entropy

MDPI

Article

A Conditional Mutual Information Estimator for Mixed Data and an Associated Conditional Independence Test

Lei Zan [1,2,*], Anouar Meynaoui [1], Charles K. Assaad [2], Emilie Devijver [1] and Eric Gaussier [1]

[1] Department of Mathematics, Information and Communication Sciences, Université Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2] R&D Department, EasyVista, 38000 Grenoble, France
* Correspondence: lei.zan@univ-grenoble-alpes.fr

Abstract: In this study, we focus on mixed data which are either observations of univariate random variables which can be quantitative or qualitative, or observations of multivariate random variables such that each variable can include both quantitative and qualitative components. We first propose a novel method, called CMIh, to estimate conditional mutual information taking advantages of the previously proposed approaches for qualitative and quantitative data. We then introduce a new local permutation test, called LocAT for local adaptive test, which is well adapted to mixed data. Our experiments illustrate the good behaviour of CMIh and LocAT, and show their respective abilities to accurately estimate conditional mutual information and to detect conditional (in)dependence for mixed data.

Keywords: mixed data; conditional mutual information; conditional independence testing; permutation tests

Citation: Zan, L.; Meynaoui, A.; Assaad, C.K.; Devijver, E.; Gaussier, E. A Conditional Mutual Information Estimator for Mixed Data and an Associated Conditional Independence Test. *Entropy* **2022**, *24*, 1234. https://doi.org/10.3390/e24091234

Academic Editor: Antonio M. Scarfone

Received: 28 July 2022
Accepted: 31 August 2022
Published: 2 September 2022

1. Introduction

Measuring the (in)dependence between random variables from data when the underlying joint distribution is unknown plays a key role in several settings, as in causal discovery [1], graphical model inference [2] or feature selection [3]. Many dependence measures have been introduced in the literature to quantify the dependence between random variables, as *Mutual Information* (MI) [4], *distance correlation* [5], kernel-based measures such as the *Hilbert–Schmidt Independence Criterion* (HSIC) [6], *COnstrained COvariance* (COCO) [7] or copula-based approaches [8]. We focus in this work on (conditional) mutual information, which has been successfully used in various contexts and has shown good practical performance in terms of the statistical power of the associated independence tests [9], and consider both quantitative and qualitative variables. A quantitative variable is a variable which has infinite support and values on which one can use more complex distances than the mere $(0-D)$ distance (which is 0 for two identical points and D for points with different values). All variables which do not satisfy these two conditions are deemed qualitative. Note that one can use the $(0-D)$ distance on any type of variables, and that this distance is the standard distance for nominal variables; one can of course use, if they exist, other distances than the $(0-D)$ on qualitative variables. Continuous variables as well as ordinal variables with infinite support are here quantitative, whereas nominal variables and ordinal variables with finite support are considered qualitative.

The conditional mutual information [10] between two quantitative random variables X and Y conditionally to a quantitative random variable Z is given by :

$$I(X;Y|Z) = \iiint P_{XYZ}(x,y,z) \log\left(\frac{P_{XY|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}\right) \mathrm{d}x\mathrm{d}y\mathrm{d}z, \quad (1)$$

where P_{XYZ} is the joint density of (X, Y, Z) and $P_{XY|Z}$ (respectively, $P_{X|Z}$ and $P_{Y|Z}$) is the density of (X, Y) (respectively, X and Y) given Z. Note that Equation (1) also applies to qualitative variables by replacing integrals by sums and densities by mass functions. The conditional mutual information can also be expressed in terms of entropies as:

$$I(X;Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z),$$

where $H(\cdot)$ is the Shannon entropy [11] defined as follows for a quantitative random variable X with density P_X:

$$H(X) = -\int P_X(x)\log(P_X(x))\,\mathrm{d}x.$$

Conditional mutual information characterizes conditional independence in the sense that $I(X;Y|Z) = 0$ if and only if X and Y are independent conditionally to Z.

Estimating conditional mutual information for purely qualitative or purely quantitative random variables is a well-studied problem [12,13]. The case of mixed datasets comprising both quantitative and qualitative variables is, however, less studied even though mixed data are present in many applications, for example as soon as one needs to threshold some quantitative values as in monitoring systems. The aim of this paper is to present a new statistical method to detect conditional (in)dependence for mixed data. To do so, we introduce both a new estimator of conditional mutual information as well as a new test to conclude on the conditional (in)dependence.

The remainder of the paper is organized as follows. Section 2 describes related work. We introduce in Section 3 our estimator, as well as some numerical comparisons with existing conditional mutual information estimators. We present in Section 4 the associated independence tests as well as numerical studies conducted on simulated and real datasets. Finally, Section 5 concludes the paper.

2. Related Work

We review here related work on (conditional) mutual information estimators as well as on conditional independence testing.

2.1. Conditional Mutual Information

A standard approach to estimate (conditional) mutual information from mixed data is to discretize the data and to approximate the distribution of the random variables by a histogram model defined on a set of intervals called bins [14]. Each bin corresponds to a single point for qualitative variables and to consecutive non-overlapping intervals for quantitative variables. Even if the approximation becomes better when dealing with smaller bins, finite sample size requires to carefully choose the number of bins. To efficiently generate adaptive histograms model from quantitative variables, Cabeli et al. [15] and Marx et al. [16] transform the problem into a model selection problem, using a criterion based on the minimum description length (MDL) principle. An iterative greedy algorithm is proposed to obtain the histogram model that minimizes the MDL score, from which one can derive joint and marginal distributions. The difference between the two methods rely on the estimation, which is conducted for each entropy term in Cabeli et al. [15] and globally in Marx et al. [16]. These approaches are very precise to estimate the value of the (conditional) mutual information even in multi-dimensional cases, but are computational costly when the dimensions increase.

To estimate entropy, two main families of approaches have been proposed. The first one is based on kernel-density estimates [17] and applies to quantitative data, whereas the second one is based on k-nearest neighbours and applies to both qualitative and quantitative data. The second one is preferred as it naturally adapts to the data density and does not require extensive tuning of kernel bandwidths. Using nearest neighbours of observations to estimate the entropy dates back to Kozachenko and Leonenko [18], which was then generalized to a k-nearest neighbour (kNN) approach by Singh et al. [19]. In this method, the distance to the k^{th} nearest neighbour is measured for each data point, the probability

density around each data point being substituted into the entropy expression. When k is fixed and the number of points is finite, each entropy term is noisy and the estimator is biased. However, this bias is distribution independent and can be subtracted out [20]. Along this line, Kraskov et al. [21] proposed an estimator for mutual information that goes beyond the sum of entropy estimators. This latter work was then extended to conditional mutual information in Frenzel and Pompe [12]. The resulting model, called FP, however, only deals with quantitative data.

More recently, Ross [22] and Gao et al. [23] introduced two approaches to estimate mutual information for mixed data, however, without any conditioning set. Following these studies, Rahimzamani et al. [24] proposed a measure of incompatibility between the joint probability $P_{XYZ}(x, y, z)$ and its factorization $P_{X|Z}(x|z)P_{Y|Z}(y|z)P_Z(z)$ called *graph divergence measure* and extended the estimator proposed in Gao et al. [23] to conditional mutual information, leading to a method called RAVK. As ties can occur with a non zero probability in mixed data, the number of neighbours has to be carefully chosen. Even more recently, Mesner and Shalizi [25] extended FP [12] to the mixed data case by introducing a qualitative distance metric for non-quantitative variables, leading to a method called MS. The choice of the qualitative and quantitative distances is a crucial point in MS [26]. FP, RAVK and MS all lead to an estimator of the form:

$$\widehat{I}(X;Y|Z) = \frac{1}{n}\sum_{i=1}^{n} \psi(k_{Ge,i}) - f(n_{Ge,XZ,i}) - f(n_{Ge,YZ,i}) + f(n_{Ge,Z,i}).$$

Ge stands for either FP, RAVK or MS, n represents the number of the observations and $\psi(.)$ is the digamma function. For FP, $k_{FP,i}$ is a constant hyper-parameter, $f(n_{FP,W,i}) = \psi(n_{FP,W,i} + 1)$ with W being either (X, Z), (Y, Z) or (Z). Denoting, as usual, the ℓ_∞ distance between i and its k-nearest-neighbour in global space by $\rho_{k,i}/2$, $n_{FP,W,i}$ represents the number of points in the joint space W that have an ℓ_∞ distance strictly smaller than $\rho_{k,i}/2$:

$$n_{FP,W,i} = \left|\left\{j : ||\omega_i - \omega_j||_\infty < \rho_{k,i}/2, j \neq i\right\}\right|, \tag{2}$$

where ω_j and ω_i represent the coordinates of the points in the space corresponding to W. For RAVK, to adapt it to mixed data, Mesner and Shalizi [25] proposed the use of $f(n_{RAVK,W,i}) = \log(n_{RAVK,W,i} + 1)$ where $n_{RAVK,W,i}$ includes boundary points:

$$n_{RAVK,W,i} = \left|\left\{j : ||\omega_i - \omega_j||_\infty \leq \rho_{k,i}/2, j \neq i\right\}\right|. \tag{3}$$

Furthermore, $k_{RAVK,i} = n_{RAVK,(X,Y,Z),i}$. Lastly, for MS, one has $f(n_{MS,W,i}) = \psi(n_{MS,W,i})$, $n_{MS,W,i}$ and $k_{MS,i}$ being defined as for RAVK.

We also want to mention the proposal made by Mukherjee et al. [27] of a two-stage estimator based on generative models and classifiers as well as the refinement introduced in Mondal et al. [28] and based on a neural network that integrates the two stages into a single training process. It is, however, not clear how to adapt to mixed data these methods primarily developed for quantitative data.

2.2. Conditional Independence Tests

To decide whether the estimated conditional mutual information value is small enough to conclude on the (in)dependence of two variables X and Y conditionally to a third variable Z in a finite sample regime, one usually relies on statistical independence tests. The null and the alternative hypotheses are, respectively, defined by

$$\mathcal{H}_0 : X \perp\!\!\!\perp Y|Z \qquad \text{and} \qquad \mathcal{H}_1 : X \not\perp\!\!\!\perp Y|Z,$$

where $\perp\!\!\!\perp$ means *independent of* and $\not\perp\!\!\!\perp$ means *not independent of*. In the independence testing literature, two main families exist, asymptotic and non-asymptotic ones (see e.g., [29], chapter 3). The former is used when the sample size is big enough and relies on the asymptotic distribution of the estimator under the null hypothesis, while the latter applies to any sample size without any prior knowledge of the asymptotic null distribution of the estimator. Note that conditional independence testing is a more difficult problem than its non-conditional counterpart [30]. In particular, the asymptotic behaviour of the conditional mutual information estimator under the null hypothesis is usually unknown.

Kernel-based tests are known for their capability to deal with nonlinearity and high dimensions. The Hilbert–Schmidt independence criterion (HSIC) has been first proposed for testing unconditional independence. Fukumizu et al. [31] extended HSIC to the conditional independence setting using the Hilbert-Schmidt norm of the conditional cross-covariance operator. Another representative of this test category is the kernel conditional independence test (KCIT) proposed by Zhang et al. [32]. It works by testing for vanishing correlation between residual functions in reproducing kernel Hilbert spaces. To reduce the computational complexity of KCIT, Strobl et al. [33] used random Fourier features to approximate KCIT and thereby proposed two tests, namely the randomised conditional independence test that explores the partial cross-covariance matrix between (X, Z) and Y, and the randomized conditional correlation test (RCoT) that tests X and Y after some transformations to remove the effect of Z. RCoT can be related to two-step conditional independence testing [34], computing first conditional expectations of feature maps and then testing the residuals. Doran et al. [35] also proposed a kernel conditional independence permutation test. They used a specific permutation of the samples to generate data from $P_{X|Z}(x|z)P_{Y|Z}(y|z)P_Z(z)$ which unfortunately requires solving a time-consuming linear program, then performed a kernel-based two-sample test [36]. However, kernel-based tests need to carefully adjust bandwidth parameters that characterise the length scales in the different subspaces of X, Y, Z and can only be implemented on purely quantitative data.

More recently, Shah and Peters [30] proposed the generalised covariance measure (GCM) test. For univariate X and Y, instead of testing for independence between the residuals from regressing X and Y on Z, the GCM tests for vanishing correlations. How to extend this approach to mixed data is, however, not clear. Tsagris et al. [37] employed likelihood-ratio tests based on regression models to devise conditional independence tests for mixed data; however, in their approach one needs to postulate a regression model.

Permutation tests [38] are popular when one wants to avoid assumptions on the data distribution. For testing the independence of X and Y conditionally to Z, permutation tests randomly permute all values in X. If this destroys the potential dependence between X and Y, as desired, this also destroys the one between X and Z, which is not desirable. In order to preserve the dependence between X and Z, Runge [39] proposed a local permutation test in which permutations within X are conducted within similar values of Z. We extend in this paper this test, designed for quantitative data, to the mixed data case.

3. Hybrid Conditional Mutual Information Estimation for Mixed Data

The two most popular approaches to estimate conditional mutual information are based on the k-nearest neighbour method [12,21], which has been mostly used on quantitative variables, or on histograms [15,16], particularly adapted to qualitative variables. We show in this section how these two approaches can be combined to derive an estimator for mixed data.

Let us consider three mixed random vectors X, Y and Z, where any of their components can be either qualitative or quantitative. Let us denote by X^t (respectively, Y^t, Z^t) the sub-vector of X (respectively, Y, Z) composed by the quantitative components. Similarly, we denote by X^ℓ (respectively, Y^ℓ, Z^ℓ)the sub-vector of qualitative components of X (respectively, Y, Z). Then, from the permutation invariance property of Shannon entropy, the conditional mutual information can be written as:

$$\begin{aligned} I(X;Y|Z) =& H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z) \\ =& H(X^t, X^\ell, Z^t, Z^\ell) + H(Y^t, Y^\ell, Z^t, Z^\ell) - H(X^t, X^\ell, Y^t, Y^\ell, Z^t, Z^\ell) - H(Z^t, Z^\ell). \end{aligned}$$

Now, from the property $H(U,V) = H(U) + H(V|U)$, which is valid for any couple of random variables (U, V), one gets:

$$\begin{aligned} I(X;Y|Z) =& H(X^t, Z^t|X^\ell, Z^\ell) + H(Y^t, Z^t|Y^\ell, Z^\ell) - H(X^t, Y^t, Z^t|X^\ell, Y^\ell, Z^\ell) \\ & - H(Z^t|Z^\ell) + H(X^\ell, Z^\ell) + H(Y^\ell, Z^\ell) - H(X^\ell, Y^\ell, Z^\ell) - H(Z^\ell). \end{aligned} \quad (4)$$

Note that here the conditioning is only expressed with respect to qualitative components, which leads to a simpler estimation than the one obtained by conditioning with quantitative variables. We now detail how the different terms in the above expression are estimated.

3.1. Proposed Hybrid Estimator

Let us now consider an independently and identically distributed sample of size n denoted $(X_i, Y_i, Z_i)_{i=1,\dots,n}$. We estimate the qualitative entropy terms of Equation (4), namely $H(X^\ell, Z^\ell)$, $H(Y^\ell, Z^\ell)$, $H(X^\ell, Y^\ell, Z^\ell)$ and $H(Z^\ell)$, using histograms in which bins are defined by the Cartesian product of qualitative values. We provide here the estimation of $H(X^\ell, Z^\ell)$, the other terms are estimated in the same way. The theoretical entropy is expressed as:

$$H(X^\ell, Z^\ell) = -\mathbb{E}\Big[\log P_{X^\ell,Z^\ell}(X^\ell, Z^\ell)\Big] = -\sum_{\substack{x^\ell \in \Omega(X^\ell) \\ z^\ell \in \Omega(Z^\ell)}} P_{X^\ell,Z^\ell}(x^\ell, z^\ell) \log\Big(P_{X^\ell,Z^\ell}(x^\ell, z^\ell)\Big),$$

where $\Omega(\cdot)$ corresponds to the probability space of a given random variable and P_{X^ℓ,Z^ℓ} is the probability distribution of (X^ℓ, Z^ℓ). The probability distribution of qualitative variables can be directly estimated via their empirical versions:

$$\widehat{P}_{X^\ell,Z^\ell}(x^\ell, z^\ell) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{(X_i^\ell, Z_i^\ell)=(x^\ell, z^\ell)\}}, \tag{5}$$

with $\mathbb{1}_{\{\cdot\}}$ is the indicator function. The resulting plug-in estimator is then given by

$$\widehat{H}(X^\ell, Z^\ell) = -\sum_{\substack{x^\ell \in \Omega(X^\ell) \\ z^\ell \in \Omega(Z^\ell)}} \widehat{P}_{X^\ell,Z^\ell}(x^\ell, z^\ell) \log\Big(\widehat{P}_{X^\ell,Z^\ell}(x^\ell, z^\ell)\Big). \tag{6}$$

Let us now turn to the conditional entropies of Equation (4) for quantitative variables conditioned on qualitative variables and let us consider the term $H(X^t, Z^t|X^\ell, Z^\ell)$. By marginalizing on (X^ℓ, Z^ℓ) one obtains:

$$H(X^t, Z^t|X^\ell, Z^\ell) = \sum_{\substack{x^\ell \in \Omega(X^\ell) \\ z^\ell \in \Omega(Z^\ell)}} H(X^t, Z^t|X^\ell = x^\ell, Z^\ell = z^\ell) P_{X^\ell,Z^\ell}(x^\ell, z^\ell). \tag{7}$$

As before, the probabilities involved in Equation (7) are estimated by their empirical versions. The estimation of the conditional entropies $H(X^t, Z^t|X^\ell = x^\ell, Z^\ell = z^\ell)$ is performed using the classical nearest neighbour estimator [19] with the constraint that $(X^\ell, Z^\ell) = (x^\ell, z^\ell)$: the estimation set consists of the sample points such that $(X^\ell, Z^\ell) = (x^\ell, z^\ell)$. The resulting estimator is given by:

$$\widehat{H}(X^t, Z^t|X^\ell = x^\ell, Z^\ell = z^\ell) = \psi(n_{xz}) - \psi(k_{xz}) + \log(v_{d_{xz}}) + \frac{d_{xz}}{n_{xz}}\sum_{i=1}^{n_{xz}} \log \xi_{xz}(i), \tag{8}$$

where ψ is the digamma function, n_{xz} is the size of the subsample space for which $(X_i^\ell, Z_i^\ell) = (x^\ell, z^\ell)$, $\xi_{xz}(i)$ is twice the distance of the i^{th} subsample point to its k_{xz} nearest neighbour, and k_{xz} is the number of nearest neighbours retained. In the sequel, we set k_{xz} to $\max(\lfloor n_{xz}/10 \rfloor, 1)$, with $\lfloor\cdot\rfloor$ the floor function, following Runge [39] which showed that this value behaves well in practice. As originally proposed in [21] and adopted in subsequent studies, we rely on the ℓ_∞-distance which is associated with the maximum norm: for a vector $w = (w_1, \dots, w_m)$ in $\mathbb{R}^m$, $\|w\|_\infty = \max(|w_1|, \dots, |w_m|)$. Finally, d_{xz} is the dimension of the vector (X^t, Z^t) and $v_{d_{xz}}$ is the volume of the unit ball for the distance metric associated with the maximum norm in the joint space associated with X^t and Z^t. The other entropy terms are estimated in the same way, the associated estimators being denoted by $\widehat{H}(Z^t|Z^\ell)$, $\widehat{H}(Y^t, Z^t|Y^\ell, Z^\ell)$ and $\widehat{H}(X^t, Y^t, Z^t|X^\ell, Y^\ell, Z^\ell)$.

The conditional mutual information estimator for mixed data, which we will refer to as *CMIh*, finally amounts to:

$$\begin{aligned}\widehat{I}(X;Y|Z) = \widehat{H}(X^t,Z^t|X^\ell,Z^\ell) + \widehat{H}(Y^t,Z^t|Y^\ell,Z^\ell) - \widehat{H}(X^t,Y^t,Z^t|X^\ell,Y^\ell,Z^\ell)\\ - \widehat{H}(Z^t|Z^\ell) + \widehat{H}(X^\ell,Z^\ell) + \widehat{H}(Y^\ell,Z^\ell) - \widehat{H}(X^\ell,Y^\ell,Z^\ell) - \widehat{H}(Z^\ell),\end{aligned} \tag{9}$$

where the different terms are obtained through Equations (5)–(8). Notice that all the volume-type terms, as for the $\log(v_{d_{xz}})$ term in Equation (8), are canceled out in Equation (9). Indeed, it is well known that the volume of the unit ball in $\mathbb{R}^p$ with respect to to the maximum norm is 2^p and this leads to the following plain equation:

$$\log(v_{d_{xyz}}) - \log(v_{d_{xz}}) - \log(v_{d_{yz}}) + \log(v_{d_z}) = \log\left(\frac{2^{d_{xyz}}2^{d_z}}{2^{d_{xz}}2^{d_{yz}}}\right) = \log\left(\frac{2^{d_x+d_y+d_z}2^{d_z}}{2^{d_x+d_y}2^{d_y+d_z}}\right) = 0.$$

The main steps to derive the hybrid estimator CMIh are summarized in Algorithm 1.

Algorithm 1 Hybrid estimator CMIh

Input $(X_i, Y_i, Z_i)_{i=1,\dots,n}$ the data, *isCat* indexes of qualitative components;
Separate qualitative and quantitative components $(X_i^t, X_i^\ell, Y_i^t, Y_i^\ell, Z_i^t, Z_i^\ell)_{i=1,\dots,n}$ using *isCat*;
$pointsInBin = \{\}$: indexes of points in each bin;
$densityOfBin = \{\}$: frequency of each bin;
$qualitativeEntropy = 0$: entropy of qualitative components;
$quantitativeEntropy = 0$: entropy of quantitative components;
if $(X^\ell, Y^\ell, Z^\ell) = \emptyset$ **then**
  $qualitativeEntropy += 0$;
  if $(X^t, Y^t, Z^t) = \emptyset$ **then**
    $quantitativeEntropy += 0$;
  else
    Compute $\widehat{H}(X^t, Y^t, Z^t)$ using the analogous of Equation (8);
    $quantitativeEntropy += \widehat{H}(X^t, Y^t, Z^t)$;
  end if
else
  for $(x^\ell, y^\ell, z^\ell) \in \Omega((X^\ell, Y^\ell, Z^\ell))$ **do**
    $pointsInBin[(x^\ell, y^\ell, z^\ell)] = \{i \in \{1, \dots, n\} : (X_i^\ell = x^\ell, Y_i^\ell = y^\ell, Z_i^\ell = z^\ell)\}$;
    $densityOfBin[(x^\ell, y^\ell, z^\ell)] = length(pointsInBin[(x^\ell, y^\ell, z^\ell)])/n$;
  end for
  $\widehat{H}(X^\ell, Y^\ell, Z^\ell) = 0$;
  for $k \in keys(densityOfBin)$ **do**
    $p = densityOfBin[k]$;
    $\widehat{H}(X^\ell, Y^\ell, Z^\ell) += -p\log(p)$;
  end for
  $qualitativeEntropy += \widehat{H}(X^\ell, Y^\ell, Z^\ell)$;
  if $(X^t, Y^t, Z^t) = \emptyset$ **then**
    $quantitativeEntropy += 0$;
  else
    for $k \in keys(densityOfBin)$ **do**
      $p = densityOfBin[k]$;
      Compute $\widehat{H}(X^t, Y^t, Z^t|(X^\ell, Y^\ell, Z^\ell) = k)$ using the analogous of Equation (8) on observations $pointsInBin[k]$;
      $\widehat{H}(X^t, Y^t, Z^t|X^\ell, Y^\ell, Z^\ell) += p\widehat{H}(X^t, Y^t, Z^t|(X^\ell, Y^\ell, Z^\ell) = k)$;
    end for
    $quantitativeEntropy += \widehat{H}(X^t, Y^t, Z^t|X^\ell, Y^\ell, Z^\ell)$;
  end if
end if
Compute other terms in Equation (9) using marginalization of the joint density;
$\widehat{I}(X;Y|Z) = quantitativeEntropy + qualitativeEntropy$;
Output $\widehat{I}(X;Y|Z)$.

Remark 1. *It is worth mentioning that our estimation of the entropy of the quantitative part is slightly different from the one usually used. In our estimation, the choice of the number of nearest neighbours is conducted independently for each entropy term and only with respect to the corresponding subsample size. This methodological choice yields more accurate estimators. Another important point is that the nearest neighbours are always computed on quantitative components as the qualitative components serve only as conditioning in Equation (9) or are involved in entropy terms estimated through Equation (6). Because of that, we can dispense with defining a distance on qualitative components, which is tricky as illustrated in Section 3.2.*

Consistency. Interestingly, the above hybrid estimator is asymptotically unbiased and consistent, as shown below.

Theorem 1. *Let (X, Y, Z) be a qualitative-quantitative mixed random vector. The estimator $\widehat{I}(X;Y|Z)$ defined in Equation (9) is consistent. Meaning that, for all $\varepsilon > 0$*

$$\lim_{n\to\infty} P(|\widehat{I}(X;Y|Z) - I(X;Y|Z)| > \varepsilon) = 0.$$

In addition, $\widehat{I}(X;Y|Z)$ is asymptotically unbiased, that is

$$\lim_{n\to\infty} \mathbb{E}[\widehat{I}(X;Y|Z) - I(X;Y|Z)] = 0.$$

Proof. It is well known that all linear combination of consistent estimators is consistent. This directly stems from Slutsky's theorem [40]. It remains to show the consistency of each term in the right-hand side of Equation (9). Histogram-based estimators $\widehat{H}(X^\ell, Z^\ell)$, $\widehat{H}(Y^\ell, Z^\ell)$, $\widehat{H}(X^\ell, Y^\ell, Z^\ell)$ and $\widehat{H}(Z^\ell)$ are consistent according to [41]. By analogy, we only show the consistency of the estimator $\widehat{H}(X^t, Z^t|X^\ell, Z^\ell)$, the same results apply to the remaining estimators. Let $\varepsilon > 0$, we write

$$\begin{aligned} & P(|\widehat{H}(X^t, Z^t|X^\ell, Z^\ell) - H(X^t, Z^t|X^\ell, Z^\ell)| > \varepsilon) \\ = & \sum_{\substack{x^\ell \in \Omega(X^\ell) \\ z^\ell \in \Omega(Z^\ell)}} P(|\widehat{H}(X^t, Z^t|X^\ell, Z^\ell) - H(X^t, Z^t|X^\ell, Z^\ell)| > \varepsilon | X^\ell = x^\ell, Z^\ell = z^\ell) \\ & \times P(X^\ell = x^\ell, Z^\ell = z^\ell). \end{aligned}$$

Now, conditionally to given values of X^ℓ and Z^ℓ, the estimator $\widehat{H}(X^t, Z^t|X^\ell, Z^\ell)$ is the traditional k-nearest neighbors built using the maximum-norm distance. This estimator is shown to be consistent, the reader can refer to [42] for more details. In other words,

$$\lim_{n\to\infty} P(|\widehat{H}(X^t, Z^t|X^\ell, Z^\ell) - H(X^t, Z^t|X^\ell, Z^\ell)| > \varepsilon | X^\ell = x^\ell, Z^\ell = z^\ell) = 0.$$

This concludes the proof of consistency. Moreover, knowing that the histogram and k-nearest neighbors estimators are asymptotically unbiased, it is plain that our estimator also has this property. □

3.2. Experimental Illustration

We compare in this section our estimator, CMIh, with several estimators described in Section 2, namely FP [12], MS [25], RAVK [24], and LH [16]. FP, MS and RAVK are methods based on the k-nearest neighbour approach. As for CMIh, the hyper-parameter k for these methods is set to the maximum value of $\lfloor n/10 \rfloor$ and 1, where n is the number of sampling points. To be consistent, we use for all three methods the widely used $(0-D_\ell)$ distance for the qualitative components: this distance is 0 for two equal qualitative values and D_ℓ otherwise. In our experiments, D_ℓ is set to 1, following [25]. Laslty, for FP, which was designed for quantitative data, we set the minimum value of $n_{FP,W,i}$ to 1 to avoid $n_{FP,W,i} = 0$ in Equation (2). Moreover, LH is a histogram method based on MDL [16]. We

use the default values for the hyper-parameters of this method: the maximum number of iterations, i_{max}, is set to 5, the threshold to detect qualitative points is also set to 5, the number of initial bins in quantitative component, K_{init}, is set to $20\log(n)$ and the maximum number of bins, K_{max}, is set to $5\log(n)$ (all entropies are computed in natural logarithm).

To assess the behaviour of the above methods, we first consider the mutual information with no conditioning ($I(X;Y)$), then with a conditioning variable which is independent of the process so that $I(X;Y|Z) = I(X;Y)$, and finally with a conditioning variable which makes the two others independent, such that $I(X;Y|Z) = 0$. We illustrate these three cases by either considering that X and Y are both quantitative or mixed, in which case they can have either balanced or unbalanced qualitative classes. Lastly, following [16,25], the conditioning variable Z is always qualitative.

Each (conditional) mutual information is computable theoretically so that one can measure the mean squared error (MSE) between the estimated value and the ground truth, which will be our evaluation measure. For each of the above experiments, we sample data with sample size n varying from 500 to 2000 and generate 100 data sets per sample size to compute statistics. More precisely, we use the following experimental settings, the first three ones being taken from [16,23,25]. The last four ones shed additional light on the different methods. Note that, as we reuse here the settings defined in [16,23,25], qualitative variables are generated either from a uniform distribution on a discrete set, a binomial distribution or a Poisson distribution, this latter case being an exception to our definition of what is a qualitative variable. We do not want to argue here on whether the Poisson variable should be considered quantitative or qualitative and simply reproduce here a setting used in previous studies for comparison purposes.

- *MI quantitative.* $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right)$ with $I(X;Y) = -\log(1-0.6^2)/2$.
- *MI mixed.* $X \sim \mathcal{U}(\{0,\dots,4\})$ and $Y|X = x \sim \mathcal{U}([x, x+2])$, we get $I(X;Y) = \log(5) - 4\log(2)/5$;
- *MI mixed imbalanced.* $X \sim Exp(1)$ and $Y|X = x \sim 0.15\delta_0 + 0.85Pois(x)$. The ground truth is $I(X;Y) = 0.85(2\log 2 - \gamma - \sum_{k=1}^{\infty} \log k 2^{-k}) \approx 0.256$, where γ is the Euler-Mascheroni constant.
- *CMI quantitative, CMI mixed and CMI mixed imbalanced.* We use the previous setting and add and independent qualitative random variable $Z \sim Bi(3, 0.5)$.
- *CMI quantitative* $\perp\!\!\!\perp$. $Z \sim Bi(9, 0.5)$, $X|Z = z \sim \mathcal{N}(z, 1)$ and $Y|Z = z \sim \mathcal{N}(z, 1)$, the ground truth is then $I(X;Y|Z) = 0$.
- *CMI mixed* $\perp\!\!\!\perp$. $Z \sim \mathcal{U}(\{0,\dots,4\})$, $X|Z = z \sim \mathcal{U}([z, z+2])$ and $Y|Z = z \sim Bi(z, 0.5)$, the ground truth is then $I(X;Y|Z) = 0$.
- *CMI mixed imbalanced* $\perp\!\!\!\perp$. $X \sim Exp(10)$, $Z|X = x \sim Pois(x)$ and $Y|Z = z \sim Bi(z + 5, 0.5)$, the ground truth is $I(X;Y|Z) = 0$.

Figure 1 displays the mean squared error (MSE) of the different methods in the different settings on a log-scale. As one can note, FP performs well in the purely quantitative case with no conditioning but is, however, not competitive in the mixed data case. MS and RAVK are close to each other and, not surprisingly, they have similar performance in most cases. MS, however, has a main drawback as it gives the value 0, or close to 0, to the estimator in some particular cases. Indeed, as noted by [25], if, for all points i, the k-nearest neighbour is always determined by Z, then, regardless of the relationship between X, Y and Z, $k_{MS,i} = n_{MS,XZ,i} = n_{MS,YZ,i} = n_{MS,Z,i}$ and the estimator equals to 0.

In addition, if the k-nearest-neighbour distance of a point i, $\rho_{k,i}/2$, is such that $\rho_{k,i}/2 \geq D_\ell$ where $D_\ell \in \mathbb{N}$ is the distance between different values of qualitative variables, then one has:

$$n_{MS,YZ,i} = n_{MS,Z,i} = n \text{ and } k_{MS,XYZ,i} = n_{MS,XZ,i}.$$

The first equality directly derives from the fact that one needs to consider points outside the qualitative class of point i (as $\rho_{k,i}/2 \geq D_\ell$) and that all points outside this class are at the same distance (D_ℓ). By definition, $n_{MS,YZ,i} \leq n_{MS,Z,i}$; furthermore, $n_{MS,Z,i} \leq n_{MS,YZ,i}$

as a neighbour of i in XZ with distance $\geq D_\ell$ is a neighbour of i in XYZ as Y cannot lead to a higher distance, which explains the second equality.

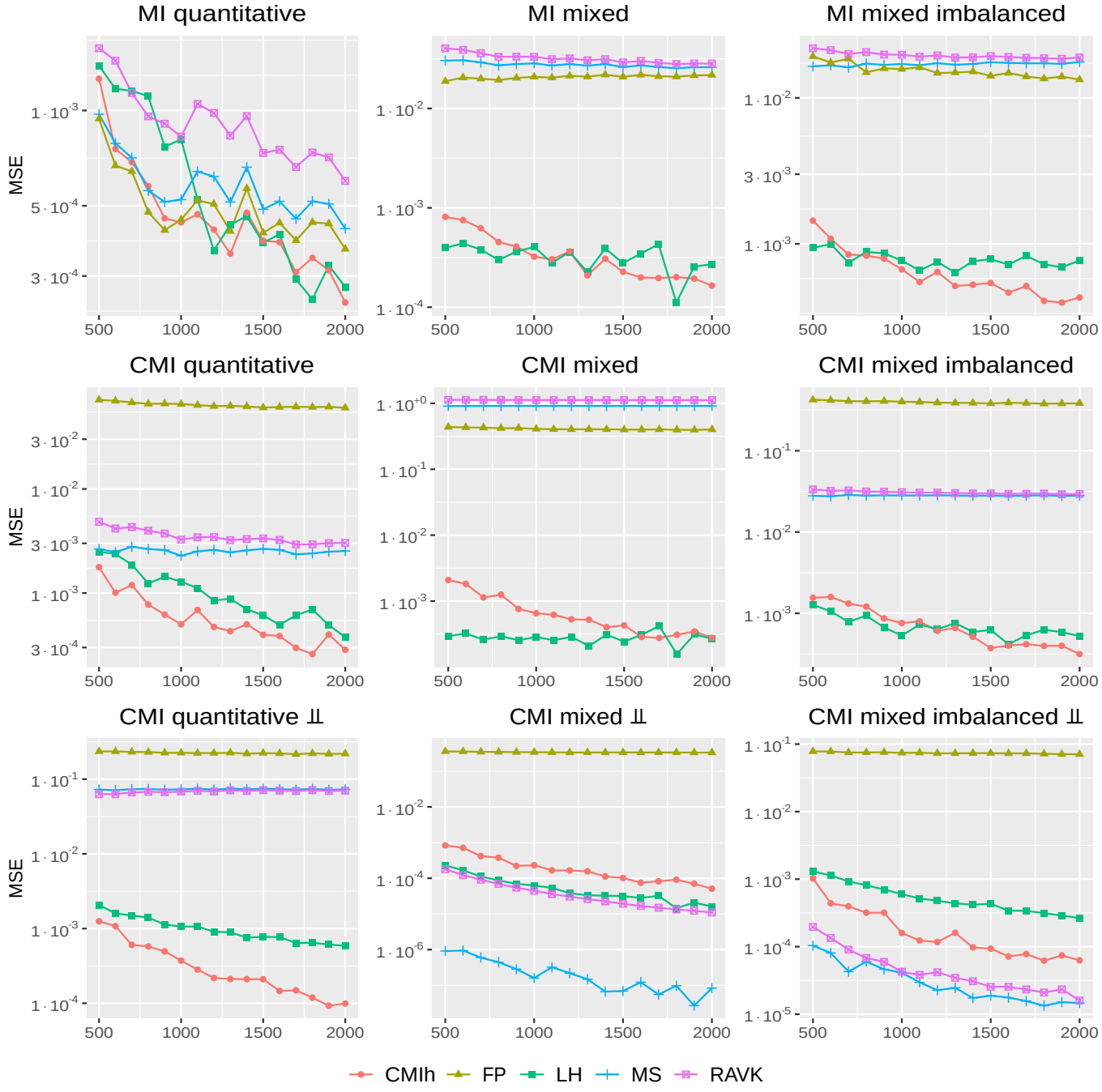

Figure 1. *Synthetic data with known ground truth.* MSE (on a log-scale) of each method with respect to the sample size (in abscissa) over the nine settings retained.

If a majority of points satisfy the above condition ($\rho_{k,i}/2 \geq D_\ell$), then MS will yield an estimator close to 0, regardless of the relation between the different variables. This is exactly what is happening in the mixed and mixed imbalance cases as the number of nearest points considered, at least 50, can be larger than the number of points in a given qualitative class. In such cases, MS will tend to provide estimators close to 0, which is the desired behaviour in the bottom-middle and bottom-right plots of Figure 1, but not in the top-middle, top-right, middle-middle and middle-right plots (in these latter cases, the ground truth is not 0 which explains the relatively large MSE value of MS and RAVK). Our

proposed estimator does not suffer from this drawback as we do not directly compare two different types of distances, one for quantitative and one for qualitative data.

Comparing LH and CMIh, one can see that, overall, these two methods are more robust than the other ones. The first and second lines of Figure 1 show that the additional independent qualitative variables Z does not have a large impact on the accuracy of the two estimators. The comparison of the second and third lines of Figure 1 furthermore suggests that, if the relationship between variables changes, the two estimators still have a stable performance.

Sensitivity to dimensionality. We conclude this comparison by testing how sensitive the different methods are to dimensionality. To do so, we first increase the dimensionality of the conditioning variable Z from 1 to 4 in a setting where X and Y are dependent and independent of Z (we refer to this setting as M-CMI for multidimensional conditional mutual information):

$$X \sim \mathcal{U}(\{0,\dots,4\}),\ Y|X=x \sim \mathcal{U}([x,x+2]),\ Z_r \sim Bi(3,0.5),\ r \in \{0,...,4\}.$$

The ground truth in this case is $I(X;Y|Z_1,\dots,Z_4) = I(X;Y) = \log(5) - 4\log(2)/5$.

The results of this first experiment, based on 100 samples of size 2000 for the different components of Z (from 0 to 4), are displayed in Figure 2 (left). As one can observe, our method achieves an MSE close to 0.001 even though the dimension increases to 4. LH has a comparable accuracy for small dimensions but deviates from the true value for higher dimensions. For MS and RAVK, as mentioned in Mesner and Shalizi [25], when X and Y have fixed-dimension, the higher the dimension of Z, the greater the probability that the estimator will give a zero value. This can explain why for dimensions above 1, the MSE remains almost constant for these two methods. Lastly, FP performs poorly when increasing the dimension of the conditioning set.

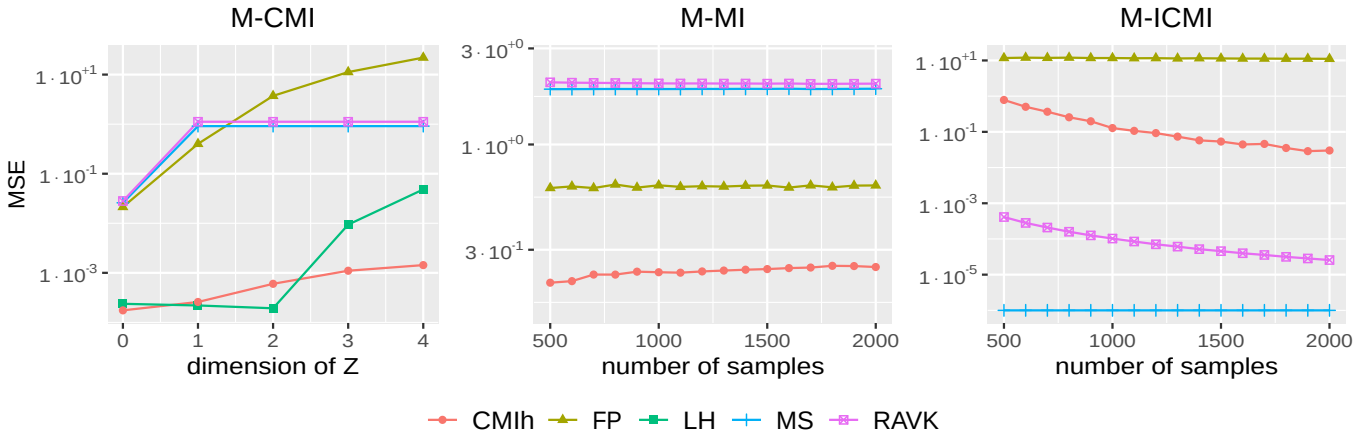

Figure 2. *Sensitivity to dimensionality* **Left**: MSE (on a log-scale) of each method for the multidimensional conditional mutual information (M-CMI) when increasing the dimension (x-axis) of the conditional variable from 0 to 4; the sample size is fixed to 2000. **Middle**: MSE (on a log-scale) of each method but LH for the multidimensional mutual information (M-MI) when increasing the number of observations. **Right**: MSE (on a log-scale) of each method but LH for the multidimensional independent conditional mutual information (M-ICMI) when increasing the number of observations.

It is also interesting to look at the computation time of each method on the above data, given in Table 1. One can note that our method is faster than the other ones and remains stable when the dimension of Z increases.

Let B denote the cardinal of the Cartesian product $X^\ell \times Y^\ell \times Z^\ell$ ($B = 1$ when all variables are quantitative and $B = 4$ in the setting retained here). The complexity of computing the four entropy terms in CMIh (Equation (9)) containing only qualitative variables is $\mathcal{O}(Bn)$ according to Equation (6). For the other entropy terms, one needs to

apply at most B times the computation in Equation (8), which has an average complexity of $\mathcal{O}((\frac{n}{B})^2 k m_t)$ (and $\mathcal{O}((\frac{n}{B})\log(\frac{n}{B})(k+m_t))$ for the approximation using KD-trees [43]), where $\frac{n}{B}$ represents the average number of sample points considered in Equation (8), $k = \max(\lfloor n/10 \rfloor, 1)$, and m_t is the number of quantitative components over all variables ($m_t = 2$ in the setting considered here). Thus, the overall complexity of CMIh is $\mathcal{O}(Bn + \frac{k m_t n^2}{B})$ (and $\mathcal{O}(Bn + (k+m_t)n\log(\frac{n}{B}))$ with KD-trees). In contrast, the complexity of MS, RAVK and FP is $\mathcal{O}(kmn^2)$ (and $\mathcal{O}((k+m)n\log(n))$ using KD-trees), where m is the number of dimensions over all variables ($m = 6$ in the setting considered here). This explains the differences observed in Table 1. Lastly, note that the complexity of LH is $\mathcal{O}(K^m_{max} K^2_{init} i_{max} m_t)$ [16], which limits its application to very small datasets.

We then focus on the multivariate version of (unconditional) mutual information for mixed data based using the following generative process (this setting is referred to as M-MI for multidimensional mutual information):

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix} \right), \; X_2 \sim \mathcal{U}(\{0,\dots,4\}), \; Y_2|X_2 = x_2 \sim \mathcal{U}([x_2, x_2+2]),$$

$$X_3 \sim Exp(1) \text{ and } Y_3|X_3 = x_3 \sim 0.15\delta_0 + 0.85\text{Pois}(x_3).$$

The ground truth in this case is $I(X_1, X_2, X_3; Y_1, Y_2, Y_3) \approx 1.534$.

Figure 2 (middle) displays the results obtained by the different methods but LH, computationally too complex to be used on datasets of a reasonable size, when the number of observations increases from 500 to 2000. As one can note, CMIh is the only method yielding an accurate estimate of the mutual information on this dataset. Both RAVK and MS suffer again from the fact that they yield estimates close to 0, which is problematic on this data. We give below another setting in which this behaviour is interesting; it remains nevertheless artificial.

Lastly, we consider the case where the two variables of interest are conditionally independent (we refer to this case as M-ICMI for multidimensional independent conditional mutual information). The generative process we used is:

$$Z_1 \sim \mathcal{U}(\{0,\dots,4\}), \; Z_2 \sim Bi(3, 0.5), \; Z_3 \sim Exp(1), \; Z_4 \sim Exp(10),$$

$$X_1, X_2|(Z_3 = z_3, Z_4 = z_4) \sim \mathcal{N}\left(\begin{pmatrix} z_3 \\ z_4 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right), \; X_3|(Z_1 = z_1, Z_2 = z_2) \sim Bi(z_1 + z_2, 0.5),$$

$$Y|(Z_1 = z_1, Z_2 = z_2) \sim Bi(z_1 + z_2, 0.5).$$

The ground truth in this case is $I(X_1, \dots, X_3; Y|Z_1, \dots, Z_4) = 0$.

Figure 2 (right) displays the results obtained on all methods but LH. As for the univariate case, both RAVK and MS obtain very good results here but this is due to their pathological behaviour discussed above. CMIh yields a reasonable estimate (with an MSE below 0.1) when the number of observations exceeds 1250. FP fails here to provide a reasonable estimate.

Table 1. We report, for each method, the mean computation time in seconds (its variance is given in parentheses), while varying the size of the conditional set from 0 to 4 with sample size fixed to 2000.

Dim of Z	0	1	2	3	4
CMIh	8.30(0.14)	5.30(0.05)	4.37(0.04)	4.16(0.04)	4.39(0.08)
FP	16.19(0.40)	22.09(0.27)	24.28(0.21)	25.91(0.08)	27.41(0.07)
LH	0.54(0.07)	1.09(0.02)	6.52(0.12)	58.58(13.74)	691.68(123.90)
MS	16.28(0.40)	22.08(0.07)	24.26(0.10)	26.07(0.06)	27.73(0.06)
RAVK	16.14(0.11)	22.07(0.07)	24.28(0.08)	25.89(0.09)	27.44(0.14)

Overall, CMIh, which can be seen as a trade-off between k-nearest neighbour and histogram methods, performs well, both in terms of the accuracy of the estimator and in

terms of the time needed to compute this estimator. Among the pure k-nearest neighbour methods, MS, despite its limitations, remains the best one overall in our experiments in terms of accuracy. Its time complexity is similar to the ones of the other methods of the same class. The pure histogram method LH performs well in terms of accuracy of the estimator, but its computation time is prohibitive. Two methods thus stand out from our analysis, namely CMIh and MS.

4. Testing Conditional Independence

Once an estimator for mutual information has been computed, it is important to assess to which extent the obtained value is sufficiently different from or sufficiently close to 0 so as to conclude on the dependence or independence of the involved variables. To do so, one usually relies on statistical tests, among which permutation tests are widely adopted as they do not require any modelling assumption [38]. We also focus on such tests here which emulate the behaviour of the estimator under the null hypothesis (corresponding to independence) by permuting values of variables. Recently, Runge [39] showed that, for conditional tests and purely quantitative data, local permutations that break any possible dependence between X and Y while preserving the dependence between X and Z and between Y and Z are to be preferred over global permutations. Our contribution here is to extend this method to mixed data.

4.1. Local-Adaptive Permutation Test for Mixed Data

Let us consider a sample of independent realisations, denoted $(X_i, Y_i, Z_i)_{i=1,\dots,n}$, generated according to the distribution P_{XYZ} where X, Y and Z are multidimensional variables with quantitative and/or qualitative components. From this sample, one can compute an estimator, denoted $\widehat{I}(X;Y|Z)$, of the conditional mutual information using the hybrid method CMIh introduced in Section 3. In order to perform a permutation test, one needs to generate samples, under the null hypothesis, from the distribution $P_{X|Z}(x|z)P_{Y|Z}(y|z)P_Z(z)$. When the conditioning variable Z is qualitative, this boils down to randomly permuting the marginal sample of X while preserving the one of Y, conditionally to each possible value of Z [35]. In the quantitative case, one proceeds in a similar way and permutes the X values of the neighbours of each point i [35,39]. In our case, as the variable Z possibly contains quantitative and qualitative components, we propose to use an adaptive distance *dist* which corresponds to the absolute value if the component is quantitative and to the $(0-\infty)$ distance (which is 0 for identical values and ∞ for different values) if the component is qualitative. For $Z_i = (Z_i^1, \dots, Z_i^m)^T$ and $Z_j = (Z_j^1, \dots, Z_j^m)^T$ two realizations of the random vector Z, where m is the dimension of the data, the distance between these two points is then defined as:

$$D(Z_i, Z_j) = \max_{r \in \{1,\dots,m\}} dist(Z_i^r, Z_j^r).$$

The neighbourhood of Z_i consists in the set of k points closest to Z_i according to D. Using the same k for all observations may, however, be problematic since it is possible that the k^{th} closest point is at a distance ∞ of a given point Z_i when k is large. In such a case, all points are in the neighbourhood of Z_i. To avoid this, we adapt k to each observation using one parameter k_i for each observation Z_i: if Z is purely quantitative, then $k_i = k$, where k is a global hyper-parameter, otherwise $k_i = \min(k, n_i^\ell)$, where n_i^ℓ is the number of sample points which have the same qualitative values as Z_i.

Then, to generate a permuted sample, for each point i one permutes X_i with the X value of a randomly chosen point in the neighbourhood of i while preserving Y_i and Z_i: a permuted sample thus takes the form $(X_{\pi(i)}, Y_i, Z_i)_{i=1,\dots,n}$, where $\pi(i)$ is a random permutation over the neighbourhood of i. By construction, a permuted sample is drawn under the null hypothesis since the possible conditional dependence is broken by the permutation. Many permuted samples finally are created, from which one can compute CMIh estimators under the null hypothesis. Comparing theses estimators to the one of the original sample allows one to determine whether the null hypothesis can be rejected

or not [38]. Note that, in practice, the permutations are drawn with replacement [44]. The main steps of our local-adaptive permutation test are summarised in Algorithm 2.

Algorithm 2 Local-Adaptive permutation test

Input $(X_i, Y_i, Z_i)_{1\leq i\leq n}$ the data, B the number of permutations, *isCat* indexes of qualitative component, k the hyper-parameter;
Compute $\widehat{I}(X, Y|Z)$ from the original data;
Separate qualitative and quantitative components of Z as $(Z_i^t, Z_i^\ell)_{1\leq i\leq n}$ using *isCat*;
for $i \in \{1, \ldots, n\}$ **do**
  if $Z^\ell \neq \emptyset$ **then**
    $n_i^\ell = length(\{m \in \{1, \ldots, n\} : Z_m^\ell = Z_i^\ell\})$;
    $k_i = \min(k, n_i^\ell)$;
  end if
  Compute $d_i^{k_i}$, the distance from Z_i to its k_i-nearest-neighbour in Z by applying two different distances metrics, respectively, to two different types of components;
  $\mathcal{N}_i = \{j \in \{1, \ldots, n\} : ||Z_j - Z_i|| \leq d_i^{k_i}\}$;
end for
for $b \in \{1, \ldots, B\}$ **do**
  Generate a sample $(X_{\pi_b(i)}, Y_i, Z_i)_{1\leq i\leq n}$ locally permuted with respect to $(\mathcal{N}_i)_{1\leq i\leq n}$;
  Compute the associated estimator $\widehat{I}(X_{\pi_b}, Y|Z)$;
end for
Estimate the p-value as

$$\widehat{p} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\widehat{I}(X_{\pi_b}, Y|Z) \geq \widehat{I}(X,Y|Z)}; \tag{10}$$

Output The p-value $\hat{p}$.

4.2. Experimental Illustration

We first propose an extensive analysis on simulated data and then perform an analysis on a real world data set. We compare our test, denoted by LocAT, with two permutation tests: the first one is the local permutation test, denoted by LocT, designed initially for purely quantitative data proposed by Runge [39] and directly extended to mixed data using the $(0-\infty)$ distance for qualitative components; the second test is the global permutation test, denoted by GloT. For LocT and LocAT, we set the hyper-parameter k_{perm} to 5 as proposed by Runge [39]. For all tests, we set the number of permutation, B, to 1000. We study the behaviour of each test with respect to the two best estimators highlighted in Section 3, CMIh and MS. It is important to note here that, in order to be consistent with the parameters of the original method, for MS we use the $(0-1)$ distance in the qualitative component to compute the estimator rather than the $(0-\infty)$ distance as in the permutation method. We use rank transformation in each quantitative component which has the advantage of preserving the order and putting all quantitative components on the same scale (the "first" method is used to break potential ties).

4.2.1. Simulated Data

We consider here that X, Y and Z are uni-dimensional but all estimators and independence tests can be used when the variables are multi-dimensional, as illustrated in the experiments conducted on the real dataset. We furthermore focus on three classical structures of causal networks: the chain ($X \to Z \to Y$), the fork ($X \leftarrow Z \to Y$), and the collider ($X \to Z \leftarrow Y$). For the chain and the fork, X and Y are dependent and independent conditionally to Z; for the collider, X and Y are independent and dependent conditionally to Z. In the sequel, the qualitative variables or components with infinite possible values are treated as quantitative ones. For each structure, one can potentially distinguishes eight configurations, depending on the type, quantitative ('t') or qualitative ('ℓ'), of each of the three variables X, Y and Z. The configuration '$t\ell t$' corresponds for example to the situation

where X and Z are quantitative and Y qualitative. Note that as X and Y play a similar role, we only consider six cases. Details of the data generating process are provided in Appendix A. All samples contain 500 points.

The results obtained for each method, each structure and each configuration are reported in Table 2. For the chain and the fork, which are conditional independence structures, the acceptance rate corresponds to the percentage of the p-values that are above the thresholds 0.01 and 0.05 for 10 repetitions of each method in each configuration. For the collider, the acceptance rate corresponds to the percentage of the p-value that is under the thresholds 0.01 and 0.05 for 10 repetitions of each method in each configuration. In all cases, the closer the acceptance rate is to 1, the better.

Table 2. 0.01 and 0.05 threshold acceptance rates computed for the statistical test $H_0 = X \perp\!\!\!\perp Y|Z$ versus $H_1 = X \not\perp\!\!\!\perp Y|Z$ using the three tests LocT, LocAT and GloT on two estimators, CMIh and MS, on synthetic data. The number of sampling points is 500. Each acceptance rate is computed over 10 repetitions.

		CMIh-LocT		**CMIh-LocAT**		**CMIh-GloT**		**MS-LocT**		**MS-LocAT**		**MS-GloT**	
		0.01	**0.05**	**0.01**	**0.05**	**0.01**	**0.05**	**0.01**	**0.05**	**0.01**	**0.05**	**0.01**	**0.05**
Chain	$t\ell t$	1	1	1	1	0	0	1	1	1	1	1	1
	ttt	1	1	1	1	0	0	1	0.9	1	0.9	0	0
	$\ell\ell t$	1	0.9	1	0.9	1	0.8	1	1	1	1	1	1
	$t\ell\ell$	1	1	1	1	1	1	1	1	1	1	1	1
	$tt\ell$	0	0	0.8	0.4	0	0	0	0	0.5	0.3	0	0
	$\ell\ell\ell$	1	0.9	1	0.9	1	1	1	1	1	1	1	1
Fork	$t\ell t$	0.9	0.9	0.9	0.9	0	0	1	1	1	1	1	1
	ttt	1	1	1	1	0	0	1	1	1	1	0	0
	$\ell\ell t$	1	1	1	1	1	1	1	1	1	1	1	1
	$t\ell\ell$	1	1	1	0.9	1	1	1	1	1	1	1	1
	$tt\ell$	0	0	0.9	0.8	0	0	0	0	0.8	0.5	0	0
	$\ell\ell\ell$	1	1	1	1	1	1	1	0.9	1	1	1	1
Collider	$t\ell t$	1	1	1	1	1	1	0	0	0	0	0	0
	ttt	1	1	1	1	0.8	0.9	1	1	1	1	1	1
	$\ell\ell t$	1	1	1	1	1	1	0	0	0	0	0	0
	$t\ell\ell$	0	0	0.4	0.7	0	0	0	0	0	0	0	0
	$tt\ell$	0.6	1	1	1	0.2	0.4	0	0	0	0	0	0
	$\ell\ell\ell$	1	1	1	1	1	1	1	1	1	1	0.4	0.9

As one can note, the global test does not perform well in the configurations 'ttt' and '$tt\ell$' of the chain and fork structures, for both CMIh-GloT and MS-GloT. In addition, it does not perform well on the '$t\ell t$' configuration of the chain and fork structures for CMIh. It nevertheless performs well for CMIh on the collider structure over all configurations, but not for MS. Overall, its global performance is relatively poor compared to the two local tests LocT and LocAT. The local test LocT performs relatively well on the chain and fork structures for all configurations but '$tt\ell$'. It performs well for CMIh and the collider structure on all configurations but '$t\ell\ell$'; it does not, however, perform well for MS on this structure as only two configurations are correctly treated, 'ttt' and '$\ell\ell\ell$'. Finally, the local adaptive test, LocAT, performs well on all configurations of all structures for CMIh. For MS, it performs well on the chain and fork structures but not on the collider structure where the results are identical to the ones obtained with the standard local test LocT. Note that the bad results obtained for all tests with MS on the collider structure are directly related to the limitations pointed out in the previous section. Indeed, the estimator given by MS on all configurations but 'ttt' and '$\ell\ell\ell$' is close to 0 as $\rho_{k,i}/2 \geq D_\ell$ if there is at least one quantitative variable (due to rank transformation).

Overall, the combination CMIh with the test LocAT allows one to correctly identify the true (in)dependence relation on all configurations of all structures.

4.2.2. Real Data

We consider here three real datasets to illustrate the behaviour of our proposed estimator and test. Given the performance of the global permutation test on the simulated data, we do not use it here and compare four estimator–test combinations: CMIh-LocT, CMIh-LocAT, MS-LocT and MS-LocAT.

Preprocessed DWD Dataset

This climate dataset was originally provided by the Deutscher Wetterdienst (DWD) and preprocessed by Mooij et al. [45]. It contains 6 variables (altitude, latitude, longitude, and annual mean values of sunshine duration over the years 1961–1990, temperature and precipitation) collected from 349 weather stations in Germany. We focus here on three variables, *latitude*, *longitude* and *temperature*, this latter variable being discretized into three balanced classes (low, medium and high) in order to create a mixed dataset. The goal here is to identify one unconditional independence (Case 1) and one conditional dependence (Case 2):

- Case 1: *latitude* is unconditionally independent of *longitude* as the 349 weather stations are distributed irregularly on the map.
- Case 2: *latitude* is dependent of *longitude* given *temperature* as both *latitude* and *longitude* act on *temperature*: moving a thermometer towards the equator will generally result in an increased temperature, and climate in West Germany is more oceanic and less continental than in East Germany.

The p-value for each method is shown in Table 3. For Case 1, the p-value should be high so that the null hypothesis is not rejected, whereas it should be small for Case 2 as the correct hypothesis is H_1. Note that as there is no conditional variable in Case 1, the permutation tests LocT and LocAT give the same results.

Table 3. DWD: p-values for the different estimator–test combinations of the statistical test, which is $H_0 = X \perp\!\!\!\perp Y$ versus $H_1 = X \not\perp\!\!\!\perp Y$ for Case 1, where X and Y correspond to *latitude* and *longitude*, and $H_0 = X \perp\!\!\!\perp Y|Z$ versus $H_1 = X \not\perp\!\!\!\perp Y|Z$ for Case 2, where X, Y and Z correspond to *latitude*, *longitude* and *temperature*. The number of sampling points is 349.

	CMIh-LocT	**CMIh-LocAT**	**MS-LocT**	**MS-LocAT**
Case 1	0.05	0.05	0.03	0.03
Case 2	0	0	0.09	0.08

As one can note from Table 3, under both thresholds 0.01 and 0.05, CMIh-LocT and CMIh-LocAT succeed in giving the correct independent and dependent relations. In contrast, MS-LocT and MS-LocAT only identify the independent relation at the threshold 0.01 and never correctly identify the conditional dependency.

ADHD-200 Dataset

This dataset contains phenotypic data on kids with ADHD (Attention Deficit Hyperactivity Disorder) [46]. It contains 23 variables. We focus here on four variables: *gender*, *attention deficit level*, *hyperactivity/impulsivity level* and *medication status*, *gender* and *medication status* being binary categorical variables. The dataset contains 426 records after removing missing data. Following previous studies, we consider two independence relations:

- Case 1: *gender* is independent of *hyperactivity/impulsivity level* given *attention deficit level*, which has been confirmed by several studies [47,48].
- Case 2: *hyperactivity/impulsivity level* is independent of *medication status* given *attention deficit level*, which has been confirmed by Cui et al. [49].

The p-values obtained for the different estimator–test combinations are reported in Table 4. For this dataset, the p-values should be sufficiently high so that the null hypothesis is not rejected in both cases.

Table 4. ADHD-200: p-values for the different estimator–test combinations of the statistical test $H_0 = X \perp\!\!\!\perp Y|Z$ versus $H_1 = X \not\perp\!\!\!\perp Y|Z$ where X, Y and Z correspond to *gender*, *hyperactivity/impulsivity level* and *attention deficit level* for Case 1 and *hyperactivity/impulsivity level*, *medication status* and *attention deficit level* for Case 2. The number of sampling points is 426.

	CMIh-LocT	**CMIh-LocAT**	**MS-LocT**	**MS-LocAT**
Case 1	0.36	0.36	1	1
Case 2	0.17	0.19	1	1

As one can note, regardless of whether the threshold is 0.01 or 0.05, all four methods reach the correct conclusion in both cases. CMIh-LocT and CMIh-LocAT have the same performance as the conditional variable is quantitative. From the previous simulated experiments we can reasonably infer that these two could perform better if more records were collected. MS-LocT and MS-LocAT give a p-value of 1 because the conditional mutual information in MS is 0; we observe here the same degenerate behaviour for this estimator as the one discussed in Section 3.

EasyVista IT Monitoring System

This dataset consists of five time series collected from an IT monitoring system with a one minute sampling rate provided by EasyVista (https://www.easyvista.com/fr/produits/servicenav, accessed on 31 August 2022). We focus on five variables: *message dispatcher* (activity of a process that orient messages to other process with respect to different types of messages), which is a quantitative variable, *metric insertion* (activity of insertion of data in a database), which is also a quantitative variable, *status metric extraction* (status of activity of extraction of metrics from messages), which is a qualitative variable with three classes, namely normal ($\approx$ 75% of the observations), warning ($\approx$ 20% of the observations) and critical ($\approx$ 5% of the observations), *group history insertion* (activity of insertion of historical status in database), which is again a quantitative variable, and *collector monitoring information* (activity of updates in a given database) another quantitative variable. We know exact lags between variables, so we synchronise the data as a preprocessing step.

For this system we consider three cases:

- Case 1 represents a conditional independence between *message dispatcher* at time t and *metric insertion* at time t given *status metric extraction* at time t and *message dispatcher* and *metric insertion* at time $t-1$.
- Case 2 represents a conditional independence between *group history insertion* at time t, *collector monitoring information* at time t given *status metric extraction* at time t and *group history insertion* and *collector monitoring information* at time $t-1$.
- Case 3 represents a conditional dependence between *status metric extraction* at time t and *group history insertion* at time t given *status metric extraction* at time $t-1$.

For each case, we consider 12 datasets with 1000 observations each. The results, reported in Table 5, are based on the acceptance rates at thresholds 0.01 and 0.05 computed as in Section 4.2.1. Again, under each threshold, the closer the result is to 1, the better. Finally note that we conditioned on the past of each time series to eliminate the effect of the autocorrelation.

As one can see, CMIh-LocT and CMIh-LocAT yield exactly the same results on this dataset. Furthermore, the results obtained by these combinations are systematically better than the ones obtained when using MS as the estimator except for Case 2 with the threshold 0.05. However, on this case, all combinations correctly identify the conditional independence. Lastly, as before, MS yields poor results on Case 3, which corresponds to a collider structure. The explanation is the same as above for this structure and suggests that MS should not be used as an estimator to conditional mutual information.

Table 5. EasyVista: 0.01 and 0.05 threshold acceptance rates for the different estimator–test combinations computed for the statistical test $H_0 = X \perp\!\!\!\perp Y|Z$ versus $H_1 = X \not\perp\!\!\!\perp Y|Z$, where X, Y and Z correspond to *message dispatcher*$_t$, *metric insertion*$_t$ and the vector (*status metric extraction*$_t$, *message dispatcher*$_{t-1}$, *metric insertion*$_{t-1}$) for Case 1, to *group history insertion*$_t$, *collector monitoring information*$_t$ and the vector (*status metric extraction*$_t$, *group history insertion*$_{t-1}$, *collector monitoring information*$_{t-1}$) for Case 2 and *status metric extraction*$_t$, *group history insertion*$_t$ and *status metric extraction*$_{t-1}$ for Case 3. The number of sampling points is 1000. Each acceptance rate is computed over 12 datasets of the same structure.

	CMIh-LocT		**CMIh-LocAT**		**MS-LocT**		**MS-LocAT**	
	0.01	**0.05**	**0.01**	**0.05**	**0.01**	**0.05**	**0.01**	**0.05**
Case 1	1	0.75	1	0.75	0.67	0.58	0.75	0.58
Case 2	1	0.67	1	0.67	0.92	0.75	1	0.83
Case 3	0.75	0.83	0.75	0.83	0	0	0	0

Overall, the experiments on simulated and real datasets indicate that the combination CMIh-LocAT is robust to different structures and data types. This combination is well adapted to mixed data and provides the best results overall in our experiments.

5. Conclusions

We propose in this paper a novel hybrid method for estimating conditional mutual information in mixed data comprising both qualitative and quantitative variables. This method relies on two classical approaches to estimate conditional mutual information: k-nearest neighbour and histograms methods. A comparison of this hybrid method to previous ones illustrated its good behaviour, both in terms of accuracy of the estimator and in terms of the time required to compute it. We have furthermore proposed a local adaptive permutation test which allows one to accept or reject null hypotheses. This test is also particularly adapted to mixed data. Our experiments, conducted on both synthetic and real data sets, show that the combination of the hybrid estimator and the local adaptive test we have introduced is able, contrary to other combinations, to identify the correct conditional (in)dependence relations in a variety of cases involving mixed data. To the best of our knowledge, this combination is the first one fully adapted to mixed data. We believe that it will become a useful ingredient for researchers and practitioners for problems, including but not limited to (1) causal discovery where one aims to identify causal relations between variables of a given system by analyzing statistical properties of purely observational data, (2) graphical model inference where one aims to establish a graphical model which describes the statistical relationships between random variables and which can be used to compute the marginal distribution of one or several variables, and (3) feature selection where one aims to reduce the number of input variables by eliminating highly dependent ones.

Author Contributions: Writing—review & editing, L.Z., A.M., C.K.A., E.D. and E.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by MIAI@Grenoble Alpes grant number ANR-19-P3IA-0003.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data presented in this study are publicly available: Simulated data are available at https://github.com/leizan/CMIh2022 (accessed on 25 August 2022); Preprocessed DWD dataset are available at https://webdav.tuebingen.mpg.de/cause-effect/ (accessed on 22 August 2022); ADHD-200 data are available at http://fcon_1000.projects.nitrc.org/indi/adhd200/index.html (accessed on 21 August 2022); IT monitoring data are available at https://easyvista2015-my.

sharepoint.com/:f:/g/personal/aait-bachir_easyvista_com/ElLiNpfCkO1JgglQcrBPP9IBxBXzaINrM5f0ILz6wbgoEQ?e=OBTsUY (accessed on 1 July 2022).

Acknowledgments: We thank Ali Aït-Bachir and Christophe de Bignicourt from EasyVista for providing us with the IT monitoring dataset along with the expected independence and conditional independence between the underlying time series.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Generative Processes Used for the Different Configurations on the Three Structures Chain, Fork and Collider

We denote here by $Round(x)$ the function that rounds x to the nearest integer, and by $Mod(number, divisor)$ the function that returns the remainder of the integer division of 'number' by 'divisor'. Note that, for generality purposes, we use both quantitative distributions as well as qualitative distributions with infinite support to generate quantitative variables. We give below the generative process used for each configuration of each structure (chain, fork and collider).

Appendix A.1. Processes for Chain

For the configuration '$t\ell t$':

$$\begin{aligned} &X \sim \mathcal{U}([0,100]) \\ &Y \sim Bi(Abs(Round(Z)), 0.5) \\ &Z = X + \xi_Z \text{ where } \xi_Z \sim \mathcal{N}(0,1). \end{aligned}$$

For the configuration 'ttt':

$$\begin{aligned} &X \sim \mathcal{U}([0,100]) \\ &Y \sim Pois(Abs(Round(Z))) \\ &Z = X + \xi_Z \text{ where } \xi_Z \sim \mathcal{N}(0,1). \end{aligned}$$

For the configuration '$\ell\ell t$':

$$\begin{aligned} &X \in \Omega(X) = \{A_X, B_X, C_X, D_X, E_X\} \text{ with probability } (0.6, 0.1, 0.1, 0.1, 0.1) \\ &Y = r(Mod(Round(Z + \xi_Y), 4)) \text{ where } \xi_Y \sim \mathcal{N}(0,1) \\ &Z \sim \mathcal{N}(n(X), 2). \end{aligned}$$

and where the function n and r are defined by

$$\begin{aligned} n(A_X) &= 1, & r(0) &= B_Y, \\ n(B_X) &= 3, & r(1) &= A_Y, \\ n(C_X) &= 0, & r(2) &= C_Y, \\ n(D_X) &= 2, & r(3) &= D_Y, \\ n(E_X) &= 4. \end{aligned}$$

For the configuration '$t\ell\ell$':

$$\begin{aligned} &X \sim \mathcal{U}([0,100]) \\ &Y \sim Bi(Z, 0.5) \\ &Z \sim \mathcal{U}(\{Round(X), \ldots, Round(X) + 2\}). \end{aligned}$$

For the configuration '$tt\ell$':

$$X \sim \mathcal{U}([0,100])$$
$$Y \sim Pois(Z)$$
$$Z \sim \mathcal{U}(\{Round(X), \ldots, Round(X)+2\}).$$

For the configuration '$\ell\ell\ell$':

$X \in \Omega(X) = \{A_X, B_X, C_X, D_X, E_X\}$ with probability (0.6,0.1,0.1,0.1,0.1)
$Y \in \Omega(Y) = \{A_Y, B_Y, C_Y, D_Y, E_Y\}$
$Y|(Z = z) = t(z)$ with probability 0.9 and the other four realizations with probability 0.025
$Z \in \Omega(Z) = \{A_Z, B_Z, C_Z, D_Z, E_Z\}$
$Z|(X = x) = s(x)$ with probability 0.9 and the other four realizations with probability 0.025

and where the function s and t are defined by

$$\begin{aligned} s(A_X) &= B_Z, & t(A_Z) &= D_Y, \\ s(B_X) &= E_Z, & t(B_Z) &= E_Y, \\ s(C_X) &= A_Z, & t(C_Z) &= B_Y, \\ s(D_X) &= C_Z, & t(D_Z) &= A_Y, \\ s(E_X) &= D_Z, & t(E_Z) &= C_Y. \end{aligned}$$

Appendix A.2. Processes for Fork

For the configuration '$t\ell t$':

$$X = Z + \xi_X \text{ where } \xi_X \sim \mathcal{N}(0,1)$$
$$Y \sim Bi(Round(Z), 0.5)$$
$$Z \sim \mathcal{U}([0,100]).$$

For the configuration 'ttt':

$$X = Z + \xi_X \text{ where } \xi_X \sim \mathcal{N}(0,1)$$
$$Y \sim Pois(Round(Z))$$
$$Z \sim \mathcal{U}([0,100]).$$

For the configuration '$\ell\ell t$':

$$X = f(Mod(Round(Z + \xi_X), 4)) \text{ where } \xi_X \sim \mathcal{N}(0,1)$$
$$Y = g(Mod(Round(Z + \xi_Y), 3)) \text{ where } \xi_Y \sim \mathcal{N}(0,1)$$
$$Z \sim Exp(0.1).$$

and where the function f and g are defined by

$$\begin{aligned} f(0) &= C_X, & g(0) &= B_Y, \\ f(1) &= A_X, & g(1) &= A_Y, \\ f(2) &= D_X, & g(2) &= C_Y, \\ f(3) &= B_X. \end{aligned}$$

For the configuration '$t\ell\ell$':

$$X = Z + \xi_X \text{ where } \xi_X \sim \mathcal{N}(0,1)$$
$$Y \sim Bi(Z, 0.5)$$
$$Z \sim \mathcal{U}(\{0, \ldots, 100\}).$$

For the configuration '$tt\ell$':

$$X = Z + \xi_X \text{ where } \xi_X \sim \mathcal{N}(0,1)$$
$$Y \sim Pois(Z)$$
$$Z \sim \mathcal{U}(\{0,\ldots,100\}).$$

For the configuration '$\ell\ell\ell$':

$X \in \Omega(X) = \{A_X, B_X, C_X, D_X, E_X\}$
$X|(Z = z) = p(z)$ with probability 0.9 and the other four realizations with probability 0.025
$Y \in \Omega(Y) = \{A_Y, B_Y, C_Y, D_Y, E_Y\}$
$Y|(Z = z) = q(z)$ with probability 0.9 and the other four realizations with probability 0.025
$Z \in \Omega(Z) = \{A_Z, B_Z, C_Z, D_Z, E_Z\}$ with probability (0.6,0.1,0.1,0.1,0.1).

and where the function p and q are defined by

$$\begin{aligned} p(A_Z) &= C_X, & q(A_Z) &= D_Y, \\ p(B_Z) &= A_X, & q(B_Z) &= E_Y, \\ p(C_Z) &= D_X, & q(C_Z) &= B_Y, \\ p(D_Z) &= E_X, & q(D_Z) &= A_Y, \\ p(E_Z) &= B_X, & q(E_Z) &= C_Y. \end{aligned}$$

Appendix A.3. Processes for Collider

For the configuration '$t\ell t$':

$$X \sim \mathcal{N}(50, 25)$$
$$Y \sim Bi(100, 0.5)$$
$$Z = X + Y + \xi_Z \text{ where } \xi_Z \sim \mathcal{N}(0,1).$$

For the configuration 'ttt':

$$X \sim \mathcal{N}(50, 25)$$
$$Y \sim Pois(100)$$
$$Z = X + Y + \xi_Z \text{ where } \xi_Z \sim \mathcal{N}(0,1).$$

For the configuration '$\ell\ell t$':

$$X \in \Omega(X) = \{A_X, B_X, C_X, D_X, E_X\} \text{ with probability } (0.6, 0.1, 0.1, 0.1, 0.1)$$
$$Y \in \Omega(Y) = \{A_Y, B_Y, C_Y, D_Y, E_Y\} \text{ with probability } (0.6, 0.1, 0.1, 0.1, 0.1)$$
$$Z \sim \mathcal{N}(h(X,Y), 1).$$

and where the function h is defined by

$$\begin{aligned}
h(A_X, A_Y) &= 0, & h(A_X, B_Y) &= 1,\\
h(A_X, C_Y) &= 2, & h(A_X, D_Y) &= 3,\\
h(A_X, E_Y) &= 4, & h(B_X, A_Y) &= 5,\\
h(B_X, B_Y) &= 6, & h(B_X, C_Y) &= 7,\\
h(B_X, D_Y) &= 8, & h(B_X, E_Y) &= 9,\\
h(C_X, A_Y) &= 10, & h(C_X, B_Y) &= 11,\\
h(C_X, C_Y) &= 12, & h(C_X, D_Y) &= 13,\\
h(C_X, E_Y) &= 14, & h(D_X, A_Y) &= 15,\\
h(D_X, B_Y) &= 16, & h(D_X, C_Y) &= 17,\\
h(D_X, D_Y) &= 18, & h(D_X, E_Y) &= 19,\\
h(E_X, A_Y) &= 20, & h(E_X, B_Y) &= 21,\\
h(E_X, C_Y) &= 22, & h(E_X, D_Y) &= 23,\\
h(E_X, E_Y) &= 24.
\end{aligned}$$

For the configuration '$t\ell\ell$':

$$\begin{aligned}
&X \sim \mathcal{N}(50, 50)\\
&Y \sim Bi(100, 0.5)\\
&Z \sim Bi(Abs(Round(X + Y + \xi_Z)), 0.5) \text{ where } \xi_Z \sim \mathcal{N}(0, 1).
\end{aligned}$$

For the configuration '$tt\ell$':

$$\begin{aligned}
&X \sim \mathcal{N}(50, 50)\\
&Y \sim Pois(100)\\
&Z \sim Bi(Abs(Round(X + Y + \xi_Z)), 0.5) \text{ where } \xi_Z \sim \mathcal{N}(0, 1).
\end{aligned}$$

For the configuration '$\ell\ell\ell$':

$X \in \Omega(X) = \{A_X, B_X, C_X, D_X, E_X\}$ with probability (0.6,0.1,0.1,0.1,0.1)
$Y \in \Omega(Y) = \{A_Y, B_Y, C_Y, D_Y, E_Y\}$ with probability (0.6,0.1,0.1,0.1,0.1)
$Z \in \Omega(Z) = \{A_Z, B_Z, C_Z, D_Z, E_Z\}$
$Z|(X = x, Y = y) = m(x, y)$ with probability 0.9 and the other four realizations with probability 0.025

and where the function m is defined by

$$\begin{aligned}
m(A_X, A_Y) = m(A_X, B_Y) = m(A_X, C_Y) = m(E_X, D_Y) = m(E_X, E_Y) &= A_Z,\\
m(A_X, D_Y) = m(A_X, E_Y) = m(B_X, A_Y) = m(B_X, B_Y) = m(B_X, C_Y) &= B_Z,\\
m(B_X, D_Y) = m(B_X, E_Y) = m(C_X, A_Y) = m(C_X, B_Y) = m(C_X, C_Y) &= C_Z,\\
m(C_X, D_Y) = m(C_X, E_Y) = m(D_X, A_Y) = m(D_X, B_Y) = m(D_X, C_Y) &= D_Z,\\
m(D_X, D_Y) = m(D_X, E_Y) = m(E_X, A_Y) = m(E_X, B_Y) = m(E_X, C_Y) &= E_Z.
\end{aligned}$$

References

1. Spirtes, P.; Glymour, C.N.; Scheines, R.; Heckerman, D. *Causation, Prediction, and Search*; MIT Press: Cambridge, MA, USA, 2000.
2. Whittaker, J. *Graphical Models in Applied Multivariate Statistics*; Wiley Publishing: New York, NY, USA, 2009.
3. Vinh, N.; Chan, J.; Bailey, J. Reconsidering mutual information based feature selection: A statistical significance view. In Proceedings of the AAAI Conference on Artificial Intelligence, Quebec City, QC, Canada, 27–31 July 2014; Volume 28.
4. Thomas, M.; Joy, A.T. *Elements of Information Theory*; Wiley-Interscience: New York, NY, USA, 2006.
5. Székely, G.J.; Rizzo, M.L.; Bakirov, N.K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **2007**, *35*, 2769–2794.

6. Gretton, A.; Bousquet, O.; Smola, A.; Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In Proceedings of the International Conference on Algorithmic Learning Theory, Singapore, 8–11 October 2005; pp. 63–77.
7. Gretton, A.; Smola, A.; Bousquet, O.; Herbrich, R.; Belitski, A.; Augath, M.; Murayama, Y.; Pauls, J.; Schölkopf, B.; Logothetis, N. Kernel constrained covariance for dependence measurement. In Proceedings of the International Workshop on Artificial Intelligence and Statistics, Hastings, Barbados, 6–8 January 2005; pp. 112–119.
8. Póczos, B.; Ghahramani, Z.; Schneider, J. Copula-based kernel dependency measures. *arXiv* **2012**, arXiv:1206.4682.
9. Berrett, T.B.; Samworth, R.J. Nonparametric independence testing via mutual information. *Biometrika* **2019**, *106*, 547–566.
10. Wyner, A.D. A definition of conditional mutual information for arbitrary ensembles. *Inf. Control.* **1978**, *38*, 51–59.
11. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 623–656.
12. Frenzel, S.; Pompe, B. Partial Mutual Information for Coupling Analysis of Multivariate Time Series. *Phys. Rev. Lett.* **2007**, *99*, 204101.
13. Vejmelka, M.; Paluš, M. Inferring the directionality of coupling with conditional mutual information. *Phys. Rev. E* **2008**, *77*, 026214.
14. Scott, D.W. *Multivariate Density Estimation: Theory, Practice, and Visualization*; John Wiley & Sons: New York, NY, USA, 2015.
15. Cabeli, V.; Verny, L.; Sella, N.; Uguzzoni, G.; Verny, M.; Isambert, H. Learning clinical networks from medical records based on information estimates in mixed-type data. *PLoS Comput. Biol.* **2020**, *16*, e1007866.
16. 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), SIAM, Virtual Event, 29 April–1 May 2021; pp. 387–395.
17. Beirlant, J.; Dudewicz, E.J.; Györfi, L.; Van der Meulen, E.C. Nonparametric entropy estimation: An overview. *Int. J. Math. Stat. Sci.* **1997**, *6*, 17–39.
18. Kozachenko, L.F.; Leonenko, N.N. Sample estimate of the entropy of a random vector. *Probl. Peredachi Informatsii* **1987**, *23*, 9–16.
19. Singh, H.; Misra, N.; Hnizdo, V.; Fedorowicz, A.; Demchuk, E. Nearest neighbor estimates of entropy. *Am. J. Math. Manag. Sci.* **2003**, *23*, 301–321.
20. Singh, S.; Póczos, B. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 5–10 December 2016; Volume 29.
21. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138.
22. Ross, B.C. Mutual Information between Discrete and Continuous Data Sets. *PLoS ONE* **2014**, *9*, e87357.
23. Gao, W.; Kannan, S.; Oh, S.; Viswanath, P. Estimating mutual information for discrete-continuous mixtures. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
24. Rahimzamani, A.; Asnani, H.; Viswanath, P.; Kannan, S. Estimators for multivariate information measures in general probability spaces. In Proceedings of the Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Montreal, QC, Canada, 3–8 December 2018; Volume 31.
25. Mesner, O.C.; Shalizi, C.R. Conditional Mutual Information Estimation for Mixed, Discrete and Continuous Data. *IEEE Trans. Inf. Theory* **2020**, *67*, 464–484.
26. Ahmad, A.; Khan, S.S. Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access* **2019**, *7*, 31883–31902.
27. Mukherjee, S.; Asnani, H.; Kannan, S. CCMI: Classifier based conditional mutual information estimation. In Proceedings of the 35th Uncertainty in Artificial Intelligence Conference, Tel Aviv, Israel, 22–25 July 2020; pp. 1083–1093.
28. Mondal, A.; Bhattacharjee, A.; Mukherjee, S.; Asnani, H.; Kannan, S.; Prathosh, A. C-MI-GAN: Estimation of conditional mutual information using minmax formulation. In Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), Virtual, 3–6 August 2020; pp. 849–858.
29. Meynaoui, A. New Developments around Dependence Measures for Sensitivity Analysis: Application to Severe Accident Studies for Generation IV Reactors. Ph.D. Thesis, INSA de Toulouse, Toulouse, France, 2019.
30. Shah, R.D.; Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *Ann. Stat.* **2020**, *48*, 1514–1538.
31. Fukumizu, K.; Gretton, A.; Sun, X.; Schölkopf, B. Kernel measures of conditional dependence. In Proceedings of the Advances in Neural Information Processing Systems 20 (NIPS 2007), Vancouver, BC, Canada, 3–6 December 2007; Volume 20.
32. 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, UAI'11, Barcelona Spain, 14–17 July 2011; pp. 804–813.
33. Strobl, E.V.; Zhang, K.; Visweswaran, S. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *J. Causal Inference* **2019**, *7*. https://doi.org/10.1515/jci-2018-0017.
34. Zhang, Q.; Filippi, S.; Flaxman, S.; Sejdinovic, D. Feature-to-Feature Regression for a Two-Step Conditional Independence Test. In Proceedings of the Association for Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, 11–15 August 2017.
35. Doran, G.; Muandet, K.; Zhang, K.; Schölkopf, B. A Permutation-Based Kernel Conditional Independence Test. In Proceedings of the Association for Uncertainty in Artificial Intelligence UAI, Quebec City, QC, Canada, 23–27 July 2014; pp. 132–141.
36. Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Schölkopf, B.; Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **2012**, *13*, 723–773.

37. Tsagris, M.; Borboudakis, G.; Lagani, V.; Tsamardinos, I. Constraint-based causal discovery with mixed data. *Int. J. Data Sci. Anal.* **2018**, *6*, 19–30.
38. Berry, K.J.; Johnston, J.E.; Mielke, P.W. Permutation statistical methods. In *The Measurement of Association*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 19–71.
39. Runge, J. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics 2018, Lanzarote, Spain, 9–11 April 2018; pp. 938–947.
40. Manoukian, E.B. *Mathematical Nonparametric Statistics*; Taylor & Francis: Tokyo, Japan, 2022.
41. Antos, A.; Kontoyiannis, I. Estimating the entropy of discrete distributions. In Proceedings of the IEEE International Symposium on Information Theory 2001, Washington, DC, USA, 24–29 June 2001; pp. 45–45.
42. Vollmer, M.; Rutter, I.; Böhm, K. On Complexity and Efficiency of Mutual Information Estimation on Static and Dynamic Data. In Proceedings of the EDBT, Vienna, Austria, 26–29 March 2018; pp. 49–60.
43. Bentley, J.L. Multidimensional binary search trees used for associative searching. *Commun. ACM* **1975**, *18*, 509–517.
44. Romano, J.P.; Wolf, M. Exact and approximate stepdown methods for multiple hypothesis testing. *J. Am. Stat. Assoc.* **2005**, *100*, 94–108.
45. Mooij, J.M.; Peters, J.; Janzing, D.; Zscheischler, J.; Schölkopf, B. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *J. Mach. Learn. Res.* **2016**, *17*, 1103–1204.
46. Cao, Q.; Zang, Y.; Sun, L.; Sui, M.; Long, X.; Zou, Q.; Wang, Y. Abnormal neural activity in children with attention deficit hyperactivity disorder: A resting-state functional magnetic resonance imaging study. *Neuroreport* **2006**, *17*, 1033–1036.
47. Bauermeister, J.J.; Shrout, P.E.; Chávez, L.; Rubio-Stipec, M.; Ramírez, R.; Padilla, L.; Anderson, A.; García, P.; Canino, G. ADHD and gender: Are risks and sequela of ADHD the same for boys and girls? *J. Child Psychol. Psychiatry* **2007**, *48*, 831–839.
48. Willcutt, E.G.; Pennington, B.F.; DeFries, J.C. Etiology of inattention and hyperactivity/impulsivity in a community sample of twins with learning difficulties. *J. Abnorm. Child Psychol.* **2000**, *28*, 149–159.
49. Cui, R.; Groot, P.; Heskes, T. Copula PC algorithm for causal discovery from mixed data. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Riva del Garda, Italy, 19–23 September 2016; pp. 377–392.