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] **Eric Gaussier** [1]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
[2] EasyVista
{lei.zan, anouar.meynaoui, emilie.devijver, eric.gaussier}@univ-grenoble-alpes.fr, kassaad@easyvista.com

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.

Introduction

Measuring the (in)dependence between random variables from data when the underlying joint distribution is unknown plays a key role in several settings (Spirtes et al. 2000; Whittaker 2009; Vinh, Chan, and Bailey 2014). Many dependence measures have been introduced in the literature to quantify the dependence between random variables, as *Mutual Information* (MI) (Thomas and Joy 2006), *distance correlation* (Székely, Rizzo, and Bakirov 2007), *Hilbert–Schmidt Independence Criterion* (HSIC) (Gretton et al. 2005a), *COnstrained COvariance* (COCO) (Gretton et al. 2005b) or copula-based approaches (Póczos, Ghahramani, and Schneider 2012). 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 (Berrett and Samworth 2019), and consider both quantitative and qualitative variables.

The conditional mutual information (Wyner 1978) between two quantitative random variables X and Y conditionally to a quantitative random variable Z is given by:

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

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. It 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 (Frenzel and Pompe 2007; Vejmelka and Paluš 2008). The case of mixed datasets comprising both quantitative and qualitative variables is however less studied even though mixed data are ubiquitous. The aim of this paper is to present a new statistical method to detect conditional (in)dependence for mixed data.

For clarity, we summarize our contributions as follows:

- We propose a novel method, called CMIh, to estimate conditional mutual information taking advantages of the previously proposed approaches for qualitative and quantitative data.
- We introduce a new local permutation test, called LocAT for local adaptive test, which is well adapted to mixed data.
- We demonstrate the good behaviour of CMIh and LocAT on both synthetic and real data sets, and show their respective abilities to accurately estimate conditional mutual information and to detect conditional (in)dependence for mixed data.

Related work

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

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 (Scott 2015). To efficiently generate adaptive histograms model from data, (Cabeli et al. 2020) and (Marx, Yang, and van Leeuwen 2021) transform the problem into a model selection problem, using the minimum description length (MDL) principle. These approaches 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 (Beirlant et al. 1997) 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 does not require extensive tuning of kernel bandwidths.

Using nearest neighbours of observations to estimate the entropy dates back to (Kozachenko and Leonenko 1987), which was then generalized to a k-nearest neighbour (kNN) approach by (Singh et al. 2003). Later, (Kraskov, Stögbauer, and Grassberger 2004) 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 2007). The resulting model, called FP, however only deals with quantitative data.

Based on *graph divergence measure* (Rahimzamani et al. 2018) extended the estimator proposed in (Gao et al. 2017) to conditional mutual information, leading to a method called RAVK. Even more recently, (Mesner and Shalizi 2020) extended FP (Frenzel and Pompe 2007) to the mixed data case by introducing a new distance for non-quantitative variables which is either 0 for two identical points or 1 for points with different values. We refer to this method as MS. However, the choice of the qualitative and quantitative distances is a crucial point in MS (Ahmad and Khan 2019).

We also want to mention the proposal made by (Mukherjee, Asnani, and Kannan 2020) of a two-stage estimator based on generative models and classifiers as well as the refinement introduced in (Mondal et al. 2020) 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.

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.*

Kernel-based tests are known for their capability to deal with nonlinearity and high dimensions. One representative of this test category is the kernel conditional independence test (KCIT) proposed by (Zhang et al. 2011). Then, (Strobl, Zhang, and Visweswaran 2019) reduced its computational complexity. (Doran et al. 2014) also proposed a kernel conditional independence permutation test. 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. (Tsagris et al. 2018) 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.

More recently, (Shah and Peters 2020) 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. 2018) 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 (Berry, Johnston, and Mielke 2018) 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 2018) proposed a local permutation test in which permutations within X are done within similar values of Z. We extend in this paper this test, designed for quantitative data, to the mixed data case.

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 (Kraskov, Stögbauer, and Grassberger 2004; Frenzel and Pompe 2007), which has been mostly used on quantitative variables, or on histograms (Cabeli et al. 2020; Marx, Yang, and van Leeuwen 2021), 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:

$$\begin{aligned} I(X;Y|Z) =& H(X^t, Z^t|X^\ell, Z^\ell) + H(Y^t, Z^t|Y^\ell, Z^\ell) \\ & - H(X^t, Y^t, Z^t|X^\ell, Y^\ell, Z^\ell) - H(Z^t|Z^\ell) \\ & + H(X^\ell, Z^\ell) + H(Y^\ell, Z^\ell) - H(X^\ell, Y^\ell, Z^\ell) \\ & - H(Z^\ell). \end{aligned} \tag{2}$$

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.

Proposed hybrid estimator

Let us now consider an independently and identically distributed sample of size n denoted $(X_i, Y_i, Z_i)_{i=1,\ldots,n}$. We estimate the qualitative entropy terms of Equation (2), 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:

$$\begin{aligned} H(X^\ell, Z^\ell) &= -\mathbb{E}\left[\log P_{X^\ell,Z^\ell}(X^\ell, Z^\ell)\right] \\ &= -\sum_{\substack{x^\ell\in\Omega(X^\ell)\\ z^\ell\in\Omega(Z^\ell)}} P_{X^\ell,Z^\ell}(x^\ell, z^\ell)\log\left(P_{X^\ell,Z^\ell}(x^\ell, z^\ell)\right), \end{aligned}$$

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:

$$\widehat{P}_{X^\ell,Z^\ell}(x^\ell, z^\ell) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\left\{(X_i^\ell, Z_i^\ell)=(x^\ell, z^\ell)\right\}}, \tag{3}$$

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

$$\widehat{H}(X^\ell, Z^\ell) = -\sum_{\substack{x^\ell\in\Omega(X^\ell)\\ z^\ell\in\Omega(Z^\ell)}} \widehat{P}_{X^\ell,Z^\ell}(x^\ell, z^\ell)\log\left(\widehat{P}_{X^\ell,Z^\ell}(x^\ell, z^\ell)\right). \tag{4}$$

Let us now turn to the conditional entropies of Equation (2) 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:

$$\begin{aligned} H(X^t, Z^t|X^\ell, Z^\ell) = \sum_{\substack{x^\ell\in\Omega(X^\ell)\\ z^\ell\in\Omega(Z^\ell)}} & H(X^t, Z^t|X^\ell = x^\ell, Z^\ell = z^\ell) \\ & P_{X^\ell,Z^\ell}(x^\ell, z^\ell). \end{aligned} \tag{5}$$

As before, the probabilities involved in Equation (5) 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 (Singh et al. 2003) 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:

$$\begin{aligned} \widehat{H}(X^t, Z^t|X^\ell = x^\ell, Z^\ell = z^\ell) =& \psi(n_{xz}) - \psi(k_{xz}) + \log\left(v_{d_{xz}}\right) \\ &+ \frac{d_{xz}}{n_{xz}}\sum_{i=1}^{n_{xz}} \log\xi_{xz}(i), \end{aligned} \tag{6}$$

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 2018) which showed that this value behaves well in practice. As originally proposed in (Kraskov, Stögbauer, and Grassberger 2004) and adopted in subsequent studies, we rely on the ℓ_∞-distance which is associated with the maximum norm: for a vector $w = (w_1, \ldots, w_m)$ in $\mathbb{R}^m$, $\|w\|_\infty = \max(|w_1|, \ldots, |w_m|)$. 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:

$$\begin{aligned} \widehat{I}(X;Y|Z) = & \widehat{H}(X^t, Z^t|X^\ell, Z^\ell) + \widehat{H}(Y^t, Z^t|Y^\ell, Z^\ell) \\ & - \widehat{H}(X^t, Y^t, Z^t|X^\ell, Y^\ell, Z^\ell) - \widehat{H}(Z^t|Z^\ell) \\ & + \widehat{H}(X^\ell, Z^\ell) + \widehat{H}(Y^\ell, Z^\ell) - \widehat{H}(X^\ell, Y^\ell, Z^\ell) \\ & - \widehat{H}(Z^\ell), \end{aligned} \tag{7}$$

where the different terms are obtained through Equations (3), (4), (5) and (6). Notice that all the volume-type terms, as for the $\log(v_{d_{xz}})$ term in Equation (6), are canceled out in Equation (7). Indeed, it is well known that the volume of the unit ball in $\mathbb{R}^p$ with respect to the maximum norm is 2^p and this leads to the following plain equation:

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

Remarks. 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 done 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 Eq. 7 or are involved in entropy terms estimated through Eq. 4. Because of that, we can dispense with defining a distance on qualitative components, which is tricky as illustrated in the experimental subsection that follows.

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

Theorem 0.1. *Let (X, Y, Z) be a qualitative-quantitative mixed random vector. The estimator $\widehat{I}(X;Y|Z)$ defined in Equation (7) 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 (Manoukian 2022). It remains to show the consistency of each term in the right-hand side of Equation (7). 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 (Antos and Kontoyiannis 2001). 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 (Vollmer, Rutter, and Böhm 2018) for more details. In other words,

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

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. □

Experimental illustration

We compare in this section our estimator, CMIh, with several estimators mentioned before, namely FP (Frenzel and Pompe 2007), MS (Mesner and Shalizi 2020), RAVK (Rahimzamani et al. 2018), and LH (Marx, Yang, and van Leeuwen 2021). FP, MS and RAVK are methods based on the k-nearest neighbour approach. As for CMIh, the hyperparameter 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 (Mesner and Shalizi 2020). Laslty, for FP, which was designed for quantitative data, we set the minimum value of $n_{FP,W,i}$ to 1 to avoid $n_{FP,W,i} = 0$, which is an invalid value for the estimator. Moreover, LH is a histogram method based on MDL (Marx, Yang, and van Leeuwen 2021). 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 (Marx, Yang, and van Leeuwen 2021; Mesner and Shalizi 2020), 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 (Marx, Yang, and van Leeuwen 2021; Gao et al. 2017; Mesner and Shalizi 2020). The last four ones shed additional light on the different methods. Note that, as we reuse here the settings defined in (Marx, Yang, and van Leeuwen 2021; Gao et al. 2017; Mesner and Shalizi 2020), 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 $Y|X = x \sim 0.15\delta_0 + 0.85 Pois(x)$. 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

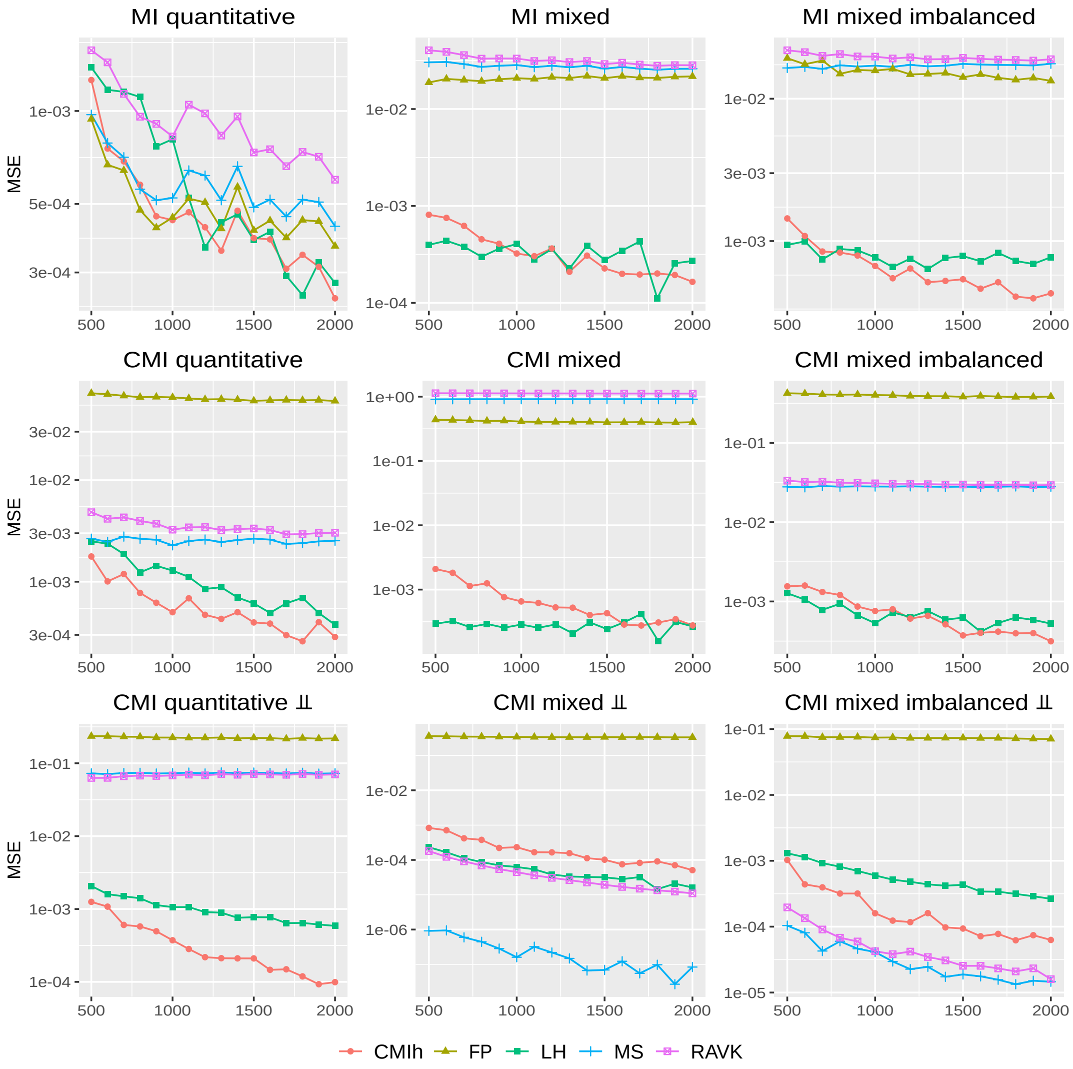

Figure 1: *Synthetic data with known ground truth.* MSE (on a log-scale) of each method with respect to the sample size (in abscissa) over the nine settings retained.

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 (Mesner and Shalizi 2020), 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 one and only one of the variables X,Y is quantitative and the others are qualitative, e.g. X is quantitative, Y and Z are qualitative (it is the same result that Y is quantitative and X and Z are qualitative) and 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, $k_{MS,XYZ,i} \leq n_{MS,XZ,i}$; furthermore, $n_{MS,XZ,i} \leq k_{MS,XYZ,i}$ 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.

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 $2,000$ 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 2020), 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.

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.

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):

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right),\ X_2 \sim \mathcal{U}(\{0,\dots,4\}),$$

$Y_2|X_2 = x_2 \sim \mathcal{U}([x_2, x_2+2]),\ X_3 \sim Exp(1)$ and $Y_3|X_3 = x_3 \sim 0.15\delta_0 + 0.85Pois(x_3)$. 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 $2,000$. 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 $1,250$. FP fails here to provide a reasonable estimate.

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.

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 (Berry, Johnston, and Mielke 2018). 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 2018) 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.

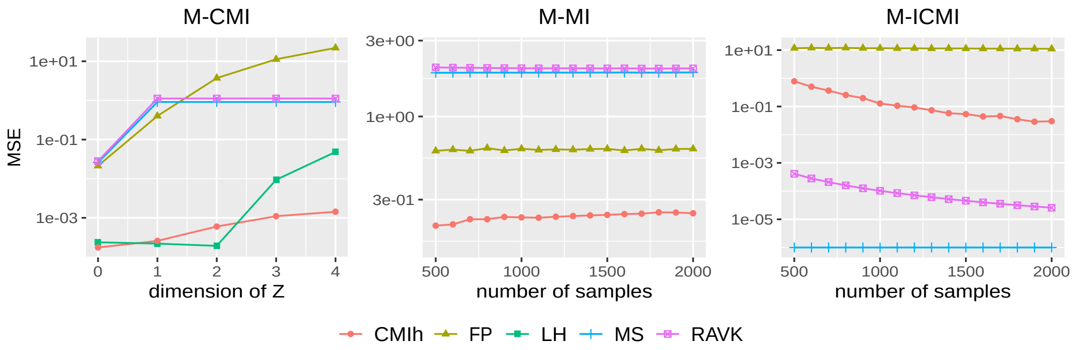

Figure 2: *Sensitivity to dimensionality* Left: MSE (on a log-scale) of each method for the multidimensional conditional mutual information (M-CMI) when increasing the dimension (x-axis) of the conditional variable from 0 to 4; the sample size is fixed to 2,000. Middle: MSE (on a log-scale) of each method 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.

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)

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 2 000.

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. 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 (Doran et al. 2014). In the quantitative case, one proceeds in a similar way and permutes the X values of the neighbours of each point i (Runge 2018; Doran et al. 2014). 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 hyperparameter, 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 (Berry, Johnston, and Mielke 2018). Note that, in practice, the permutations are drawn with replacement (Romano and Wolf 2005).

Experimental illustration

We directly perform an analysis on real world data sets (Links are available in Appendix). We compare our test, denoted by LocAT, with the local permutation test, denoted by LocT, designed initially for purely quantitative data proposed by (Runge 2018) and directly extended to mixed data using the $(0-\infty)$ distance for qualitative components. For LocT and LocAT, we set the hyper-parameter k_{perm} to 5 as proposed by (Runge 2018). 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 the previous section, CMIh and MS. MS-G is not considered here, as most random variables in experiment settings are unidimensional, so MS-G makes no difference with MS. Totally, we have four estimator-test combinations: CMIh-LocT, CMIh-LocAT, MS-LocT and MS-LocAT. 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). We consider here two real datasets to illustrate the behaviour of our proposed estimator and test.

Preprocessed DWD dataset This climate dataset was originally provided by the Deutscher Wetterdienst (DWD) and preprocessed by (Mooij et al. 2016). 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 2. 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.

As one can note from Table 2, 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.

	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

Table 2: 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.

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 the company EasyVista. 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 3, are based on the acceptance rates at thresholds 0.01 and 0.05. For conditional independent cases, 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 conditional dependent case, 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. 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 re-

	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

Table 3: 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 $\textit{message dispatcher}_t$, $\textit{metric insertion}_t$ and the vector $(\textit{status metric extraction}_t, \textit{message dispatcher}_{t-1}, \textit{metric insertion}_{t-1})$ for Case 1, to $\textit{group history insertion}_t$, $\textit{collector monitoring information}_t$ and the vector $(\textit{status metric extraction}_t, \textit{group history insertion}_{t-1}, \textit{collector monitoring information}_{t-1})$ for Case 2 and $\textit{status metric extraction}_t$, $\textit{group history insertion}_t$ and $\textit{status metric extraction}_{t-1}$ for Case 3. The number of sampling points is 1000. Each acceptance rate is computed over 12 datasets of the same structure.

sults 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.

Overall, the experiments on synthetic 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.

Conclusion

We have proposed 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.

Data availability:

- Synthetic data are available at https://github.com/leizan/CMIh2022;
- Preprocessed DWD dataset are available at https://webdav.tuebingen.mpg.de/cause-effect/;
- IT monitoring data are available at https://easyvista2015-my.sharepoint.com/:f:/g/personal/aait-bachir_easyvista_com/ElLiNpfCkO1JgglQcrBPP9IBxBXzaINrM5f0ILz6wbgoEQ?e=OBTsUY.

Acknowledgments

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

References

Ahmad, A.; and Khan, S. S. 2019. Survey of state-of-the-art mixed data clustering algorithms. *Ieee Access*, 7: 31883–31902.

Antos, A.; and Kontoyiannis, I. 2001. Estimating the entropy of discrete distributions. In *IEEE International Symposium on Information Theory*, 45–45.

Beirlant, J.; Dudewicz, E. J.; Györfi, L.; Van der Meulen, E. C.; et al. 1997. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1): 17–39.

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

Berry, K. J.; Johnston, J. E.; and Mielke, P. W. 2018. Permutation statistical methods. In *The Measurement of Association*, 19–71. Springer.

Cabeli, V.; Verny, L.; Sella, N.; Uguzzoni, G.; Verny, M.; and Isambert, H. 2020. Learning clinical networks from medical records based on information estimates in mixed-type data. *PLOS Computational Biology*, 16(5): 1–19.

Doran, G.; Muandet, K.; Zhang, K.; and Schölkopf, B. 2014. A Permutation-Based Kernel Conditional Independence Test. In *UAI*, 132–141. Citeseer.

Frenzel, S.; and Pompe, B. 2007. Partial Mutual Information for Coupling Analysis of Multivariate Time Series. *Physical review letters*, 99: 204101.

Gao, W.; Kannan, S.; Oh, S.; and Viswanath, P. 2017. Estimating mutual information for discrete-continuous mixtures. *Advances in neural information processing systems*, 30.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005a. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, 63–77. Springer.

Gretton, A.; Smola, A.; Bousquet, O.; Herbrich, R.; Belitski, A.; Augath, M.; Murayama, Y.; Pauls, J.; Schölkopf, B.; and Logothetis, N. 2005b. Kernel constrained covariance for dependence measurement. In *International Workshop on Artificial Intelligence and Statistics*, 112–119. PMLR.

Kozachenko, L. F.; and Leonenko, N. N. 1987. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2): 9–16.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Physical review E*, 69(6): 066138.

Manoukian, E. B. 2022. *Mathematical nonparametric statistics*. Taylor & Francis.

Marx, A.; Yang, L.; and van Leeuwen, M. 2021. 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)*, 387–395. SIAM.

Mesner, O. C.; and Shalizi, C. R. 2020. Conditional Mutual Information Estimation for Mixed, Discrete and Continuous Data. *IEEE Transactions on Information Theory*, 67(1): 464–484.

Mondal, A.; Bhattacharjee, A.; Mukherjee, S.; Asnani, H.; Kannan, S.; and Prathosh, A. 2020. C-MI-GAN: Estimation of conditional mutual information using minmax formulation. In *Conference on Uncertainty in Artificial Intelligence*, 849–858. PMLR.

Mooij, J. M.; Peters, J.; Janzing, D.; Zscheischler, J.; and Schölkopf, B. 2016. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research*, 17(32): 1–102.

Mukherjee, S.; Asnani, H.; and Kannan, S. 2020. CCMI: Classifier based conditional mutual information estimation. In *Uncertainty in artificial intelligence*, 1083–1093. PMLR.

Póczos, B.; Ghahramani, Z.; and Schneider, J. 2012. Copula-based kernel dependency measures. *arXiv preprint arXiv:1206.4682*.

Rahimzamani, A.; Asnani, H.; Viswanath, P.; and Kannan, S. 2018. Estimators for multivariate information measures in general probability spaces. *Advances in Neural Information Processing Systems*, 31.

Romano, J. P.; and Wolf, M. 2005. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469): 94–108.

Runge, J. 2018. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, 938–947. PMLR.

Scott, D. W. 2015. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Shah, R. D.; and Peters, J. 2020. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3): 1514–1538.

Singh, H.; Misra, N.; Hnizdo, V.; Fedorowicz, A.; and Demchuk, E. 2003. Nearest neighbor estimates of entropy. *American journal of mathematical and management sciences*, 23(3-4): 301–321.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

Strobl, E. V.; Zhang, K.; and Visweswaran, S. 2019. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1).

Székely, G. J.; Rizzo, M. L.; and Bakirov, N. K. 2007. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6): 2769–2794.

Thomas, M.; and Joy, A. T. 2006. *Elements of information theory*. Wiley-Interscience.

Tsagris, M.; Borboudakis, G.; Lagani, V.; and Tsamardinos, I. 2018. Constraint-based causal discovery with mixed data. *International journal of data science and analytics*, 6(1): 19–30.

Vejmelka, M.; and Paluš, M. 2008. Inferring the directionality of coupling with conditional mutual information. *Physical Review E*, 77(2): 026214.

Vinh, N.; Chan, J.; and Bailey, J. 2014. Reconsidering mutual information based feature selection: A statistical significance view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Vollmer, M.; Rutter, I.; and Böhm, K. 2018. On Complexity and Efficiency of Mutual Information Estimation on Static and Dynamic Data. In *EDBT*, 49–60.

Whittaker, J. 2009. *Graphical models in applied multivariate statistics*. Wiley Publishing.

Wyner, A. D. 1978. A definition of conditional mutual information for arbitrary ensembles. *Information and Control*, 38(1): 51–59.

Zhang, K.; Peters, J.; Janzing, D.; and Schölkopf, B. 2011. Kernel-Based Conditional Independence Test and Application in Causal Discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, 804–813. Arlington, Virginia, USA: AUAI Press. ISBN 9780974903972.