Z}}{d(P_{X|Z} \times P_{Y|Z})}$ *is well-defined, and assume, for some* $C > 0$, $f(x, y, z) < C$ *for all* $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$;
- *(A2)* $\{(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} : P_{XYZ}((x, y, z)) > 0\}$ *countable and nowhere dense in* $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$;
- *(A3)* $k_{CMI} = k_{CMI,n} \to \infty$ *and* $\frac{k_{CMI,n}}{n} \to 0$ *as* $n \to \infty$;
- *(A4)* $k_{perm} = k_{perm,n} \to \infty$ *and* $\frac{k_{perm,n}}{n} \to 0$ *as* $n \to \infty$.

In the following, we examine the above assumptions in more detail.

(A1): While rather technical, non-singularity is helpful for practice as it provides a sufficient condition that can be verified in data analysis.

Definition 1 (Non-singularity of $P_{XY|Z}$).
Let $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}, \mathcal{B}, P_{XYZ}$ be a probability space with marginal conditional probability measures, $P_{X|Z}$ and $P_{Y|Z}$. $P_{XY|Z}$ is non-singular if for any measurable set, $E \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, $a \in X \times Z$ and $b \in Y \times Z$, such that $P_{X|Z}(E_b) = 0$ and $P_{Y|Z}(E_a) = 0$, then $P_{XY|Z}(E) = 0$, where $E_b = \{(x, z) : (x, b, z) \in E\}$ and $E_a = \{(y, z) : (a, y, z) \in E\}$.

Assuming non-singularity of $P_{XY|Z}$ ensures absolute continuity $P_{XY|Z} \ll P_{X|Z} \times P_{Y|Z}$, i.e., the existence of the Radon-Nikodym derivative $f \equiv \frac{dP_{XY|Z}}{d(P_{X|Z} \times P_{Y|Z})}$ using Fubini's theorem and Radon-Nikodym's theorem, see [26, Thm. 2.2]. Further, given that f is well-defined, the existence of a $C > 0$ such that $f(x, y, z) < 0$ for all $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ is satisfied whenever the distribution is (i) (finitely) discrete, (ii) continuous, (iii) some dimensions are (countably) discrete and some are continuous, and (iv) a mixture of the previous cases, see [11]. Hence, for practice, checking the sufficient condition of non-singularity can be done by ensuring that there exists no set $E \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ such that $P_{XY|Z}(E) = 0$ while $P_{Y|Z}(E_a) = 0$ for $E_b = \{(x, z) : (x, b, z) \in E\}$ and $E_a = \{(y, z) : (a, y, z) \in E\}$, cf. [10,44].

(A2): Assumption (A2) is satisfied whenever the distribution of P_{XYZ} is (i) (finitely) discrete, (ii) continuous, (iii) some dimensions are (countably) discrete and some are continuous, and (iv) a mixture of the previous cases, which covers most real-world data [11]. This mild assumption simplifies the application of `mCMIkNN` in practice. In contrast, stronger assumptions such as lower bounds on corresponding probabilities for discrete points, cf., [3,19], or further smoothness assumptions for continuous variables, cf. [43,4], allow examining tighter bounds on type I and II error control for the finite case.

(A3): The kNN-parameter k_{CMI} can be seen as the lower bound of a locally data-adaptive "bandwidth" parameter used in the local kNN-based estimation of the Shannon entropies, see [33]. In contrast to global bandwidths of kernel-based measures, which require a careful adjustment, particularly in mixed discrete-continuous data, k_{CMI} is locally adapted within the density estimation for each sample point (see Alg. 1 and Alg. 2) providing easier calibration of the CI test. In this context, k_{CMI} can be chosen given the data characteristic[1]. In particular, higher ratios of discrete variables require smaller values of k_{CMI}. Overall, k_{CMI} can be increased for increasing n, e.g., using Runge's rule of thumb $k_{CMI} \approx 0.1n, \ldots, 0.2n$, particularly for low discrete variable ratios, cf. [33]. For more information, see the detailed evaluation results on `mCMIkNN`'s calibration in Appendix D.2.

[1] For more details on the impact of k_{CMI} and k_{perm}, we refer to the illustrative examples covering the continuous case provided by Runge [33].

Therefore, although (A3) requires $k_{CMI} \to \infty$ for $n \to \infty$, needed to receive asymptotic results, small values of k_{CMI} already suffice to approximate the densities well. In particular, the experimental results for $n \in \{50, \dots, 1\,000\}$ provided in Appendix D indicate that fixing the value to $k_{CMI} = 25$ yields well-calibrated tests while not affecting power much for the finite case.

(A4): The kNN-parameter k_{perm} is used to simulate the null distribution as local permutations are drawn within the k_{perm}-nearest distance regarding the neighborhood of $Z = z$. Therefore, k_{perm} should be chosen given the data characteristics similar to k_{perm}. For more information, see the detailed evaluation results on `mCMIkNN`'s calibration in Appendix D.2.
In this context, too large values of k_{perm} (or even a fully non-local permutation with $k_{perm} \approx n$) destroy the conditional marginal distributions under H_0, hence, increase type I errors, and too small values of k_{perm} are not sufficient to simulate H_0 given H_1 accurately, hence, increase type II errors[1].
In our experimental results for $n \in \{50, \dots, 1\,000\}$ provided in Appendix D, we find that small values of $k_{perm} \approx 5$ already suffice to simulate the null distribution reliably, as the local data-adaptiveness yields robustness.

Second, we consider the local CP parameter M_{perm} and examine `mCMIkNN`'s computational complexity in more detail.

$\mathbf{M_{perm}}$: As commonly done for permutation tests, the number of permutations M_{perm} used for the `mCMIkNN` (see Alg. 2) is chosen according to the desired nominal value α and respective requirements on the derived p-value $p_{perm,n}$. Further, note that according to Thm. 2, the power $1 - \beta$ is naturally bounded by $1 - \frac{1}{1+M_{perm}}$. For example, choosing $M_{perm} = 100$ allows for a smallest possible p-value of approx. 0.099 and bounds the power to be smaller than approx. 0.901. Hence, $M_{perm} = 100$ provides a good starting point given a nominal value $\alpha = 0.05$ and may be increased to receive more power or to examine smaller nominal values. For example, we choose $M_{perm} = 1\,000$ for all experiments with $\alpha = 0.01$, cf. CI testing in Sec. 5.2 and Sec. 5.3, and $M_{perm} = 100$ for all experiments with $\alpha = 0.05$, cf. causal discovery testing in Sec. 5.4. For more information on the choice of M_{perm}, we refer to the work of [8,30].

Computational Complexity: The main computational cost of `mCMIkNN` comes from the kNN searches in Alg. 1 and Alg. 2, which is $\mathcal{O}(n^2)$ in the worst case. To speed up the searches, `mCMIkNN` uses k-d trees, reducing the computational complexity to $\mathcal{O}(n \times log(n))$ when searching over all n samples. For a detailed evaluation of runtimes and a discussion on execution strategies, see Appendix D.5.

B `mCMIkNN`: Detailed Proofs of Robustness

In this section, we provide the detailed proofs of Thm. 1 (Appendix B.1) and Thm. 2 (Appendix B.2), which have been introduced in Appendix 3.2 *CI Test: Local Conditional Permutation Scheme*.

B.1 Proof of Theorem 1: Type I Error Control

We show that `mCMIkNN` is able to control type I error. For more information on the assumptions, see Appendix A.

Theorem 1 (Validity: Type I Error Control of $\Phi_{perm,n}$).
Let $(x_i, y_i, z_i)_{i=1}^n$ be i.i.d. samples from P_{XYZ}, and assume

(A1) *$P_{XY|Z}$ is non-singular such that $f \equiv \frac{dP_{XY|Z}}{d(P_{X|Z} \times P_{Y|Z})}$ is well-defined, and assume, for some $C > 0$, $f(x, y, z) < C$ for all $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$,*

(A2) *$\{(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} : P_{XYZ}((x, y, z)) > 0\}$ countable and nowhere dense in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$,*

(A4) *$k_{perm} = k_{perm,n} \to \infty$ and $\frac{k_{perm,n}}{n} \to 0$ as $n \to \infty$,*

then $\Phi_{perm,n}$ with p-value estimated according to Alg. 2 is able to control type I error, i.e., for any desired nominal value $\alpha \in [0, 1]$, when H_0 is true, then

$$\mathbb{E}_{P_{XYZ}}[\Phi_{perm,n}] \leq \alpha. \tag{1}$$

Proof. First, we use similar arguments as in the proof of [4, Thm. 4] to show that, under H_0, the type I error of $\Phi_{perm,n}$ (cp. Alg. 2) can be bounded in the finite case by the total variation distance of $P^n_{X|Z}$ and $\widetilde{P}^n_{X|Z}$ simulated with the local CP scheme. Note that (A1) and (A2) ensure the well-definiteness of all marginal distributions for a regular probability space throughout this proof.

Therefore, given that $P_{XY|Z}$ is non-singular (A1), regularity ensures that $P_{XY|Z} \ll P_{X|Z} \times P_{Y|Z}$ such that, under H_0, we have $P_{XYZ} \equiv P_{X|Z} \times P_{Y|Z} \times P_Z$, cf. [26, Thm. 2.2]. Further, we define the simulated product measure $\widetilde{P}_{XYZ} = \widetilde{P}_{X|Z} \times P_{Y|Z} \times P_Z$, where $P_{Y|Z}$ and P_Z are the marginals of $P_{XY|Z}$ and P_{XYZ}, respectively, and where $\widetilde{P}_{X|Z}$ is the approximated conditional probability distribution of permuted samples in Alg. 2. In this context, note that, in the finite case, the distribution of $\widetilde{P}_{X|Z}$ depends on the distribution of the observed samples $(x_i, z_i)_{i=1}^n$. We write $\widetilde{P}^n_{X|Z}$ and $P^n_{X|Z}$ to denote the samples' distribution in the finite case, respectively. In particular, let $\mathcal{S}_n$ denote the set of permutations on the indices $\{1, \ldots, n\}$ such that, for a permutation $\pi_m \in \mathcal{S}_n$ sampled according to the local CP scheme in Alg. 2, $\widetilde{P}^n_{X|Z}$ denotes the samples' conditional distribution where samples of X are permuted according to $\pi_m \in \mathcal{S}_n$, i.e., where $x^{(m)} = (x_{\pi_m(i)})_{i=1}^n$ with $(x_{\pi_m(i)}, z_i) \sim \widetilde{P}^n_{X|Z=z_i}$, $m = 1, \ldots, M_{perm}$.

Let $\tilde{x} = (\tilde{x}_i)_{i=1}^n$ be drawn from $\widetilde{P}_{X|Z}$, and let M_{perm} permutations $\tilde{x}^{(m)} = (\tilde{x}_{\pi_m(i)})_{i=1}^n$, $m = 1, \dots, M_{perm}$ be drawn from the local CP scheme of Alg. 1 sampled from $\tilde{x}$ instead of the true values in x. Now, we define

$$A_\alpha := \left\{ (x,y,z), (x^{(1)},y,z), ..., (x^{(M_{perm})},y,z) \colon \frac{1 + \sum_{m=1}^{M_{perm}} \mathbb{I}\{\hat{I}_n^{(m)} \geq \hat{I}_n\}}{1 + M_{perm}} \leq \alpha \right\},$$

where $\hat{I}_n = \hat{I}_n(x; y|z)$ and $\hat{I}_n^{(m)} = \hat{I}_n(x^{(m)}; y|z)$ i.e., the set where $p_{perm,n} \leq \alpha$. Then, by definition of A_α, we have that

$$\begin{aligned}
&\mathbb{E}_{P_{XYZ}}[\Phi_{perm,n}] \\
&\quad = \mathbb{P}_{P_{XYZ}}\Big((x,y,z), (x^{(1)},y,z), ..., (x^{(M_{perm})},y,z) \in A_\alpha\Big) \\
&\quad \leq \mathbb{P}_{P_{XYZ}}\Big((\tilde{x},y,z), (\tilde{x}^{(1)},y,z), ..., (\tilde{x}^{(M_{perm})},y,z) \in A_\alpha\Big) + \mathcal{D}_{TV}\Big(P_{XYZ}^n, \widetilde{P}_{XYZ}^n\Big),
\end{aligned}$$

with total variation distance $\mathcal{D}_{TV}\Big(P_{XYZ}^n, \widetilde{P}_{XYZ}^n\Big) = \sup_{A \in \mathcal{B}} |P_{XYZ}^n(A) - \widetilde{P}_{XYZ}^n(A)|$. Since $(\tilde{x}^{(1)},y,z), ..., (\tilde{x}^{(M_{perm})},y,z)$ are clearly i.i.d. sampled according to $\widetilde{P}_{XYZ}$, and are therefore exchangeable, by definition of A_α, we must have

$$\mathbb{P}_{P_{XYZ}}\Big((\tilde{x},y,z), (\tilde{x}^{(1)},y,z), ..., (\tilde{x}^{(M_{perm})},y,z) \in A_\alpha\Big) \leq \alpha.$$

Further, by construction of $\widetilde{P}_{XYZ}$, it holds that

$$\mathcal{D}_{TV}\Big(P_{XYZ}^n, \widetilde{P}_{XYZ}^n\Big) = \mathcal{D}_{TV}\Big(P_{X|Z}^n, \widetilde{P}_{X|Z}^n\Big).$$

Hence, $\mathbb{E}_{P_{XYZ}}[\Phi_{perm,n}] \leq \alpha + \mathcal{D}_{TV}\Big(P_{X|Z}^n, \widetilde{P}_{X|Z}^n\Big)$.

Next, we show that $\mathcal{D}_{TV}\Big(P_{X|Z}^n, \widetilde{P}_{X|Z}^n\Big)$ diminishes for $\frac{k_{perm,n}}{n} \to 0$ as $n \to \infty$. In this context, we relate the total variation distance to the Kullback-Leibler divergence using Pinsker's inequality, namely

$$\mathcal{D}_{TV}\Big(P_{X|Z}^n, \widetilde{P}_{X|Z}^n\Big) \leq \sqrt{\frac{1}{2}\mathcal{D}_{KL}\Big(P_{X|Z}^n \Big\| \widetilde{P}_{X|Z}^n\Big)}, \tag{2}$$

where $\mathcal{D}_{KL}\big(P_{X|Z}^n \big\| \widetilde{P}_{X|Z}^n\big) = \int \log(\frac{dP_{X|Z}^n}{d\widetilde{P}_{X|Z}^n}) dP_{X|Z}^n$ denotes the Kullback-Leibler divergence. Notice that, by construction, $P_{X|Z}^n \ll \widetilde{P}_{X|Z}^n$ such that the Radon-Nikodym derivative $f \equiv \frac{dP_{X|Z}^n}{d\widetilde{P}_{X|Z}^n}$ is well-defined. Notice that $\widetilde{P}_{X|Z}^n = \widetilde{P}_{X|Z=z_1}^n \times \cdots \times \widetilde{P}_{X|Z=z_n}^n$ and $P_{X|Z}^n = P_{X|Z=z_1}^n \times \cdots \times P_{X|Z=z_n}^n$ such that

$$\mathcal{D}_{KL}\Big(P_{X|Z}^n \Big\| \widetilde{P}_{X|Z}^n\Big) = \sum_{i=1}^n \mathcal{D}_{KL}\Big(P_{X|Z=z_i}^n \Big\| \widetilde{P}_{X|Z=z_i}^n\Big). \tag{3}$$

It is therefore sufficient to show that $\mathcal{D}_{KL}\left(P^n_{X|Z=z_i}\middle\|\widetilde{P}^n_{X|Z=z_i}\right)$ diminishes for one point z_i for increasing sample sizes. Therefore, for X in $\mathcal{X}$ sampled according to $P^n_{X|Z=z_i}$ or $\widetilde{P}^n_{X|Z=z_i}$, respectively, for a $r \geq 0$, we define

$$\begin{aligned} P^n_{X|Z=z_i}(x, z_i, r) &= P^n_{X|Z=z_i}\left(\{x \in \mathcal{X} : \|(x, z_i) - (x, z_i)\|_\infty \leq r\right), \text{ or} \\ \widetilde{P}^n_{X|Z=z_i}(x, z_i, r) &= \widetilde{P}^n_{X|Z=z_i}\left(\{x \in \mathcal{X} : \|(x, z_i) - (x, z_i)\|_\infty \leq r\right), \end{aligned} \tag{4}$$

respectively, which is possible due to (A2). Then, we partition $\mathcal{X} \times \mathcal{Z}$ into three disjoint sets:

1) $\Omega_1 = \{(x, z_i) \in \mathcal{X} \times \mathcal{Z} : f = 0\}$;
2) $\Omega_2 = \{(x, z_i) \in \mathcal{X} \times \mathcal{Z} : f > 0, P^n_{X|Z=z_i}(x, z_i, 0) > 0\}$;
3) $\Omega_3 = \{(x, z_i) \in \mathcal{X} \times \mathcal{Z} : f > 0, P^n_{X|Z=z_i}(x, z_i, 0) = 0\}$;

such that $\mathcal{X} \times \mathcal{Z} = \Omega_1 \cup \Omega_2 \cup \Omega_3$. Using the law of total expectation and properties of integrals, we have

$$\mathcal{D}_{KL}\left(P^n_{X|Z=z_i}\middle\|\widetilde{P}^n_{X|Z=z_i}\right) = \int \log(f(x, z_i))\ dP^n_{X|Z=z_i}(x, z_i) \tag{5}$$

$$= \int_{\Omega_1} \log(f(x, z_i))\ dP^n_{X|Z=z_i}(x, z_i) \tag{6}$$

$$+ \int_{\Omega_2} \log(f(x, z_i))\ dP^n_{X|Z=z_i}(x, z_i) \tag{7}$$

$$+ \int_{\Omega_3} \log(f(x, z_i))\ dP^n_{X|Z=z_i}(x, z_i). \tag{8}$$

Next, we consider each $\Omega_1, \Omega_2,$, and Ω_3 in three cases, respectively.

Case 1: Let $(x, z_i) \in \Omega_1$ and $\omega_X(\Omega_1) = \{(x) : (x, z_i) \in \Omega_1\}$ be the projection onto the the first coordinate of Ω_1. Using the definition of f as the Radon-Nikodym derivative, we have

$$P^n_{X|Z=z_i}(\omega_X(\Omega_1)) = \int_{\omega_X(\Omega_1)} f\ dP^n_{X|Z=z_i} = \int_{\omega_X(\Omega_1)} 0\ dP^n_{X|Z=z_i} = 0,$$

so $\int_{\Omega_1} \log(f(x, z_i))\ dP^n_{X|Z=z_i}(x, z_i) = 0$, cp. (6).

Case 2: Let $(x, z_i) \in \Omega_2$, i.e., we consider the partition of discrete points as the singletons have a positive measure in $\mathcal{X} \times \mathcal{Z}$. In this context, analogously to [26, Lem. 8], we have

$$f(x, z_i) = \frac{P^n_{X|Z=z_i}(x, z_i, 0)}{\widetilde{P}^n_{X|Z=z_i}(x, z_i, 0)}.$$

Hence, it remains to show that, for $\frac{k_{perm,n}}{n} \to 0$ as $n \to \infty$, $P^n_{X|Z=z_i}(x, z_i, 0) \equiv \widetilde{P}^n_{X|Z=z_i}(x, z_i, 0)$. Let σ_i be the distance from z_i to its k_{perm}-nearest neighbors, see Alg. 2, line 3. We proceed in the two cases $\sigma_i > 0$ and $\sigma_i = 0$.

First, for $\sigma_i > 0$, i.e., there are less than k_{perm} points in the sample equal to z_i, we show that $\mathbb{P}(\sigma_i > 0) \to 0$, as $n \to \infty$. In particular, the number of points exactly equal to z_i has a binomial distribution with parameters, $n-1$ and $P_Z(z_i)$, Binomial$(n-1, P_Z(z_i))$. Because $\frac{k_{perm,n}}{n} \to 0$ as $n \to \infty$ (A4), there must be n sufficiently large such that $\frac{k_{perm,n}-1}{n-1} \leq P_Z(z_i)$. Therefore, $\mathbb{P}(\sigma_i > 0) = \mathbb{P}(\text{Binomial}(n-1, P_Z(z_i)) \leq k_{perm,n} - 1) \to 0$ as $n \to \infty$.

Second, for $\sigma_i = 0$, there must be k_{perm} or more points exactly equal to z_i. In this context, $|\tilde{\mathbf{z}}_i|$ is the total number of points equal to z_i, see Alg. 2, line 4. Then, we draw samples according to $\widetilde{P}^n_{X|Z=z_i}$ by locally permuting only the $|\tilde{\mathbf{z}}_i|$ samples of x in (x, z) for which $z_j = z_i$, $j \neq i$, i.e., $(x_{\pi^i_m(i)})_{i=1}^n$ where π^i_m is the permutation of indices $\{j : \|(z_j) - (z_i)\|_\infty = 0, j \neq i\}$. Therefore, for all $j \in \tilde{\mathbf{z}}_i$, it holds that $\|(x_{\pi^i_m(j)}, z_i) - (x_i, z_i)\|_\infty = \|(x_j, z_i) - (x_i, z_i)\|_\infty$, cp. (4), i.e., $P^n_{X|Z=z_i}(x, z_i, 0) \equiv \widetilde{P}^n_{X|Z=z_i}(x, z_i, 0)$. Hence, the local CP scheme locally preserves the distribution of X such that, for $\frac{k_{perm,n}}{n} \to 0$ as $n \to \infty$, $P^n_{X|Z=z_i}(x, z_i, 0) \equiv \widetilde{P}^n_{X|Z=z_i}(x, z_i, 0)$ locally for $Z = z_i$. Using basic probability rules it follows that, for $n \to \infty$, $f = 1$ almost surely such that $\int_{\Omega_2} \log(f(x, z_i))\, dP^n_{X|Z=z_i}(x, z_i) \to 0$ as $n \to \infty$, cp. (7).

Case 3: Let $(x, z_i) \in \Omega_3$, i.e., we consider the continuous partition because the singletons have a zero measure in $\mathcal{X} \times \mathcal{Z}$. In this context, for $n \to \infty$, there are enough samples such that $P^n_{XZ}(\{(x, z_i) \in \Omega_3 : |\tilde{\mathbf{z}}_i| \to k_{perm}\}) = 1$, cf. [26, Lem. 5]. As (A1) and (A2) holds, analogously to [26, Lem. 7], we have that, for all $\epsilon > 0$,

$$\lim_{r \to 0} \mathbb{P}\left(\left|\frac{P^n_{X|Z=z_i}(x, z_i, r)}{\widetilde{P}^n_{X|Z=z_i}(x, z_i, r)} - f(x, z_i)\right| \leq \epsilon\right) = 1.$$

Hence, for all $\epsilon > 0$, there exists an $r_\epsilon > 0$ such that for all $r \leq r_\epsilon$ it holds that $\mathbb{P}\left(\left|\frac{P^n_{X|Z=z_i}(x,z_i,r)}{\widetilde{P}^n_{X|Z=z_i}(x,z_i,r)} - f(x, z_i)\right| \leq \epsilon\right) = 1$. Notice that $\|(x_j, z_j) - (x_i, z_i)\|_\infty \geq \|(z_j) - (z_i)\|_\infty$ for $j \neq i$. Therefore, let σ_i be the distance of z_i to its nearest neighbors such that $\|(x_{\pi^i_m(i)}, z_i) - (x_i, z_i)\|_\infty = r_\epsilon$. Then, we proceed in the two cases $\sigma_i : \|(x_{\pi^i_m(j)}, z_i) - (x_i, z_i)\|_\infty > r_\epsilon$ and $\sigma_i : \|(x_{\pi^i_m(i)}, z_i) - (x_i, z_i)\|_\infty = r_\epsilon$.

First, we consider $\sigma_{r_\epsilon,i} : \|(x_{\pi^i_m(j)}, z_i) - (x_i, z_i)\|_\infty > r_\epsilon$, i.e., that shuffling x within the distance of $\sigma_{r_\epsilon,i}$ in Z yields a distance greater than r_ϵ. Then, we show that $\mathbb{P}(\{\sigma_i : \|(x_{\pi^i_m(j)}, z_i) - (x_i, z_i)\|_\infty > r_\epsilon\}) \to 0$ as $n \to \infty$. This can only happen when $k_{perm} - 1$ or less neighbors fall within the radius of $\sigma_{r_\epsilon,i}$ such that $\sigma_{r_\epsilon,i} > \sigma_{k_{perm},i}$ where $\sigma_{k_{perm},i}$ denotes the distance of z_i to its k_{perm} nearest neighors, see Alg. 2, line 3 In this case, $|\tilde{\mathbf{z}}_i| < k_{perm}$. As there are $n-1$ i.i.d. points z_j, $j \neq i$, that can potentially fall into this region with probability $P^n_Z(z_i, \sigma_{r_\epsilon,i}) = P^n_Z(\{z_j \in \mathcal{Z} : \|(z_j) - (z_i)\|_\infty \leq \sigma_{r_\epsilon,i}\})$. Hence, this follows a binomial distribution with parameters $n-1$ and $P^n_Z(z_i, \sigma_{r_\epsilon,i})$. Because $\frac{k_{perm,n}}{n} \to 0$ as $n \to \infty$ (A4), there must be n sufficiently large such that $\frac{k_{perm,n}-1}{n-1} \leq P^n_Z(z_i, \sigma_{r_\epsilon,i})$. Therefore, $\mathbb{P}(\sigma_{k_{perm},i} > \sigma_{r_\epsilon,i}) = \mathbb{P}(\text{Binomial}(n-1, P^n_Z(z_i, \sigma_{r_\epsilon,i})) \leq k_{perm,n} - 1) \to 0$ as $n \to \infty$.

Second, for $\sigma_{r_\epsilon,i} : \|(x_{\pi^i_m(j)}, z_i) - (x_i, z_i)\|_\infty = r_\epsilon$, we have that $\sigma_{k_{perm},i} \leq \sigma_{r_\epsilon,i}$. Hence, there must be exactly k_{perm}-nearest neighbors j, $j \neq i$, of z_i, i.e., for which $\|(z_j) - (z_i)\|_\infty \leq \sigma_{k_{perm},i}, j \neq i$, holds. In this context, we draw samples according to $\widetilde{P}^n_{X|Z=z_i}$ by locally permuting only the $|\tilde{\mathbf{z}}_i|$ samples of x in (x, z) for which $\big\{j : \|(z_i) - (z_j)\|_\infty \leq \sigma_{k_{perm},i}, j \neq i\big\}$. Therefore, for all $j \in \tilde{\mathbf{z}}_i$, it holds that $\|(x_{\pi^i_m(j)}, z_i) - (x_i, z_i)\|_\infty = \|(x_j, z_i) - (x_i, z_i)\|_\infty$, cp. (4), i.e., $P^n_{X|Z=z_i}(x, z_i, 0) \equiv \widetilde{P}^n_{X|Z=z_i}(x, z_i, 0)$. Hence, the local CP scheme locally preserves the distribution of X such that, for $\frac{k_{perm,n}}{n} \to 0$ as $n \to \infty$, $P^n_{X|Z=z_i}(x, z_i, 0) \equiv \widetilde{P}^n_{X|Z=z_i}(x, z_i, 0)$ locally for $Z = z_i$. Therefore, using basic probability rules, we have that, for $n \to \infty$, $f = 1$ almost surely such that $\int_{\Omega_3} \log(f(x, z_i))\, dP^n_{X|Z=z_i}(x, z_i) \to 0$ as $n \to \infty$, cp. (8). □

Note that the second part of of the proof shows that the local CP scheme of Alg. 2 allows to asymptotically estimate $P^n_{X|Z}$, i.e., $P^n_{X|Z} \equiv \widetilde{P}^n_{X|Z}$ for $n \to \infty$.

B.2 Proof of Theorem 2: Type II Error Control

We show that `mCMIkNN` has non-trivial power, i.e., is able able to control type II error. For more information on the assumptions, see Appendix A.

Theorem 2 (Power: Type II Error Control of $\Phi_{perm,n}$).
Let $(x_i, y_i, z_i)^n_{i=1}$ be i.i.d. samples from P_{XYZ}, and assume:

- *(A1) $P_{XY|Z}$ is non-singular such that $f \equiv \frac{dP_{XY|Z}}{d(P_{X|Z} \times P_{Y|Z})}$ is well-defined, and assume, for some $C > 0$, $f(x, y, z) < C$ for all $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$;*
- *(A2) $\{(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} : P_{XYZ}((x, y, z)) > 0\}$ countable and nowhere dense in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$;*
- *(A3) $k_{CMI} = k_{CMI,n} \to \infty$ and $\frac{k_{CMI,n}}{n} \to 0$ as $n \to \infty$;*
- *(A4) $k_{perm} = k_{perm,n} \to \infty$ and $\frac{k_{perm,n}}{n} \to 0$ as $n \to \infty$,*

hold. Then $\Phi_{perm,n}$, with p-value estimated according to Alg. 2, is able to control type II error, i.e., for any desired nominal value $\beta \in \left[\frac{1}{1+M_{perm}}, 1\right]$, when H_1 is true,

$$\lim_{n\to\infty} \mathbb{E}_{P_{XYZ}}[1 - \Phi_{perm,n}] = 0. \tag{9}$$

Proof. Let $(x, y, z) = (x_i, y_i, z_i)^n_{i=1}$ be drawn from P_{XYZ}, let the M_{perm} permutations $(x^{(m)}, y, z) = (x_{\pi_m(i)}, y_i, z_i)^n_{i=1}$, $m = 1, \ldots, M_{perm}$ be drawn from $\widetilde{P}_{XYZ}$ according to the local CP scheme of Alg. 2. Then, under $H_1 : X \not\perp\!\!\!\perp Y \,|Z$, let $I(X; Y|Z) = c > 0$, such that the consistency of $\hat{I}_n(x; y|z)$ of Cor. 3. guarantees that, for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}_{P_{XYZ}}\left(\left|\hat{I}_n(x; y|z) - c\right| > \epsilon\right) = 0$. Similarly, for all $\epsilon > 0$ and $m = 1, \ldots, M_{perm}$, $\lim_{n\to\infty} \mathbb{P}_{P_{XYZ}}\left(\left|\hat{I}_n(x^{(m)}; y|z)\right| > \epsilon\right) = 0$ as $I(X^{(m)}; Y|Z) = 0$ by construction of $\widetilde{P}_{XYZ}$ as $k_{perm} = k_{perm,n} \to \infty$ for $n \to \infty$

(A4) and as $P_{X|Z} \equiv \widetilde{P}_{X|Z}$ for $\frac{k_{perm,n}}{n} \to 0$. Therefore, $(\hat{I}_n(x;y|z), I(X^{(m)};Y|Z)) \xrightarrow{P} (c,0)$, such that the continuous mapping theorem with $\phi(x,y) = |x-y|$ implies $|\hat{I}_n(x;y|z) - I(X^{(m)};Y|Z)| \xrightarrow{P} c$, for $I(X;Y|Z) = c > 0$. Now, we define

$$A_\beta := \left\{ (x,y,z), (x^{(1)},y,z), ..., (x^{(M_{perm})},y,z) \colon \frac{1 + \sum_{m=1}^{M_{perm}} \mathbb{I}\{\hat{I}_n^{(m)} \geq \hat{I}_n\}}{1 + M_{perm}} \leq \beta \right\},$$

where $\hat{I}_n = \hat{I}_n(x;y|z)$ and $\hat{I}_n^{(m)} = \hat{I}_n(x^{(m)};y|z)$ i.e., the set where $\Phi_{perm,n} = 1$. Then, by definition of A_β, we have that

$$\mathbb{E}_{P_{XYZ}}[1 - \Phi_{perm,n}] = 1 - \mathbb{P}_{P_{XYZ}}\left((x,y,z), (x^{(1)},y,z), ..., (x^{(M_{perm})},y,z) \in A_\beta\right).$$

As $|\hat{I}_n(x;y|z) - I(x^{(m)};y|y)| \xrightarrow{P} c$ for $(x,y,z), (x^{(1)},y,z), ..., (x^{(M_{perm})},y,z) \in A_\beta$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left\{ \frac{1 + \sum_{m=1}^{M_{perm}} \mathbb{I}\{\hat{I}_n^{(m)} \geq \hat{I}_n\}}{1 + M_{perm}} \leq \beta \right\}\right) = \mathbb{P}\left(\left\{ \frac{1}{1 + M_{perm}} \leq \beta \right\}\right).$$

This completes the proof, as we can conclude that $\lim_{n\to\infty} \mathbb{E}_{P_{XYZ}}[1 - \Phi_{perm,n}] = 1 - 1 = 0$ for all $\beta \in \left[\frac{1}{1+M_{perm}}, 1\right]$. □

Hence, note that the power $1-\beta$ of `mCMIkNN` is, as common for permutation tests, naturally bounded according to the number of permutations M_{perm}, i.e., $1 - \beta \leq 1 - \frac{1}{1+M_{perm}}$. Therefore, increasing M_{perm} yields more power but comes along with a longer run-time.

Note, although Thm. 2 shows that `mCMIkNN` is asymptotically able to control type II errors, the dimensionality-biasedness of $\hat{I}_n(x;y,|z)$ for $d_Z \to \infty$ affects the robustness in finite sample sizes. In particular, for finite n, $\hat{I}_n(x;y,|z)$ converges in probability towards zero for $d_Z \to \infty$, hence, increasing type II errors, cp. A_β. In this context, the extensive synthetic evaluation in Appendix D indicates that `mCMIkNN` is robust regarding type II errors in the finite case, too.

C mCMIkNN-based Constraint-based Causal Discovery

In this section, we provide more information on constraint-based causal discovery (Appendix C.1) and a detailed proof of Thm. 3 (Appendix C.2) introduced in Sec. 3.3 *mCMIkNN-based Constraint-based Causal Discovery.*

C.1 Constraint-Based Causal Discovery

In the framework of causal discovery, the causal structures between N random variables $\mathbf{V} = \{X, Y, \dots\}$ are described as a *directed acyclic graph* (DAG) $\mathcal{G}$, where an edge $X \to Y$ depicts a direct causal mechanism between the two respective variables X and Y, for $X, Y \in \mathbf{V}$ [28]. If $\mathbf{V}$ entails all relevant variables, i.e., there are no latent variables, then the set $\mathbf{V}$ is said to be *causally sufficient* [34]. In this context, the *causal Markov* condition states a coincidence between the conditional independence structure of the variable's joint distribution $P_{\mathbf{V}}$ through a graphical condition on the structure of $\mathcal{G}$. In particular, the so-called *d-separation* criterion states that two variables $X, Y \in \mathbf{V}$ are conditionally independent given a set of variables $Z \subseteq \mathbf{V} \setminus \{X, Y\}$, denoted by $X \perp\!\!\!\perp Y \mid Z$, if and only if all paths between X and Y are blocked by Z in the corresponding DAG $\mathcal{G}$, see [28]. Further, the assumption of *causal faithfulness* implies that any CI characteristic of $P_{\mathbb{V}}$ is required to be entailed in $\mathcal{G}$ [34].

Methods for causal discovery build upon this coincidence between the causal structures of $\mathcal{G}$ and the conditional independence (CI) characteristics of the joint distribution $P_{\mathbf{V}}$, cf. [34]. In particular, constraint-based methods for causal discovery, such as the well-known PC algorithm, apply CI tests to derive as many underlying causal structures in $\mathcal{G}$ from n i.i.d. observations $(x_i, y_i, \dots)_{i=1}^{n}$ sampled from $P_{\mathbb{V}}$ as possible, e.g., see [7]. For more information on causal discovery in general and recent advances, we refer to [12,29,36,42].

C.2 Proof of Theorem 3: Consistent Causal Discovery

Theorem 3 (Consistency of mCMIkNN-based Causal Discovery). *Let $\mathbf{V}$ be a finite set variables with joint distribution $P_{\mathbf{V}}$ and assume (A1) - (A4) hold. Further, assume the general assumptions of the PC algorithm hold, i.e., causal faithfulness and causal Markov condition, see [34]. Let $\hat{\mathcal{G}}_{CPDAG,n}(\alpha_n)$ be the estimated CPDAG of the PC algorithm and $\mathcal{G}_{CPDAG}$ the true CPDAG from $\mathcal{G}$. Then, for $\alpha_n = \frac{1}{1+M_{perm,n}}$ with $M_{perm,n} \to \infty$ as $n \to \infty$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\mathcal{G}}_{CPDAG,n}(\alpha_n) = \mathcal{G}_{CPDAG}\right) = 1. \tag{10}$$

Proof. The idea of the proof is inspired by Kalisch et al. [17] and considers wrongly detected edges due to incorrect CI decisions of mCMIkNN. In contrast, we show that the errors due to incorrect CI decisions can be controlled asymptotically by choosing $\alpha_n = \frac{1}{1+M_{perm,n}}$.

In the adjacency search, the first part of the PC algorithm, an error occurs if, for nodes X and Y and conditioning set Z, an error event $E_{X,Y|Z}$ occurs. Thus,

$$\begin{aligned}\mathbb{P}(\text{error occurs in the first part of PC}) &\leq \mathbb{P}\left(\bigcup_{X,Y,Z} E_{X,Y|Z}\right)\\ &\leq \sum_{X,Y,Z} \mathbb{P}(E_{X,Y|Z} \text{ occurs})\\ &\leq N^{N-2} \sup_{X,Y,Z} \mathbb{P}(E_{X,Y|Z} \text{ occurs}),\end{aligned}$$

as the number of combinations of X, Y and Z in $\mathbf{V}$ is bounded by N^{N-2}. Now, we split error events into type I and II errors, i.e., $E_{X,Y|Z} = E^I_{X,Y|Z} \cup E^{II}_{X,Y|Z}$,

$$\begin{aligned}\text{type I error } E^I_{X,Y|Z}: &\quad p_{perm,n} \leq \alpha_n, \text{ and } X \perp\!\!\!\perp Y \mid Z;\\ \text{type II error } E^{II}_{X,Y|Z}: &\quad p_{perm,n} > \alpha_n, \text{ and } X \not\perp\!\!\!\perp Y \mid Z.\end{aligned}$$

Then, the statistical validity of `mCMIkNN` according to Thm. 1 ensures that, for any $\alpha_n \in [0,1]$, we have that $\mathbb{P}(E^I_{X,Y|Z} \text{ occurs}) \leq \alpha_n$ for $n \to \infty$. Further, the power of `mCMIkNN` according to Thm. 2 ensures, that for any $\alpha_n \in [\frac{1}{1+M_{perm}}, 1]$, $\mathbb{P}(E^{II}_{X,Y|Z} \text{ occurs}) = 0$ for $n \to \infty$. Hence, choosing $\alpha_n = \frac{1}{1+M_{perm}}$ with $M_{perm} = M_{perm,n} \to \infty$ as $n \to \infty$ we have that,

$$\begin{aligned}\mathbb{P}(\text{error occurs in the first part of PC}) &\leq N^{N-2} \sup_{X,Y,Z} P(E_{X,Y|Z} \text{ occurs})\\ &\leq N^{N-2} \sup_{X,Y,Z} \Big(P(E^I_{X,Y|Z} \text{ occurs}) + P(E^I_{X,Y|Z} \text{ occurs})\Big)\\ &\leq N^{N-2} \frac{1}{1+M_{perm,n}}\\ &= 0, \text{ as } n \to \infty.\end{aligned}$$

Therefore, the undirected skeleton of $\mathcal{G}$ and separation sets are correctly estimated for $n \to \infty$, which proves Thm. 3, as the edge orientation (second part) of the PC algorithm will never fail, see [25]. □

Hence, using `mCMIkNN` for constraint-based causal discovery allows consistently estimating the CPDAG $\mathcal{G}_{CPDAG}$ of the true underlying DAG $\mathcal{G}_{CPDAG}$ for $n \to \infty$.

D Synthetic Empirical Evaluation

In this section, we provide a more detailed description of the synthetic evaluation of Sec. 5 *Empirical Evaluation*. In particular, we describe the assumed MANM and its parameters used for synthetic data generation (Appendix D.1). Further, we include additional measurement results and a more detailed analysis regarding the calibration of `mCMIkNN` (Appendix D.2), robustness of `mCMIkNN` (Appendix D.3), and the accuracy of CI decision in comparison to state-of-the-art competitors (Appendix D.4). Morevover, we compare the CI tests' runtimes and sketch parallel execution strategies to speed up `mCMIkNN` (Appendix D.5). Finally, we provide additional results in the context of constraint-based causal discovery (Appendix D.6).

D.1 Synthetic Data Generation

As recommended by Huegle et al. [15], we consider the mixed additive noise (MANM) model for evaluating approaches for CI testing and constraint-based causal discovery from mixed discrete-continuous data. In particular, for all $X \in \mathbf{V}$, let X be generated from its J discrete parents $\mathcal{P}^{dis}(X) \subseteq \mathbf{V} \setminus X$, where $J := \#\mathcal{P}^{dis}(X)$, its K continuous parents $\mathcal{P}^{con}(X) \subseteq \mathbf{V} \setminus X$, where $K := \#\mathcal{P}^{con}(X)$, and an independent noise term N_X according to

$$X = \frac{1}{J} \sum_{j=1,\dots,J} f_j(Z_j) + \left(\sum_{k=1,\dots,K} f_k(Z_k) \right) \bmod d_X + N_X. \tag{11}$$

We restrict our attention to the cyclic model, where the domain of a discrete variable is the modulo ring $\mathbb{Z}/d_X\mathbb{Z}$ to restrict the support of discrete variables, i.e., X can take values from $\{0, \dots, d_X - 1\}$. Moreover, the independent noise variable N_X either is a continuous distributed random variable, i.e., $N_X \sim \mathcal{N}(0, 1)$, or discrete distributed over $\mathbb{Z}/d_X\mathbb{Z}$ with $\mathbf{P}(N_X = 0) \geq \mathbf{P}(N_X = l)$ for all $l \in \{0, \dots, d_X - 1\}$ if X is continuous or discrete, respectively. Therefore, $f_j : \mathbb{Z}/d_j\mathbb{Z} \to \mathbb{Z}/d_X\mathbb{Z}$ and $f_k : \mathbb{R} \to \mathbb{Z}/d_X\mathbb{Z}$ if X is discrete, or $f_j : \mathbb{Z}/d_j\mathbb{Z} \to \mathbb{R}$ and $f_k : \mathbb{R} \to \mathbb{R}$ if X is continuous. In particular, functions $f_j : \mathbb{R} \to \mathbb{Z}/d_X\mathbb{Z}$ assign a probability to each realization within the support $\{0, \dots, d_X - 1\}$ of X using a softmax function while $f_k : \mathbb{R} \to \mathbb{R}$ is a continuous function. Note that we scale the parents' signals, see (11), to reduce the noise for subsequent variables avoiding high varsortability [31], and max-min normalize all continuous variables. For more information on the MANM, we refer to [15].

Table D.1 describes the parameters and their values used to generate synthetic data with the `MANM-CS` library. For the first three experiments, i.e., Calibration (Appendix D.2) Robustness (Appendix D.3), and CI testing (Appendix D.4), a CGM is generated according to $X \perp\!\!\!\perp Y | Z$ or $X \not\perp\!\!\!\perp Y | Z$, and the MANM is sampled according to the parameters at the top. For the experimental evaluation of causal discovery (Appendix D.6), the parameters at the bottom are used to generate the structure of the CGM, too.

Table D.1. Parameters of `MANM-CS` used for synthetic data generation.

Parameter Description	Values
ratio of discrete variables	$\{0.0, 0.25, 0.5, 0.75, 1.0\}$
range for discrete classes	$\{2, 3, 4\}$
discrete signal to noise ratio	$\{0.85\}$
continuous functions with sample probabilities	$\{(\frac{1}{3}, id(\cdot)), (\frac{1}{3}, (\cdot)^2), (\frac{1}{3}, \cos(\cdot))\}$
standard deviation of continuous Gaussian noise	$\{1.0\}$
scale parents	$\{1\}$
number of samples	$\{50, 100, 250, 500, 1\,000\}$
variables scaling	{`normal`}
number of variables N	$\{10, 20, 30\}$
edge density of the CGMs	$\{0.1, 0.2, 0.3, 0.4\}$

D.2 Calibration

We start with a detailed evaluation of `mCMIkNN`'s parameters to provide recommendations for calibration (Sec. 5.2). In this context, we restrict our attention to two simple CGMs $\mathcal{G}$ with variables $\mathbf{V} = \{X, Y, Z_1, ..., Z_{d_Z}\}$, where first, X and Y have common parents $Z = \{Z_1, ..., Z_{d_Z}\}$ in $\mathcal{G}$, i.e., $H_0 : X \perp\!\!\!\perp Y | \boldsymbol{Z}$, and second, there exists an additional edge connecting X and Y in $\mathcal{G}$, i.e., $H_1 : X \not\perp\!\!\!\perp Y | Z$.

Table D.2. ROC AUC scores (higher better) for different combinations of k_{CMI}, k_{perm}, and samples n with fixed $M_{perm} = 100$ and $\alpha = 0.05$ derived from CI decisions over multiple settings, e.g., sampled with a varying dimension of Z, $d_Z \in \{1, 3, 5, 7\}$, continuous functions, or discrete variable ratios (see Table D.1).

samples n	k_{CMI} \ k_{perm}	5	25	100	200
50	5	**0.58**	**0.58**	-	-
	25	0.56	0.55	-	-
100	5	**0.64**	**0.64**	-	-
	25	**0.64**	**0.64**	-	-
250	5	0.72	0.72	0.72	0.72
	25	**0.73**	**0.73**	**0.73**	**0.73**
	100	0.66	0.65	0.64	0.64
	200	0.55	0.54	0.53	0.53
500	5	**0.77**	**0.77**	**0.77**	**0.77**
	25	**0.77**	0.76	0.76	0.76
	100	0.73	0.71	0.71	0.7
	200	0.67	0.66	0.65	0.65
1000	5	0.8	0.8	0.8	0.8
	25	**0.81**	0.8	0.8	0.8
	100	0.77	0.75	0.74	0.74
	200	0.73	0.71	0.7	0.69

Table D.3. Type I (top) and type II (bottom) error rates (smaller better) for different combinations of k_{CMI}, k_{perm}, and samples n with fixed $M_{perm} = 100$ derived from CI decisions ($\alpha = 0.05$) over multiple settings, e.g., sampled with a varying dimension of Z, $d_Z \in \{1, 3, 5, 7\}$, continuous functions, or discrete variable ratios (see Table D.1).

Type I Error Rates					
samples n	k_{CMI} \ k_{perm}	5	25	100	200
50	5	0.06	0.06	-	-
	25	**0.04**	0.06	-	-
100	5	**0.05**	0.06	-	-
	25	**0.05**	0.06	-	-
250	5	0.07	0.07	0.07	0.07
	25	0.07	0.08	0.09	0.09
	100	0.06	0.08	0.09	0.09
	200	**0.04**	0.07	0.09	0.09
500	5	**0.07**	**0.07**	**0.07**	**0.07**
	25	0.08	0.11	0.11	0.12
	100	0.08	0.11	0.12	0.13
	200	**0.07**	0.1	0.12	0.11
1000	5	**0.08**	**0.08**	0.09	**0.08**
	25	0.12	0.15	0.15	0.15
	100	0.12	0.17	0.18	0.19
	200	0.1	0.15	0.18	0.18

Type II Error Rates					
samples n	k_{CMI} \ k_{perm}	5	25	100	200
50	5	**0.78**	**0.78**	-	-
	25	0.84	0.84	-	-
100	5	**0.66**	**0.66**	-	-
	25	0.67	**0.66**	-	-
250	5	0.49	0.49	0.5	0.49
	25	0.47	**0.46**	**0.46**	**0.46**
	100	0.63	0.62	0.62	0.62
	200	0.86	0.85	0.85	0.85
500	5	0.39	0.38	0.38	0.38
	25	**0.37**	**0.37**	**0.37**	**0.37**
	100	0.47	0.46	0.46	0.46
	200	0.58	0.58	0.58	0.58
1000	5	0.31	0.31	0.31	0.31
	25	**0.26**	**0.26**	**0.26**	**0.26**
	100	0.34	0.34	0.34	0.34
	200	0.43	0.43	0.44	0.43

To get a balanced view on type I and type II errors, we compare the area under the receiver operating curve (ROC AUC) given varying parameters k_{CMI} and k_{perm}. In Table D.2, we present a detailed comparison of the ROC AUCs for different combinations of $k_{CMI} \in \{5, 25, 100, 200\}$ and $k_{perm} \in \{5, 25, 100, 200\}$ with sample sizes ranging from 50 to 1 000. Note that consider $\alpha = 0.01$ and set $M_{perm} = 100$, which provides a good starting point (cf. Appendix A). Furthermore, Table D.3 presents the type I and type II error measures for the same set of CI decisions used in Table D.2.

For k_{CMI}, the detailed evaluation show that small values of k_{CMI}, e.g., $k_{CMI} \leq 25$, are sufficient to estimate the true CMI value achieving appropriate accuracy (cf. Table D.3), i.e., yield higher ROC AUCs for all sample sizes n and k_{perm} (cf. Table D.2). This is in line with the results of [26]. Note, that for a significantly large number of samples, a common rule-of-thump of $k_{CMI} \approx 0.1n, \ldots, 0.2n$, e.g., see [33]. As the runtime of `mCMIkNN` increases log-linear with k_{CMI}, fixing k_{CMI} to small values keeps the experimental evaluation practical without affecting the power much.

For k_{perm}, the ROC AUCs marginally decrease with larger values of k_{perm} (cf. Table D.2), such that small values already suffice to simulate the null distribution reliably (cf. Table D.3), which is in line with results from [33]. Hence, our experimental evaluation indicates that the power of `mCMIkNN` is relatively robust regarding the choice of k_{perm} (of course, as long as $k_{perm} < n$). In this context, note that the runtime is only marginally affected by k_{perm}.

Hence, the experimental results indicate that fixing the values to $k_{CMI} = 25$ and $k_{perm} = 5$ yields well-calibrated CI tests while not affecting accuracy much for the finite case.

D.3 Robustness of `mCMIkNN`: Type I and II Error Control

Further, evaluate `mCMIkNN`'s robustness regarding validity and power in the finite case by examining the type I and II error rates as depicted in Fig. D.1 (cp. Fig. 1 in Sec. 5.2).

We see that `mCMIkNN` is able to control type I errors for all discrete variable ratios dvr (Fig. D.1 vertical) and sizes of the conditioning set d_Z (Fig. D.1 horizontal). Moreover, for an increasing number of samples n the type II error rates decrease (Fig. D.1 in each subplot), hence, `mCMIkNN` achieves non-trivial power, particularly for small sizes of the conditioning sets d_Z. In this context, higher type II errors in the case of higher dimensions d_Z point out that `mCMIkNN` suffers from the curse of dimensionality, cp. the dimonsionality-biasedness of $\hat{I}_n(X; Y|Z)$ for increasing d_Z as shown in Cor. 3.

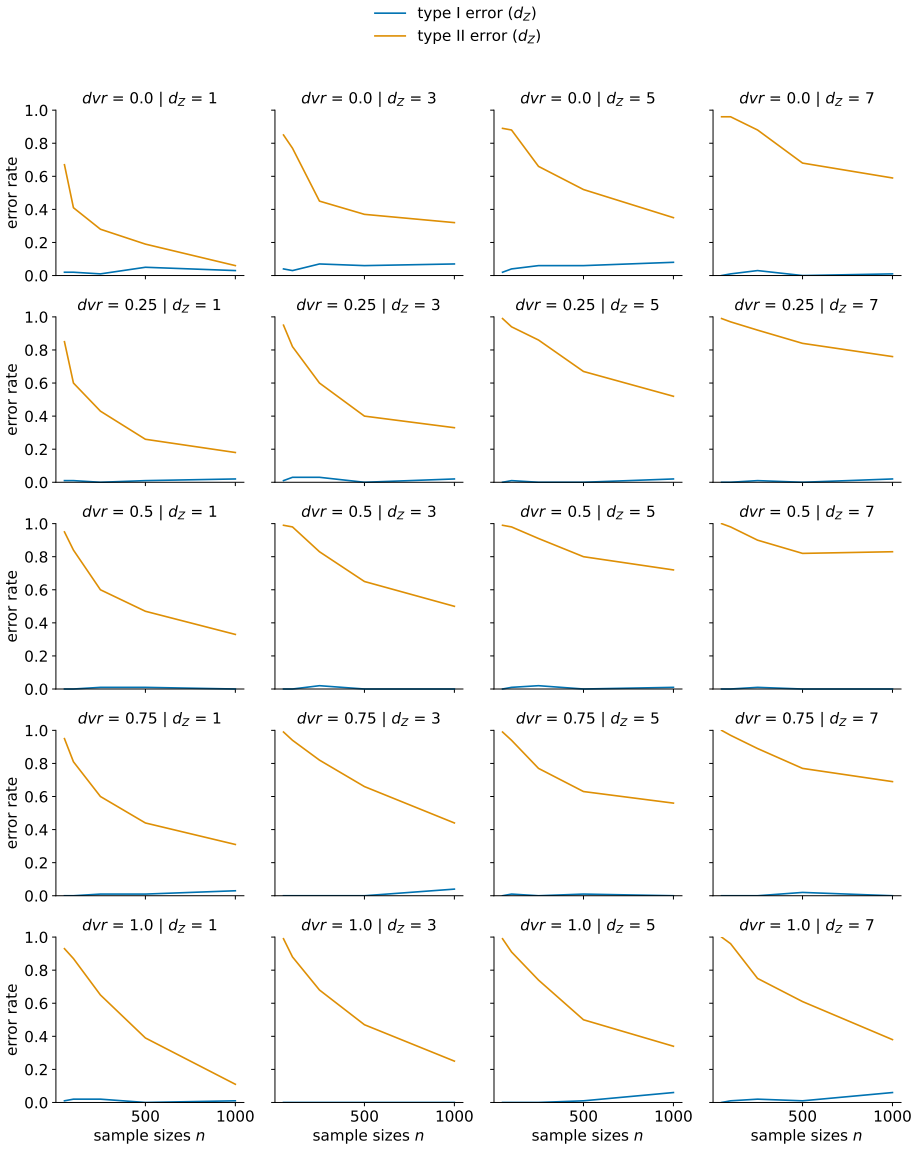

Fig. D.1. Type I and II error rates of `mCMIkNN` (smaller better) given varying sample sizes n, each subplot illustrates one combination of a dimension of Z, i.e., $d_Z \in \{1, 3, 5, 7\}$, and a distinct discrete variable ratio $dvr \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$.

D.4 Conditional Independence Testing

According to Sec. 5.3, we examine `mCMIkNN`'s empirical performance compared to state-of-the-art competitors in more detail. In particular, we chose the following CI tests and implementations:

- `mCMIkNN` our kNN-based CI using CMI and the local permutation scheme;
- `CG` a likelihood ratio test assuming conditional Gaussianity [2] implemented as function `mixCItest` in the R package `micd` [9];
- `discχ²` a discretization-based approach, where we discretize continuous variables using `Ckmeans.1d.dp` from the same-named R package [39] (using BIC to choose an appropriate number of clusters $k \in \{4, ..., 9\}$), before applying Pearson's χ^2 test from the R package `pcalg` [18];
- `aHistχ²` a non-parametric CI test, where CMI is estimated based upon adaptive histograms [24] and a CI test is derived via a pseudo-p-value using a χ^2 correction coded for $\alpha = 0.01$ (cf. `CMIp.Chisq95` [24])[2];
- `KCIT` the well-known kernel-based CI test from Zhang et al. [41].

In this experiment, we again consider the two CGMs used for the calibration in Appendix D.2 and examine the respective type I and type II errors associated with the ROC AUC scores in Fig. 2 from 20 000 CI decisions ($\alpha = 0.01$).

We compare the CI test's type I errors (Fig. D.2) and type II errors (Fig. D.3) concerning various sample sizes (left), different sizes of conditioning sets d_Z (center), and different ratios of discrete variables (right).

To achieve statistical validity, type I error rates should be close to the required nominal value $\alpha = 0.01$, see Thm. 1. As depicted in Fig. D.2, statistical validity for all settings can only achieved by `mCMIkNN`, while all other CI tests struggle with its well-known weaknesses regarding the curse of dimensionality or inadequate assumptions.

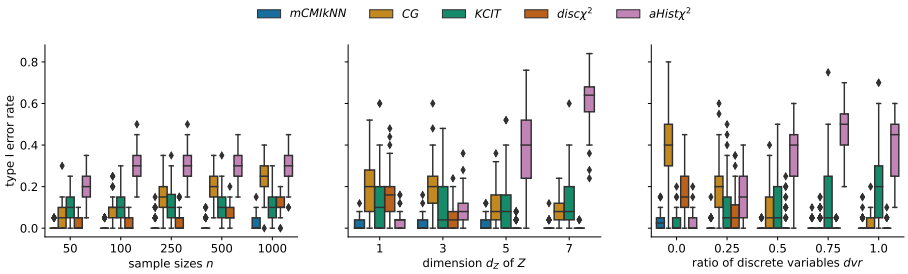

Fig. D.2. Type I error rate (smaller better) of 20 000 CI decisions of the CI tests `mCMIkNN`, `CG`, `KCIT`, `discχ²`, and `aHistχ²` with varying sample sizes n (left), dimensions of the conditioning sets d_Z (center), and ratios of discrete variables (right)[2].

[2] Note that runs of `aHistχ²` are limited to an execution time of 10 minutes and approx. 4 900 out of 20 000 runs for the CI experiment did not complete in time. Therefore, its

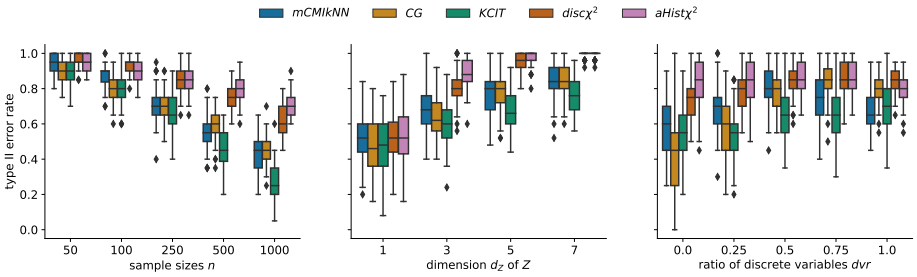

Fig. D.3. Type II error rate (smaller better) of 20 000 CI decisions of the CI tests `mCMIkNN`, `CG`, `KCIT`, `disc`χ^2, and `aHist`χ^2 with varying sample sizes n (left), dimensions of the conditioning sets d_Z (center), and ratios of discrete variables dvr (right)[2].

For example, `aHist`χ^2 suffers strongly from the curse of dimensionality (Fig. D.2 center), which yields weaknesses in type I error control when examining the aggregated view for increasing number of samples (Fig. D.2 left). Further, for a low discrete variable ratio (Fig. D.2 right), `CG` has high type I error rates as the linearity assumption of the conditional Gaussianity is not fulfilled in the continuous case. Similarly, for low discrete variable ratios(Fig. D.2 right), type I error rates of `KCIT` are low as kernel-based methods demonstrate their strength in nonlinear continuous data but increase for increasing ratios of discrete variables.

Regarding type II errors (Fig. D.3), the results are in line with the ROC AUC scores of Fig. 2. The type II error rates of all CI tests decrease as n grows (Fig. D.3 left) with the well-known weaknesses regarding the curse of dimensionality (Fig. D.3 center) and inadequate assumptions (Fig. D.3 right).

Concerning an increasing size of the conditioning sets d_Z (Fig. D.3 center), we observe that all methods suffer from the curse of dimensionality, while `KCIT`, directly followed by `mCMIkNN`, is able to control type II errors for higher dimensions of conditioning sets d_Z. In this context, `aHist`χ^2 and `disc`χ^2 suffer strongly from the curse of dimensionality as adaptive histogram-based and discretization-based approaches require much more sample sizes to achieve an appropriate power (cp. Fig. D.3 left and center).

For varying ratios of discrete variables dvr (Fig. D.3 right), we observe that `mCMIkNN` and `KCIT` achieve stable and low type II errors for all dvr. In this context, for the restricted number of samples $n \leq 1\,000$, `disc`χ^2 and `aHist`χ^2 suffer from the curse of dimensionality, which yields high type II error rates. For the continuous case, the linearity assumption of `CG` approximately covers some dependencies such that it achieves lower type II error rates. In contrast, for the discrete case, `CG` suffers from the combination of high degrees of freedom in combination with low sample sizes (similar to `disc`χ^2 and `aHist`χ^2), in particular for high d_Z which yields high type II error rates.

long execution time impedes usage in constraint-based causal discovery and `aHist`χ^2 is excluded in the respective experiment.

D.5 Runtime Comparison

Lastly, we compare the mean runtimes of the different methods for CI testing over 2 400 CI decisions. In this context, we restrict the runtime measurements to the execution of the CI tests, i.e., omitting any data preparation, such as discretization in the case of `discχ²`. Furthermore, we performed each CI test single-threaded on a server system with an Intel® Xeon® E7-4850 v4 CPU. Moreover, due to the long runtime of `aHistχ²` we limit the execution time to 600 seconds. In particular, we compare the runtimes for varying sample sizes n, dimensions of the conditioning sets d_Z, and discrete variables ratios dvr in Tables D.4, D.5, and D.6, respectively.

Table D.4. Mean runtimes in seconds of 2 400 CI decisions of `mCMIkNN` ($M_{perm} = 100$), `CG`, `KCIT`, `discχ²`, and `aHistχ²` for an increasing number of samples n covering different sizes of conditioning sets $d_Z \in \{1, 3, 5, 7\}$ and discrete variables ratios $dvr \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$.

Method \ n	50	100	250	500	1000	2000
$mCMIkNN$	0.411	0.914	2.929	7.713	20.995	58.109
CG	0.029	0.044	0.077	0.115	0.173	0.274
$KCIT$	0.014	0.035	0.255	1.648	12.114	102.012
$disc\chi^2$	0.001	0.001	0.001	0.001	0.001	0.001
$aHist\chi^2$	119.327	165.654	179.724	180.526	178.684	201.159

Examining runtimes of the CI tests for increasing sample sizes (see Table D.4) shows expected behavior according to the methods' general computational complexity. For example, `CG` and `discχ²` achieve the fastest runtimes with fractions of seconds, even for a high number of samples[3]. The adaptive histogram-based `aHistχ²` suffers the longest runtimes for all considered number of samples, which is also due to the curse of dimensionality yielding a worse performance for high-dimensional conditioning sets, cf. Table D.5. While `KCIT` achieves a fast execution for small sample sizes, its cubic complexity yields long runtimes for high sample sizes. In contrast, `mCMIkNN`'s log-linear complexity, hence, achieves reasonable performance for small sample sizes and outperforms `KCIT` for higher sample sizes $n \geq 2000$.

Regarding an increase of the dimensionality of the conditioning set d_Z (see Table D.5), the runtimes indicate that `aHistχ²` struggles strongly in high-dimensional settings. Similarly, but to a lesser extent, the runtimes of `CG` and `mCMIkNN` increase when increasing the size of the conditioning sets. In contrast, `discχ²` and `KCIT` achieve stable runtimes for all dimensions[3].

[3] Note, that `discχ²` requires the discretization of continuous variables which is excluded in the runtime measurements.

Table D.5. Mean runtimes in seconds of 2 400 CI decisions of `mCMIkNN` ($M_{perm} = 100$), `CG`, `KCIT`, `disc`χ^2, and `aHist`χ^2 for increasing sizes conditioning sets d_Z covering different sample sizes $n \in \{50, 100, 250, 500, 1000, 2000\}$ and discrete variables ratios $dvr \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$.

Method \ d_Z	1	3	5	7
$mCMIkNN$	7.849	9.887	17.474	25.505
CG	0.006	0.032	0.144	0.292
$KCIT$	19.425	19.249	19.313	19.397
$disc\chi^2$	0.001	0.001	0.001	0.002
$aHist\chi^2$	2.214	62.267	307.724	311.178

Table D.6. Mean runtimes in seconds of 2 400 CI decisions of `mCMIkNN` ($M_{perm} = 100$), `CG`, `KCIT`, `disc`χ^2, and `aHist`χ^2 for increasing discrete variables ratios dvr covering different sample sizes $n \in \{50, 100, 250, 500, 1000, 2000\}$ and sizes of conditioning sets $d_Z \in \{1, 3, 5, 7\}$.

Method \ dvr	0.0	0.25	0.5	0.75	1.0
$mCMIkNN$	12.554	11.677	13.067	17.938	20.656
CG	0.002	0.02	0.091	0.291	0.189
$KCIT$	19.615	19.33	19.257	19.379	19.151
$disc\chi^2$	0.001	0.001	0.001	0.001	0.002
$aHist\chi^2$	21.506	138.145	181.078	328.064	185.435

For varying discrete variables ratios dvr (see Table D.6), `disc`χ^2 and `KCIT` achieve stable runtimes for all settings, too[3]. In contrast, the runtimes of `mCMIkNN` and `CG` increase for increasing discrete variables ratios and `aHist`χ^2 struggles in mixed cases with $dvr \in \{0.25, 0.5, 0.75\}$. In the case of `mCMIkNN`, the poorer performance for higher discrete variable ratios dvr is caused by the use of k-d trees when computing the kNN, as k-d trees are less efficient for discrete data, compared to continuous data.

Note that the permutation scheme of `mCMIkNN` is well suited for parallel execution strategies to speed up the computation. In particular, the computational expensive M_{perm} permutations in Alg. 2 can be embarrassingly parallelized, e.g., see [33]. Further, recent research on hardware acceleration has shown that kNN-based CI tests can be efficiently executed on GPUs, particularly when used in constraint-based causal discovery [14].

D.6 Causal Discovery

In the following, we provide more details on the consistency evaluation of causal discovery using the PC-stable algorithm, cf. Sec. 5.3. Therefore, we examine

wrongly detected edges in the skeletons of $\hat{\mathcal{G}}_{CPDAG,n}(0.05)$ estimated with PC-stable using the respective CI tests in comparison to the true skeleton of $\mathcal{G}$, cf. [6]. In this context, we choose $\alpha = 0.05$, which takes the more complex CI characteristics present in causal discovery (e.g., confounders, colliders, paths) into account, such that $M_{perm} = 100$ is sufficient for the chosen nominal value.

Note that we excluded `aHist`χ^2 in the evaluation of causal discovery as its long execution time and restricted implementation to non-discrete data do not allow for usage in constraint-based causal discovery.

In Fig. 3 of the paper, we evaluated the F1 score regarding varying ratios of discrete variables, densities of the DAGs, and different numbers of variables N. To provide a complete examination, we also present the ROC AUC scores of wrongly detected edges in the skeletons of $\hat{\mathcal{G}}_{CPDAG,n}(0.05)$ estimated with PC-stable using the respective CI tests in comparison to the true skeleton of $\mathcal{G}$. In this context, examining type I and type II errors more balanced using ROC AUC scores (Fig. D.4) shows no noteworthy differences.

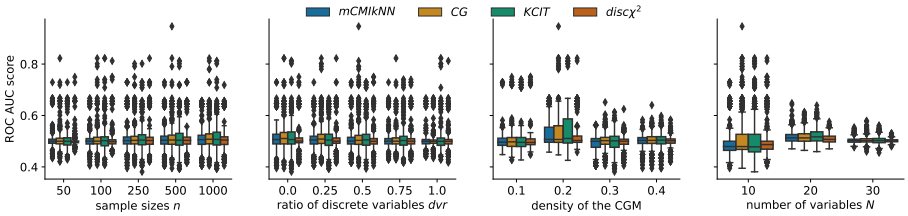

Fig. D.4. ROC AUC scores (higher better) of CPDAGs estimated with PC-stable using CI tests `mCMIkNN`, `CG`, `KCIT`, and `discχ`^{2} computed over 3 000 CGMs for varying sample sizes n, discrete variable ratios dvr, densities of CGMs, and numbers of variables N (left to right)[2].

E Real-World Evaluation

In this section, we provide detailed information on the real-world industrial manufacturing scenario sketched in Sec. 5.5 *Real-World Scenario: Discrete Manufacturing*. In particular, we motivate the use-case and describe challenges in practice (Appendix E.1). Further, we describe the simplified manufacturing scenario and an empirical evaluation of `mCMIkNN`-based causal discovery (Appendix E.2). We complete our real-world scenario with concluding remarks and point out limitations (Appendix E.3).

E.1 Introduction to Causal Discovery in Discrete Manufacturing

Motivation: Modern discrete manufacturing enterprises are faced with growing demands for increased product quality, diversified products that collide with shortened product life-cycles, reduced costs, and global competition [21]. In this context, production quality performance during the machinery production process is of fundamental relevance [5,38]. Therefore, enhancing productivity and effective quality improvement can be instrumental in increasing an enterprise's competitive power [27]. Moreover, the machinery's configurations have a profound impact on the performance of the manufacturing system in terms of productivity, product quality, capacity, scalability, and costs [20]. Therefore, understanding causal structures between machinery configurations, product characteristics, and the respective product quality is essential for enhancing the productivity of the manufacturing process [1]. However, understanding causal structures in the industrial domain usually relies on domain knowledge and intuition as industrial manufacturing processes become more complex, with sometimes hundreds of possible factors [23]. In this context, the emergence of methods for causal discovery creates the basis for attempts of a data-driven assessment of the causal structures from observational data of manufacturing processes, e.g., see [13,16,23].

Background: A typical machinery production process can be schematically divided into the *configuration phase* and the *production phase*, cf. Figure E.1.

While the production process is highly automated, the configuration phase requires substantial human participation as the machinery technician has to configure the machinery settings. In particular, the machinery has to be properly configured to ensure a minimum of rejected goods that do not meet required quality targets. Therefore, the technician aims to start the machinery production process in the configuration phase to obtain the quality targets through an iterative adaption of possible machinery configuration settings such as speed. In this context, all products that do not meet the quality targets are rejected. In case quality standards are met, machinery configurations S_{con}, achieved quality results Q_{con}, and the number of rejected goods R_{con} within the configuration phase are logged. Then, the discrete production process enters the production phase with a high throughput of produced products over several units of similar design using the derived machinery configurations. Finally, the number of rejected goods within the production phase R_{prod} is logged, too.

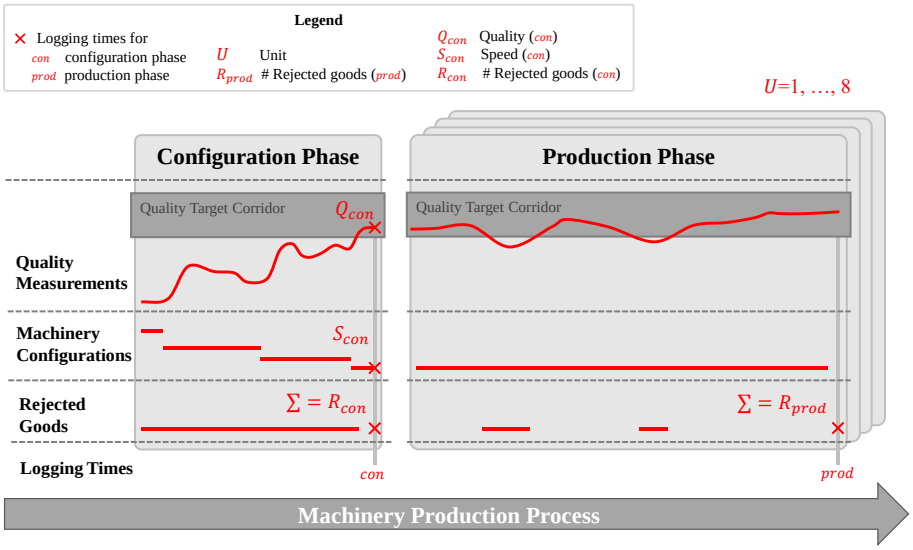

Fig. E.1. A schematic overview of a machinery production process, where machinery configurations, e.g. speed S_{con}, are adapted within the *configuration phase* to obtain the required quality target for a quality measurement Q_{con} avoiding rejected goods R_{con}. This configuration aims to reduce the number of rejected goods R_{prod} within a high throughput *production phase* over several units U.

Goal and Challenges: As causal structural knowledge serves as the basis for data-driven decision support, e.g., see [13,16,23], we aim to estimate the underlying causal structures of the above-described discrete manufacturing process. In practice, we face the following well-known challenges:

C1) High-dimensional settings, as hundreds of factors are involved in controlling and monitoring the machinery production process, e.g., see [13,22];

C2) Semi-structured and erroneous logged data due to inconsistencies of an insufficient logging procedure [35,40];

C3) Mixed discrete-continuous data, e.g., continuous quality measurements or discrete machinery configurations, as common in practice, e.g., see [13,16].

To allow for a comprehensive real-world example, cf. C1) we restrict our attention to the main factors within the production process as proposed by domain experts. In particular, we consider the variables described in the schematic overview in Fig. E.1. In this context, the manufacturing of one good in the production phase can be described by the variables Q_{con}, S_{con}, and R_{con} determined in the configuration phase, and R_{prod} as well as the locality, i.e., unit U, the good is produced. Further, we apply the transformation rules proposed by Hagedorn et al. [13] and leverage the knowledge of domain experts to receive sound observational data, cf. C2). On this basis, we evaluate the accuracy of `mCMIkNN` in mixed discrete-continuous real-world data compared to commonly applied discretization-based approaches for causal discovery, cf. C3).

E.2 Empirical Evaluation of the Discrete Manufacturing Scenario

Data: In line with the manufacturing process described in Appendix E.1, we consider the continuous variables Q_{con}, S_{con}, R_{con}, and R_{prod} (which may follow a mixture distribution), as well as the discrete U. The transformation of the semi-structured log data yields approximately 1 300 sound samples of the discrete manufacturing process, each associated with one produced good with thousands of pieces made in the production phase. In line with all experiments in the paper, we max-min normalize continuous variables.

Assumed DAG: With the help of domain experts, we define the DAG depicted in Fig. E.2 that serves as ground truth representing the underlying causal structures of the discrete manufacturing process in Fig. E.1. The causal structures arise as follows. Quality measurements Q_{con} and rejections R_{con} are used for adjustment of the processing speed S_{con} in the configuration phase with respective edges $Q_{con} \to R_{con}$, $R_{con} \to S_{con}$. Then, for all units U, the high throughput production process is started according to the configuration. Hence, we assume the following edges hold true $Q_{con} \to R_{prod}$, $R_{con} \to R_{prod}$, $S_{con} \to R_{prod}$, and $U \to R_{prod}$.

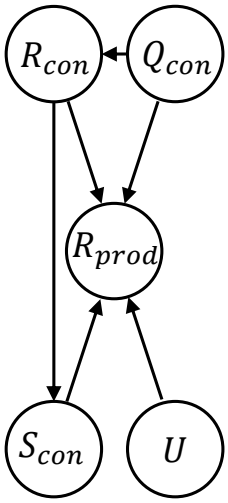

Fig. E.2. True underlying DAG.

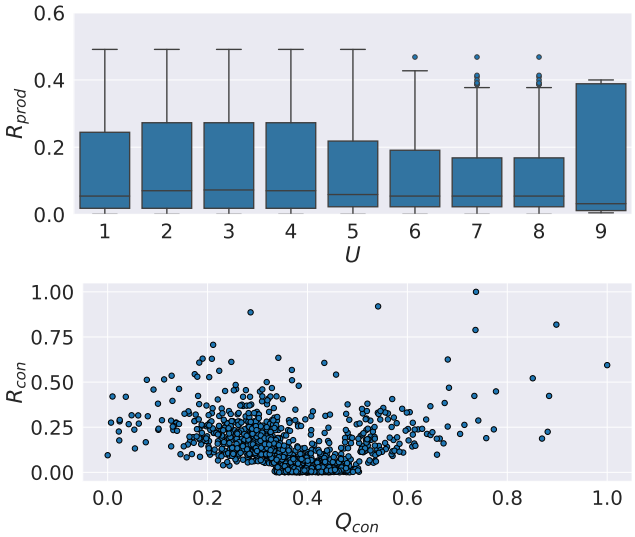

Fig. E.3. Real-world data characteristics for $U \to R_{prod}$ (top) and for $Q_{con} \to R_{prod}$ (bottom).

Our small real-world scenario covers the omnipresent characteristics of mixed discrete-continuous data in discrete manufacturing, cf. [13]. For example, see Fig. E.3, our data contains the mixed discrete-continuous relationship $U \to R_{prod}$ (top) and the non-linear relationship $Q_{con} \to R_{prod}$.

In this context, note that we cannot guarantee causal faithfulness, i.e., there may exist confounders not considered within our small example.

Empirical Evaluation: In line with the experiments on constraint-based causal discovery in Sec. 5.4, we apply the PC-stable algorithm ($\alpha = 0.05$) from [7] to estimate the Markov equivalence class $\mathcal{G}_{CPDAG}$ of the DAG $\mathcal{G}$. In this context, we compare `mCMIkNN` ($k_{perm} = 5$, $k_{CMI} = 25$, $M_{perm} = 100$) against the commonly applied discretization-based CI test `discχ²`.

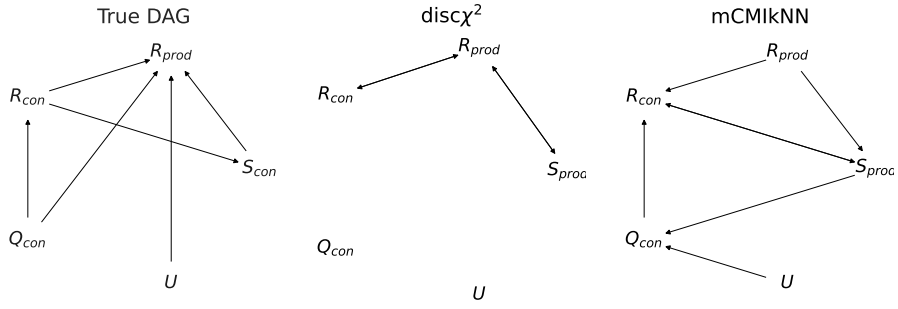

Fig. E.4. Assumed DAG (left) and the estimated CPDAGs using the PC-stable algorithm with `discχ²`, F1 = 0.4, (center) and `mCMIkNN`, F1 = 0.57 (right).

As depicted in Fig. E.4, the CPDAG estimated with PC-stable using `mCMIkNN` (right) is closer to the assumed true DAG (left) compared to the CPDAG estimated with PC-stable using `discχ²` (center). The performance difference is reflected in the F1 scores calculated on the respective skeletons, i.e., F1 = 0.57 for `mCMIkNN` vs. F1 = 0.4 for `discχ²`.

E.3 Summary and Limitations

Summary: We demonstrated that `mCMIkNN` outperforms the commonly used discretization-based approach for constraint-based causal discovery in a real-world discrete manufacturing scenario.

Limitations: In this context, note that the accuracy of the estimated CPDAGs is affected by latent confounding variables not present in the data. In particular, within many cycles with domain experts, we extended the small scenario to cover other driving factors within the machinery production process. In this context, a similar behavior was visible within more complex graphs of up to 50 variables where `mCMIkNN` outperformed `discχ²`, too. Overall, `mCMIkNN` was able to capture the CI characteristics of the mixed discrete-continuous discrete manufacturing data compared to `discχ²`. For more information on challenges of causal discovery in practice and constraint-based methods that allow for latent variables, we refer to [12,13,22], and [32,34,37], respectively.

References

1. Abellan-Nebot, J.V., Subirón, F.R.: A review of machining monitoring systems based on artificial intelligence process models. The International Journal of Advanced Manufacturing Technology **47**(1-4), 237–257 (2010)
2. Andrews, B., Ramsey, J., Cooper, G.F.: Scoring bayesian networks of mixed variables. International Journal of Data Science and Analytics **6**(1), 3–18 (Aug 2018)
3. Antos, A., Kontoyiannis, I.: Convergence properties of functional estimates for discrete distributions. Random Structures & Algorithms **19**(3-4), 163–193 (2001)
4. Berrett, T.B., Wang, Y., Barber, R.F., Samworth, R.J.: The conditional permutation test for independence while controlling for confounders. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **82**(1), 175–197 (2020)
5. Chen, K.S., Huang, M.: Performance measurement for a manufacturing system based on quality, cost and time. International Journal of Production Research **44**(11), 2221–2243 (2006)
6. Cheng, L., Guo, R., Moraffah, R., Sheth, P., Candan, K.S., Liu, H.: Evaluation methods and measures for causal learning algorithms. IEEE Transactions on Artificial Intelligence **3**, 924–943 (2022)
7. Colombo, D., Maathuis, M.H.: Order-independent constraint-based causal structure learning. Journal of Machine Learning Research **15**(116), 3921–3962 (2014)
8. Ernst, M.D.: Permutation methods: A basis for exact inference. Statistical Science **19**(4), 676–685 (2004)
9. Foraita, R., Friemel, J., Günther, K., Behrens, T., Bullerdiek, J., Nimzyk, R., Ahrens, W., Didelez, V.: Causal Discovery of Gene Regulation with Incomplete Data. Journal of the Royal Statistical Society Series A: Statistics in Society **183**(4), 1747–1775 (04 2020)
10. Frigyesi, A., Hössjer, O.: A test for singularity. Statistics & probability letters **40**(3), 215–226 (1998)
11. Gao, W., Kannan, S., Oh, S., Viswanath, P.: Estimating mutual information for discrete-continuous mixtures. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 5988–5999 (2017)
12. Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. Frontiers in genetics **10** (2019)
13. Hagedorn, C., Huegle, J., Schlosser, R.: Understanding unforeseen production downtimes in manufacturing processes using log data-driven causal reasoning. Journal of Intelligent Manufacturing **33**(7), 2027–2043 (Oct 2022)
14. Hagedorn, C., Lange, C., Huegle, J., Schlosser, R.: GPU acceleration for information-theoretic constraint-based causal discovery. In: Proceedings of The KDD'22 Workshop on Causal Discovery. pp. 30–60 (2022)
15. Huegle, J., Hagedorn, C., Boehme, L., Poerschke, M., Umland, J., Schlosser, R.: MANM-CS: Data generation for benchmarking causal structure learning from mixed discrete-continuous and nonlinear data. In: WHY-21 @ NeurIPS 2021 (2021)
16. Huegle, J., Hagedorn, C., Uflacker, M.: How causal structural knowledge adds decision-support in monitoring of automotive body shop assembly lines. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. pp. 5246–5248 (2020)
17. Kalisch, M., Bühlmann, P.: Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research **8**, 613–636 (2007)
18. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H., Bühlmann, P.: Causal inference using graphical models with the R package pcalg. Journal of Statistical Software, Articles **47**(11), 1–26 (2012)

19. Kim, I., Neykov, M., Balakrishnan, S., Wasserman, L.: Local permutation tests for conditional independence. The Annals of Statistics **50**(6), 3388–3414 (2022)
20. Koren, Y., Hu, S.J., Weber, T.W.: Impact of manufacturing system configuration on performance. CIRP annals **47**(1), 369–372 (1998)
21. Liang, S.Y., Hecker, R.L., Landers, R.G.: Machining process monitoring and control: The state-of-the-art. Journal of Manufacturing Science and Engineering **126**(2), 297–310 (2004)
22. Malinsky, D., Danks, D.: Causal discovery algorithms: A practical guide. Philosophy Compass **13**(1), e12470 (2018)
23. Marazopoulou, K., Ghosh, R., Lade, P., Jensen, D.D.: Causal discovery for manufacturing domains. CoRR **abs/1605.04056** (2016)
24. Marx, A., Yang, L., van Leeuwen, M.: Estimating conditional mutual information for discrete-continuous mixtures using multi-dimensional adaptive histograms. In: Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). pp. 387–395 (2021)
25. Meek, C.: Causal inference and causal explanation with background knowledge. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 403–410 (1995)
26. Mesner, O.C., Shalizi, C.R.: Conditional mutual information estimation for mixed, discrete and continuous data. IEEE Transactions on Information Theory **67**(1), 464–484 (2021)
27. Montgomery, D.C.: Introduction to Statistical Quality Control. John Wiley & Sons (2007)
28. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, 1st edn. (2000)
29. Peters, J., Janzing, D., Schölkopf, B.: Elements of Causal Inference: Foundations and Learning Algorithms. MIT Press (2017)
30. Phipson, B., Smyth, G.K.: Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. Statistical Applications in Genetics and Molecular Biology **9**(1) (2010)
31. Reisach, A., Seiler, C., Weichwald, S.: Beware of the simulated dag! causal discovery benchmarks may be easy to game. In: Advances in Neural Information Processing Systems. vol. 34, pp. 27772–27784 (2021)
32. Rohekar, R.Y., Nisimov, S., Gurwicz, Y., Novik, G.: Iterative causal discovery in the possible presence of latent confounders and selection bias. Advances in Neural Information Processing Systems **34**, 2454–2465 (2021)
33. Runge, J.: Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In: International Conference on Artificial Intelligence and Statistics. pp. 938–947. PMLR (2018)
34. Spirtes, P., Glymour, C.N., Scheines, R.: Causation, Prediction, and Search. Adaptive Computation and Machine Learning, MIT Press (2000)
35. Spirtes, P., Zhang, K.: Causal discovery and inference: Concepts and recent methodological advances. Applied Informatics **3**(1), 1–28 (2016)
36. Spirtes, P., Zhang, K.: Search for causal models. In: Handbook of Graphical Models, pp. 439–470. CRC Press (2018)
37. Strobl, E.V.: A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. International Journal of Data Science and Analytics **8**(1), 33–56 (2019)
38. Takata, S., Kirnura, F., van Houten, F.J., Westkamper, E., Shpitalni, M., Ceglarek, D., Lee, J.: Maintenance: changing role in life cycle management. CIRP annals **53**(2), 643–655 (2004)

39. Wang, H., Song, M.: Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic programming. The R journal **3 2**, 29–33 (2011)
40. Wuest, T., Weimer, D., Irgens, C., Thoben, K.D.: Machine learning in manufacturing: Advantages, challenges, and applications. Production and Manufacturing Research **4**(1), 23–45 (2016)
41. Zhang, K., Peters, J., Janzing, D., Schölkopf, B.: Kernel-based conditional independence test and application in causal discovery. In: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence. pp. 804–813 (2011)
42. Zhang, K., Schölkopf, B., Spirtes, P., Glymour, C.: Learning causality and causality-related learning: some recent progress. National science review **5**(1), 26–29 (2018)
43. Zhao, P., Lai, L.: Analysis of knn information estimators for smooth distributions. IEEE Transactions on Information Theory **66**(6), 3798–3826 (2019)
44. Zinde-Walsh, V., Galbraith, J.W.: A test of singularity for distribution functions. CIRANO-Scientific Publications 2011s-06 (2011)