`fastMI`: A fast and consistent copula-based nonparametric estimator of mutual information

Soumik Purkayastha[1], Peter X.-K. Song [*,1]

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

ARTICLE INFO

ABSTRACT

As a fundamental concept in information theory, mutual information (MI) has been commonly applied to quantify association between random vectors. Most existing nonparametric estimators of MI have unstable statistical performance since they involve parameter tuning. We develop a consistent and powerful estimator, called `fastMI`, that does not incur any parameter tuning. Based on a copula formulation, `fastMI` estimates MI by leveraging Fast Fourier transform-based estimation of the underlying density. Extensive simulation studies reveal that `fastMI` outperforms state-of-the-art estimators with improved estimation accuracy and reduced run time for large data sets. `fastMI` provides a powerful test for independence that exhibits satisfactory type I error control. Anticipating that it will be a powerful tool in estimating mutual information in a broad range of data, we develop an R package `fastMI` for broader dissemination.

1. Introduction

Investigating dependence between two random variables is a key issue in statistical science. It is known that classical measures like Pearson's r, Kendall's τ, and Spearman's ρ [9], although widely used in practice, are incapable of capturing a non-linear, non-monotonic association, which cannot be properly estimated using ranks and their monotonic transformations. To address this practical need, more sophisticated measures, like the distance correlation (dCor) [31], the Heller–Heller–Gorfine (HHG) [13] statistic, and the maximal information coefficient (MIC) [24] have been introduced to study more complex association patterns such as the dependence of two random vectors.

Among many non-linear association metrics, mutual information (MI) [5] has recently re-emerged in the statistics and machine learning literature with many exciting applications [34]. Originally introduced in a communication theory context [27], MI presents a remarkably general and intuitive measure of dependence. Its widespread use in practice is due largely to the self-equitability [15] of MI - the ability to characterize dependency strength for both linear and non-linear relationships.

For continuous data, which shall be the focus of this paper, there are three distinct MI estimation approaches. The first approach implements a binning method to group continuous data into different bins and estimates MI from the binned data [22,29]. The success of this simple method depends heavily on proper specification of both number and position of said bins. Another approach is based on a k-nearest neighbors (kNN) estimation method, utilized by the Kraskov–Stögbauer–Grassberger (KSG) estimator [16]. As is the case with all kNN-based methods, the KSG estimator greatly depends on properly specifying the number of neighbors. The third approach is based on estimates of probability density functions (PDFs), using histograms, kernel density estimation (KDE) [20], B-splines [8], or wavelets [23]. This nonparametric approach typically relies on a tuning parameter (e.g. bandwidth) that needs to be specified in the estimation routine.

---

* Corresponding author.
E-mail address: pxsong@umich.edu (P.X.-K. Song).
[1] Equal contribution.

Although the approaches mentioned above demonstrate good properties [8,16,20,22,23], they are all sensitive to proper specification of tuning parameter(s) in a chosen smoothing technique. As a consequence, the resulting estimators may be numerically unstable and/or suffer from serious bias. A recent study [38] presents a tuning-free KDE approach to estimate MI in which the bandwidth parameter is automatically set to maximize the jackknifed version of MI (henceforth referred to as JMI). Not only does this method exhibit better estimation efficiency than other existing MI estimation approaches but it also provides a stable hypothesis test for independence that is shown to be more powerful than its competitors such as the dCor, HHG, or MIC. According to existing literature, JMI serves as the current gold standard for estimating MI as well as is the choice of test for independence [38].

Along the line of tuning-free estimation, an improvement on JMI is proposed in this paper. This new estimator is motivated by the self-consistent density estimator proposed by [2,21] to minimize the mean integrated squared error (MISE) between the estimated density and the true density without incurring any manual parameter tuning. The estimation process relies on fast Fourier transforms (FFT).

Furthermore, utilizing this 'optimal' density estimator, we propose a plug-in estimator of MI, termed as the `fastMI`, which is shown to be consistent and manyfold faster than the original JMI for large data sets. This proposed `fastMI` automatically determines the bandwidth by minimizing the MISE objective function in a data-adaptive way.

Simulation studies comparing estimation accuracy reveal `fastMI` outperforms the current gold standard JMI estimator. Through extensive numerical experiments, `fastMI` demonstrates improved estimation efficiency, higher empirical power when testing for independence, and reduced computation time. All these lead to a recommendation of our new methodology to practitioners.

The rest of the paper is organized as follows. In Section 2 we define MI and underline a key connection of MI with copula. In Section 3 we describe the self-consistent density estimation method in detail. In Section 4 we provide theoretical guarantees on the asymptotic behavior of the `fastMI` estimator. We present extensive simulation studies in Section 5 to highlight the strengths and advantages of using `fastMI` over the existing state-of-the-art estimator, the JMI estimator. Further, we benchmark our findings by means of an empirical copula estimator-based MI as well. Finally, we present an application of `fastMI` to real data in Section 6 before summarizing our findings in the concluding Section 7. All heavy technical details are included in Section 8.

2. Methods

To formalize the problem, suppose that $d \in \mathbb{N}$ can be written as $d = p + q$ for some $p, q \in \mathbb{N}$, that X and Y are random vectors taking values in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively, and that $Z = (X^\top, Y^\top)^\top$ has density f_{XY} with respect to Lebesgue measure on $\mathbb{R}^d$. We write f_X and f_Y for the marginal Lebesgue densities of the random vectors X and Y, respectively. Given independent and identically distributed copies $\{Z_1, \ldots, Z_n\}$ of Z, one of our primary tasks is to test the null hypothesis H_0 that the vectors X and Y are independent against the alternative that X and Y are not independent.

2.1. Mutual information and its copula-based formulation

In order to test the independence of X and Y we consider mutual information $MI(X, Y)$, defined by

$$MI(X, Y) = \mathbb{E}_{XY}\left[\ln\left\{\frac{f_{XY}(X, Y)}{f_X(X)f_Y(Y)}\right\}\right],$$

where $\mathbb{E}_{XY}$ denotes expectation under the joint density f_{XY}. It is known that MI is a valid measure of association; that is, it is equal to zero if and only if X and Y are independent and positive otherwise. Larger values of MI indicate a stronger association. MI is invariant under monotonic transformations, which is an important property allowing various rank-based transformations. Further, MI satisfies the self-equitability condition [38], implying that it detects associations without a bias for specific association patterns, unlike the MIC [15].

Interestingly, MI may be rewritten as a function of copulas, a class of dependence models [14]. Note that by using Sklar's theorem [7], we can apply monotonic marginal transformations to reduce the technical complexity by using uniform transformations $U_X = \left(F_{X_1}(X_1), \ldots, F_{X_p}(X_p)\right)^\top$ and $U_Y = \left(F_{Y_1}(Y_1), \ldots, F_{Y_q}(Y_q)\right)^\top$, where F_{X_i} and F_{Y_j} are the cumulative distribution functions (CDFs) of X_i and Y_j, respectively, for $1 \leq i \leq p$ and $1 \leq j \leq q$. We use $c_{U_X U_Y}$, c_{U_X}, and c_{U_Y} to denote the copula density functions of $(U_X^\top, U_Y^\top)^\top$, U_X, and U_Y respectively. Using this marginal uniform transformation trick, we arrive at an alternative copula-based formulation of MI:

$$MI(X, Y) = MI(U_X, U_Y) = \int_{[0,1]^p} \int_{[0,1]^q} \ln\left\{\frac{c_{U_X U_Y}(u_X, u_Y)}{c_{U_X}(u_X)c_{U_Y}(u_Y)}\right\} c_{U_X U_Y}(u_X, u_Y) du_X du_Y. \tag{1}$$

MI in (1) may be estimated by plug-in copula density estimators. The marginal uniform transformation trick allows us to consider ranks instead of raw data, making copula-based estimation methods robust to any marginal irregularity, in contrast with methods which use raw data to estimate MI, such as the kNN-based KSG estimator. Recognizing MI as an integral in (1), we may invoke data generative methods, including classical Monte Carlo methods [25] for estimation purposes. The estimated copula density functions $\hat{c}_{U_X U_Y}$, $\hat{c}_{U_X}$, and $\hat{c}_{U_Y}$ are used to obtain an estimate of the underlying MI. Using a fast Fourier transform-based density estimation technique [2,21], which is described in greater detail in Sections 3 and 2.2, we obtain said copula density estimates. This estimation routine will be utilized in a powerful MI-based hypothesis test for independence of X and Y.

2.2. Improved nonparametric estimation of copula density using probit transformation

From (1) we note the need to estimate three multivariate copula density functions, namely $c_{U_X U_Y}$, c_{U_X}, and c_{U_Y}. Here we describe the nonparametric KDE of a generic d-variate copula density function for vector $\boldsymbol{U} = (U_1, \ldots, U_d)^\top$ on unit hypercube $[0,1]^d$. The naive KDE [28], is not suitable for unit hypercube-supported copula densities, mainly because it is known to be heavily affected by boundary bias issues. Most kernel estimators, for instance, have problems with such bounded support because, for points close to the boundaries, they typically place some positive mass outside of the support. In addition, according to [10], kernel estimators may fail to provide consistent estimation for some copula density values at corners of the unitary cubic because of the unboundedness issue (i.e., infinite density values). For example, in the bivariate Gaussian copula case with moderate correlation, the copula density is unbounded in two corners of the unitary square $[0,1]^2$. Therefore, to address this issue, a variable transformation approach in the kernel estimation of copula densities was proposed [10,19], which we describe below.

We define a vector of normal scores $\boldsymbol{V} = (V_1, \ldots, V_d)^\top$ where $V_i = \Phi^{-1}(U_i), 1 \le i \le d$, with Φ denoting the standard normal CDF and Φ^{-1} denoting its quantile (probit) function. Given that each U_i is uniformly distributed on $[0,1]$, we have each V_i distributed as a standard normal variable, although this does not force the joint distribution of $\boldsymbol{V}$ to be multivariate normal. Then, the joint density function of $\boldsymbol{V}$, given by g, may be expressed as follows:

$$c(u_1, \ldots, u_d) = \frac{g\left\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right\}}{\phi\left\{\Phi^{-1}(u_1)\right\} \ldots \phi\left\{\Phi^{-1}(u_d)\right\}}, \quad \boldsymbol{u} \in [0,1]^d.$$

The motivation for this probit transformation is as follows: if $c(\boldsymbol{u}) > 0$ Lebesgue-a.e. over $[0,1]^d$, $\boldsymbol{V}$ has unconstrained support over $\mathbb{R}^d$ and estimating its density g no longer suffers from boundary issues. In addition, due to its normal margins, one may expect g to be smooth and well-behaved, and its estimation becomes relatively easy and accurate. It is clear that any estimator $\hat{g}$ of g on $\mathbb{R}^d$ will produce a corresponding estimator of the copula density on $(0,1)^d$:

$$\hat{c}(u_1, \ldots, u_d) = \frac{\hat{g}\left\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right\}}{\phi\left\{\Phi^{-1}(u_1)\right\} \ldots \phi\left\{\Phi^{-1}(u_d)\right\}}, \quad \boldsymbol{u} \in (0,1)^d. \tag{2}$$

Further, when necessary, $\hat{c}$ can also be defined at the boundaries of $[0,1]^d$ by continuity [10]. This probit transformation trick confers many advantages to the estimator $\hat{c}$, including (i) $\hat{g}$, being an unconstrained density estimator on $\mathbb{R}^d$ does not suffer from boundary bias issues and does not allocate any probability to $\hat{c}$ outside $[0,1]^d$; (ii) if $\hat{g}$ is a *bona fide* density estimator, i.e., $\hat{g}(\boldsymbol{v}) \ge 0 \ \forall \boldsymbol{v} \in \mathbb{R}^d$ and $\int_{\mathbb{R}^d} \hat{g}(\boldsymbol{v}) d\boldsymbol{v} = 1$, then by change of variable $U_i = \Phi(V_i)$ for $1 \le i \le d$ we have $\hat{c}(\boldsymbol{u}) \ge 0 \ \forall \boldsymbol{u} \in [0,1]^d$ and $\int_{[0,1]^d} \hat{c}(\boldsymbol{u}) d\boldsymbol{u} = 1$; (iii) if $\hat{g}$ is a uniformly (weak or strong) consistent estimator for g, we note that the corresponding $\hat{c}$ inherits the same behavior on any compact subset of $[0,1]^d$.

The key challenge now lies in obtaining a 'good' PDF estimator $\hat{g}$. Kernel density estimators (KDEs) are commonly used for estimating PDFs. A well-known technical challenge in the KDE method is to determine a kind of optimal bandwidth H in addition to a specifically chosen kernel density [38]. Albeit a vast literature exists on this issue of bandwidth tuning, selecting optimal H remains case-dependent and computationally burdensome, and oftentimes this task involves a manual user intervention [28]. A review of automatic selection methods [12] recommends a variety of different approaches that are dependent on data set characteristics (including sample size, smoothness, and skewness) and thus are hard to implement properly in practice. In Section 3, we consider an alternative approach to density estimation that relies on fast Fourier transforms.

In summary, using the probit-transformation trick and an FFT-based density estimation method, we are able to obtain the estimated copula density functions $\hat{c}_{U_X U_Y}$, $\hat{c}_{U_X}$, and $\hat{c}_{U_Y}$ via (2). These estimated densities are used to compute our estimator `fastMI` for MI as described in Section 2.3.

2.3. fastMI: fast nonparametric estimation of MI

Let $\boldsymbol{Z}_i = (\boldsymbol{X}_i^\top, \boldsymbol{Y}_i^\top)^\top$, with $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{pi})^\top$ and $\boldsymbol{Y}_i = (Y_{1i}, \ldots, Y_{qi})^\top$ for $i \le i \le n$ be a random sample drawn from a d-variate distribution f_{XY}. As a preliminary processing step, we define the vectors of empirical probability integral transforms $\boldsymbol{U}_{Xi} = (\hat{F}_{X_1}(X_{1i}), \ldots, \hat{F}_{X_p}(X_{pi}))^\top$ and $\boldsymbol{U}_{Yi} = (\hat{F}_{Y_1}(Y_{1i}), \ldots, \hat{F}_{Y_q}(Y_{qi}))^\top$. Here, $\hat{F}_{X_i}$ and $\hat{F}_{Y_j}$ are the empirical CDFs of X_i and Y_j respectively for $1 \le i \le p$ and $1 \le j \le q$. Next, we invoke the probit transformation as described in Section 2.2 to obtain the estimated copula density functions: $\hat{c}_{U_X U_Y}$, $\hat{c}_{U_X}$, and $\hat{c}_{U_Y}$, as described by (2). These estimated densities are used to compute our plug-in estimator `fastMI` estimator using (1). Consequently, `fastMI` is given by

$$\widehat{MI}_{fast} = n^{-1} \sum_{i=1}^{n} \ln\left\{ \frac{\hat{c}_{U_X U_Y}(\boldsymbol{U}_{Xi}, \boldsymbol{U}_{Yi})}{\hat{c}_{U_X}(\boldsymbol{U}_{Xi}) \hat{c}_{U_Y}(\boldsymbol{U}_{Yi})} \right\}. \tag{3}$$

Note that we use `fastMI` and $\widehat{MI}_{fast}$ interchangeably in this article. In Section 3 we discuss a data-driven fast Fourier transform-based density estimation technique that yields consistent and fast estimates of the underlying density function without being encumbered by bandwidth selection issues.

3. Self-consistent density estimation via fast Fourier transforms

A density estimation technique that has no need for user-selected parameter tuning was introduced by [2] for univariate continuous random variables and later extended to higher dimensions by [21]. The density estimator called the self-consistent (SC) estimator, was shown to converge almost surely to the true underlying density for the univariate case. In this section, we present the derivation of the SC estimator for the multivariate case and present proof of almost sure consistency of the SC estimator to the true d-dimensional density under mild assumptions. To our knowledge, while proof of consistency for the SC estimator exists for the univariate case, this is the first attempt at proving the consistency of the estimator for the multivariate case.

The SC estimator has desirable large-sample properties in addition to enjoying greatly improved computations speeds. Using the SC estimator and (2), we obtain the estimated copula density functions $\hat{c}_{U_X U_Y}$, $\hat{c}_{U_X}$, and $\hat{c}_{U_Y}$, which are then used to compute `fastMI` using (3).

3.1. Formulation

Let $S = \{Z_j \in \mathcal{Z} \subseteq \mathbb{R}^d, j = 1, \ldots, n\}$ denote a random sample drawn from PDF f_Z. Assume that the true PDF f_Z belongs to the Hilbert space of square-integrable functions, namely $\mathcal{L}^2 = \{g : \int g^2(z)dz < \infty\}$. We propose to estimate $\hat{f}_Z$ by the following convolution of an arbitrary kernel K and the set of delta functions centered on the dataset for $z \in \mathbb{R}^d$:

$$\hat{f}_Z(z) := n^{-1} \sum_{j=1}^{n} K(z - Z_j) = n^{-1} \sum_{j=1}^{n} \int_{\mathbb{R}^d} K(t)\delta(z - Z_j - t)dt = n^{-1} \sum_{j=1}^{n} (K * \delta)(z - Z_j), \qquad (4)$$

where $\delta(z)$ is the Dirac delta function [17] and $f * g$ denotes the convolution of two functions f and g: $(f * g)(t) := \int_{\mathbb{R}^d} f(\tau)g(t-\tau)d\tau$. Here K is chosen such that the resulting estimator $\hat{f}_Z \in \mathcal{L}^2$.

Denote the space of candidate kernel functions by

$$\mathcal{K} := \left\{ K : K(z) \geq 0; K(z) = K(-z); \int K(z)dz = 1 \right\}.$$

Given a random sample S, our aim is to identify an optimal kernel $\hat{K} \in \mathcal{K}$, that achieves the minimal mean integrated square error (MISE) between the true density f_Z and an estimator $\hat{f}_Z$:

$$\hat{K} := \operatorname*{argmin}_{K \in \mathcal{K}} \operatorname{MISE}(\hat{f}_Z, f_Z) = \operatorname*{argmin}_{K \in \mathcal{K}} \mathbb{E}\left[\int_{\mathbb{R}^d} \{\hat{f}_Z(z) - f_Z(z)\}^2 dz\right], \qquad (5)$$

where the $\mathbb{E}$ operator denotes expectation taken over the entire support of f_Z.

To perform minimization in (5), we follow a procedure for signal deconvolution [37] via Fourier transforms. First, consider the Fourier transform of the true density f_Z, namely the characteristic function (CF) given by

$$\phi(t) := \int_{\mathbb{R}^d} f_Z(z) \exp(it^\top z)dz, \quad t \in \mathbb{R}^d.$$

Then, using the Fourier convolution theorem we obtain

$$\mathcal{F}\{K(z - Z_j)\}(t) = \mathcal{F}\{K(z) * \delta(z - Z_j)\}(t) = \kappa(t) \exp(it^\top Z_j), \quad t \in \mathbb{R}^d, \qquad (6)$$

where $\mathcal{F}$ is the Fourier transform operator and $\kappa(t) := \mathcal{F}\{K(z)\}(t)$ denotes the Fourier transform of the kernel K. From (6) and the linearity of Fourier transforms in (4), we obtain $\hat{\phi}(t)$, the Fourier transform of the estimate $\hat{f}_Z$, as follows for $t \in \mathbb{R}^d$:

$$\hat{\phi}(t) = \mathcal{F}\{\hat{f}_z(z)\}(t) = \kappa(t) \, n^{-1} \sum_{j=1}^{n} \exp(it^\top Z_j) = \kappa(t)C(t), \qquad (7)$$

where $C(t) = n^{-1} \sum_{j=1}^{n} \exp(it^\top Z_j)$ is the empirical characteristic function (ECF). Since the data $\{Z_1, \ldots, Z_n\}$ are i.i.d., it is easy to see that the ECF is an unbiased estimator of the corresponding CF, i.e., $\mathbb{E}\{C(t)\} = \phi(t)$. It can also be shown that $\mathbb{E}|C(t) - \phi(t)|^2 = (1 - |\phi(t)|^2)/n$. We refer the reader to [33, Chapter 3] for more details on the ECF and its properties. It follows that $\mathbb{E}\{\hat{\phi}(t)\} = \kappa(t)\phi(t)$ for $t \in \mathbb{R}^d$.

The MISE in (5) corresponds to the mean-square distance between the true density f_Z and the estimate $\hat{f}_Z$, in terms of the Euclidean metric in the Hilbert space $\mathcal{L}^2$. We rewrite (5) in Fourier space using Parseval's theorem as follows:

$$\operatorname{MISE}(\hat{f}_Z, f_Z) = (2\pi)^{-d} \, \mathbb{E}\left[\int_{\mathbb{R}^d} |\hat{\phi}(t) - \phi(t)|^2 dt\right]. \qquad (8)$$

Since $f_Z(z), \hat{f}_Z(z), \phi(t), \hat{\phi}(t) \in \mathcal{L}^2$, we may interchange the expectation and integral operations [3] and rewrite the MISE in (8) as follows:

$$\operatorname{MISE}(\hat{f}_Z, f_Z) = (2\pi)^{-d} \int_{\mathbb{R}^d} \left[\mathbb{E}|\hat{\phi}(t) - \phi(t)|^2 dt\right] = (2\pi)^{-d} \int_{\mathbb{R}^d} \left[\mathbb{E}|\hat{\phi}(t) - \mathbb{E}\{\hat{\phi}(t)\} + \mathbb{E}\{\hat{\phi}(t)\} - \phi(t)|^2 dt\right]$$

$$= (2\pi)^{-d} \int_{\mathbb{R}^d} \left[n^{-1}|\kappa(t)|^2 \left\{1 - |\phi(t)|^2\right\} + |\phi(t)|^2 |1 - \kappa(t)|^2\right] dt. \qquad (9)$$

Since the integrand in the last line of (9) is quadratic in κ it is straightforward to find the optimal Fourier-transformed kernel κ_{OPT} that minimizes MISE, by equating the functional derivative of MISE with respect to κ to zero and solving for κ_{OPT}, given by

$$\kappa_{OPT}(t) = \frac{n}{n - 1 + |\phi(t)|^{-2}}, \quad t \in \mathbb{R}^d. \tag{10}$$

(10) reveals that a unique 'optimal' Fourier transformed kernel can be derived as a function of the power spectrum of the (unknown) density that is to be estimated. Although this finding was first reported in [36], the result alone is unfortunately of little use since the power spectrum ϕ of the true density f_Z is not known. We follow the suggestions in [2] and take a step further to plug in to (7), the transformed kernel κ_{OPT} obtained in (10) to write the Fourier transform $\hat{\phi}$ of the density estimate $\hat{f}_z$ as follows

$$\hat{\phi}(t) = C(t)\kappa_{OPT}(t) = \frac{nC(t)}{n - 1 + |\phi(t)|^{-2}}, \quad t \in \mathbb{R}^d. \tag{11}$$

Then, an iterative procedure proposed by [2,21] may be applied to determine the exact fixed point of (11), with the details described in Section 3.2.

3.2. Multivariate self-consistent estimation

An iterative procedure is defined by the following sequence of estimates, begun with an initial guess $\tilde{\phi}_0(t)$:

$$\tilde{\phi}_{n+1}(t) = \frac{nC(t)}{n - 1 + |\tilde{\phi}_n(t)|^{-2}}, \quad t \in \mathbb{R}^d. \tag{12}$$

The convergent point from (12) yields an estimate $\tilde{\phi}_{SC}(t)$ which satisfies

$$\tilde{\phi}_{SC}(t) = \frac{nC(t)}{n - 1 + |\tilde{\phi}_{SC}(t)|^{-2}}, \quad t \in \mathbb{R}^d. \tag{13}$$

The existence of $\tilde{\phi}_{SC}(t)$ is warranted as (13) has two fixed and unique non-null solutions. We show that (13) has only one stable non-null solution. Noting that $\tilde{\phi}_{SC}$ is complex-valued, we take the absolute value of (13). When the null solution $\tilde{\phi}_{SC} = 0$ is removed, we obtain a simple quadratic equation:

$$(n - 1) |\tilde{\phi}_{SC}(t)|^2 = n|C(t)||\tilde{\phi}_{SC}(t)|.$$

Provided $|C(t)| \geq 4(n - 1)/n^2$, the above equation has two solutions denoted by the superscripted $|\tilde{\phi}_{SC}(t)^{\pm}|$,

$$|\tilde{\phi}_{SC}(t)^{\pm}| = \frac{n|C(t)|}{2(n - 1)} \left[1 \pm \sqrt{\left\{ 1 - \frac{4(n - 1)}{n^2|C(t)|^2} \right\}} \right].$$

When substituted in (13), they return the solution for $\tilde{\phi}_{SC}(t)^{\pm}$ of the form:

$$\tilde{\phi}_{SC}(t)^{\pm} = \frac{nC(t)}{2(n - 1)} \left[1 \pm \sqrt{\left\{ 1 - \frac{4(n - 1)}{n^2|C(t)|^2} \right\}} \right], \quad t \in \mathbb{R}^d \cap \left\{ t : |C(t)| \geq 4(n - 1)/n^2 \right\}.$$

Since these two solutions only depend on the ECF C, they may be used to estimate ϕ. Below, we identify a stable solution to be used in (11).

Whereas $\tilde{\phi}_{SC}^+$ is normalized, $\tilde{\phi}_{SC}^-$ is not; that is, $C(0) = 1$ implies $\tilde{\phi}_{SC}^+(0) = 1$, which is desirable; in contrast $\tilde{\phi}_{SC}^-(0) = 1/(n - 1)$, which is undesirable. Furthermore, we compute the derivative

$$\frac{d|\tilde{\phi}_{n+1}|}{d|\tilde{\phi}_n|}\Bigg|_{|\tilde{\phi}_n| = |\tilde{\phi}_{SC}^{\pm}|} = 1 \mp \sqrt{\left\{ 1 - \frac{4(n - 1)}{n^2|C(t)|^2} \right\}}.$$

Thus, under $|C(t)| \geq 4(n - 1)/n^2$, $\tilde{\phi}_{SC}^+$ has a derivative smaller than one, indicating a stable equilibrium point, whereas $\tilde{\phi}_{SC}^-$ has a derivative larger than one, indicating instability. In summary, we choose $\hat{\phi}_{SC} = \tilde{\phi}_{SC}^+$ as the solution to (11) in the construction of the estimate. That is,

$$\hat{\phi}_{SC}(t) = \frac{nC(t)}{2(n - 1)} \left[1 - \sqrt{\left\{ 1 - \frac{4(n - 1)}{n^2|C(t)|^2} \right\}} \right] \mathbb{I}_{A_n}(t), \quad t \in \mathbb{R}^d. \tag{14}$$

where A_n serves as a low-pass filter that ensures the stability of the estimation process. Since $\hat{\phi}_{SC}$ and A_n are bounded (see remarks below), (14) can be antitransformed back to the estimate in real space, given as follows:

$$\hat{f}_{SC}(x) = (2\pi)^{-d} \int_{\mathbb{R}^d} \hat{\phi}_{SC}(t) \exp(-it^{\top}x)dt. \tag{15}$$

Remark 1. The purpose of the filter $\mathbb{I}_{A_n}(t)$ is to define a Fourier-based low-pass filter on the ECF $C(t)$ that yields a stable optimal estimate in the minimum MISE sense. Primarily, the set A_n is specified such that:

$$A_n = \left\{ t \in \mathbb{R}^d : |C(t)|^2 \geq C_{\min}^2 = \frac{4(n - 1)}{n^2} \right\}. \tag{16}$$

Here, the primary filter is necessary for the stability of the iteration method, since the lower bound $C_{\min}$ can ensure a well-defined square root term in (14). Moreover, according to [2], the set A_n may exclude an additional small subset of frequencies to produce a smoother density estimate $\hat{f}_Z$. In order for $\hat{f}_Z$ to converge to the true density f_Z as n increases, we require that this set of additionally excluded frequencies must shrink, so that the set A_n of included frequencies grows with increasing n.

Remark 2. According to [21], the multivariate ECF $C(t)$ consists of a finite set of contiguous hypervolumes denoted by $\left\{ HV_1^n, \ldots, HV_{k_n}^n \right\}$ where k_n is a finite integer. Each hypervolume permits 'above-threshold' frequency values t for which the constraint in (16) holds. Note that at least one such contiguous hypervolume containing $t = 0$ is guaranteed to exist since $C(0) = 1$ due to normalization and the primary filter A_n has a lower bound $C_{\min} \leq 1$. Following the suggestion by [21] we employ the lowest contiguous hypervolume filter, choosing the only hypervolume centered at $t = 0$, which we denote as HV_1^n for notational convenience. We make the following observations about HV_1^n:

1. The set of frequencies included in the lowest contiguous hypervolume filter is bounded above since the frequencies will always be contained within a finite-sized hypervolume around the origin.
2. The volume of the lowest contiguous hypervolume filter grows as the number of data points n increases, implying more frequencies are included for larger sample sizes.

The resulting filter satisfies the convergence conditions described by [2]. Hence, we set $A_n = HV_1^n$, and study convergence of $\hat{f}_{SC}$ to the true f_z as n increases. For notational convenience, let $\bar{A}_n$ denote the complement set of A_n and $\mathcal{V}(A_n)$ denote the volume of A_n.

When compared to the classical kernel estimation approach [28] that assumes a specific form of the kernel with a need of tuning bandwidth, the advantage of the above SCE lies in making minimal assumptions on the functional form of K, while determining the estimate $\hat{\phi}_{SC}$ as a function of the ECF C, along with a well-specified low-pass frequency filter A_n entirely by a data-driven approach. The proposed optimization via (11) enjoys computational efficiency when 'finding' the functional form of the optimal kernel.

4. Theoretical guarantees

4.1. Large sample properties

We now present key large-sample properties for the SCE estimator and the subsequent plug-in estimator of MI, beginning with Theorem 1 that establishes strong consistency of the SCE estimator $\hat{f}_{SC}$ at all points on the support of f_Z.

Theorem 1. *Let the true density f_Z be square integrable and its corresponding Fourier transform ϕ be integrable, then the self consistent estimator $\hat{f}_{SC}$, which is defined by (15), converges almost surely to f_Z as $n \to \infty$, under the additional assumptions $\mathcal{V}(A_n) \to \infty$, $\mathcal{V}(A_n)/\sqrt{n} \to 0$ and $\mathcal{V}(\bar{A}_n) \to 0$ as $n \to \infty$. Further, assuming f_Z to be continuous on support $\mathbb{R}^d$, we have uniform almost sure convergence of $\hat{f}_{SC}$ to f_Z as $n \to \infty$.*

The proof of Theorem 1 is given in Section 8. Note that $\hat{f}_{SC}$ can be used in conjunction with the probit transformation trick described in Section 2.2 in order to obtain uniformly strong consistent estimators for the copula density functions $\hat{c}_{U_X U_Y}$, $\hat{c}_{U_X}$, and $\hat{c}_{U_Y}$ and subsequently obtain the fastMI estimator described by (3). fastMI is shown to be consistent, as established by Theorem 2 below.

Theorem 2. *Let the assumptions of Theorem 1 hold. Further, we assume the true underlying copula functions $c_{U_X U_Y}$, c_{U_X}, and c_{U_Y} are bounded away from zero and infinity on their support. Then, the estimator $fastMI$ given by (3) converges in probability to the true MI given in (1) as $n \to \infty$.*

The proof of Theorem 2 is given in Section 8.

4.2. Test for independence

On the basis of Theorem 2, using the estimator $\widehat{MI}_{fast}$, we propose a permutation-based test for independence, i.e., a test for $H_0 : MI = 0$ against the alternative $H_a : MI > 0$.

Typically, rejection rules of a test based on large-sample theory require data with large sample sizes, which may not be always available in practice. Leveraging the fast computational speed of our SC estimation method, in this paper we consider a hypothesis testing method that is deemed stable and reliable. Thus, we implement a permutation-based test as it is known to give a precise finite-sample distribution of the test statistic for even small samples [18]. For a random sample of n observations, $S = \left\{ (X_1, Y_1), \ldots, (X_n, Y_n) \right\}$, let $\{\delta(1), \ldots, \delta(n)\}$ be a random permutation of $\{1, \ldots, n\}$. Based on δ-permuted data set $S_\delta = \left\{ (X_1, Y_{\delta(1)}), \ldots, (X_n, Y_{\delta(n)}) \right\}$, we calculate the corresponding estimate $\widehat{MI}_{fast}^\delta$. On repeating the above procedure r times, $\mathcal{T}_{perm} = \left\{ \widehat{MI}_{fast}^{\delta_1}, \ldots, \widehat{MI}_{fast}^{\delta_r} \right\}$, a collection of estimates is obtained, which may be used to approximate the null distribution of $\widehat{MI}_{fast}$, under the null hypothesis H_0 : X and Y are independent. At significance level α, we reject the null hypothesis when $\widehat{MI}_{fast}$ based on the original data is greater than the $(1 - \alpha)$th empirical quantile of $\mathcal{T}_{perm}$.

Table 1

Percentage decrease in mean squared error values of `fastMI` relative to $ECMI$ (`fastMI` relative to JMI) for sample sizes $n = 256$ for all copula models across various levels of association, controlled by Kendall's τ.

τ	Copula		
	Clayton	*Gaussian*	*Gumbel*
0.0	97 (98)	28 (81)	83 (73)
0.1	83 (72)	71 (71)	46 (72)
0.2	17 (68)	12 (75)	14 (80)
0.3	17 (90)	97 (99)	17 (61)
0.4	78 (70)	72 (64)	64 (76)
0.5	40 (68)	25 (68)	22 (73)
0.6	27 (79)	36 (91)	97 (99)
0.7	5 (54)	79 (71)	67 (68)
0.8	58 (71)	45 (79)	32 (73)
0.9	46 (81)	42 (83)	6 (73)

5. Simulation experiments

Previous replicable studies [38] have compared the performance of the JMI-based test with other popular methods, including dCor, HHG and MIC-based tests, and concluded that JMI appears to be the most stable test for independence. Further, as a benchmark we compare the performance of our `fastMI` estimator with the empirical copula-based estimator of MI (named $ECMI$), in which a naive bandwidth-driven kernel density estimation route is taken when estimating the copula functions described in (1).

In this section, our proposed nonparametric estimator is compared with the $ECMI$ and JMI from three distinct but inter-related perspectives: (i) estimation accuracy, (ii) ability to test for independence, and (iii) computation time. We design and implement exhaustive simulation-based experiments for our study. In brief: relative to the $ECMI$ as well as the JMI, the `fastMI` is the more accurate estimator with reduced estimation error and is able to conduct a hypothesis test for independence with a higher power while exhibiting satisfactory type I error control. Reduced computation times further indicate it is more amenable to applications in large data sets when compared to $ECMI$ and JMI.

5.1. Estimation accuracy of MI

Through extensive simulation studies, we compare the MSE performance of `fastMI` with $ECMI$ and JMI for a wide range of association patterns for different sample sizes. We generate a sample of n observations drawn from PDF f_{XY} on $\mathbb{R}^d$, specified by the underlying copula and marginal densities. In the data simulation, choices of marginal distributions have little effect, and for numerical convenience, normal marginals are chosen for data generation in our empirical studies.

We first consider the bivariate case ($d = 2$) and restrict ourselves to three separate classes of copula models [7,14] – the symmetric Gaussian copula and two asymmetric Archimedean copula — the Clayton and Gumbel copulas. While the Clayton copula exhibits greater dependence in the negative tail, the Gumbel copula exhibits greater dependence in the positive tail. Each of these copula classes may be specified by fixing the underlying value of Kendall's τ, which in turn, may be used to compute the underlying true MI [4]. For each of the three copula classes considered, we fix $\tau \in \{0, 0.1, 0.2, \dots, 0.9\}$ and generate $n \in \{64, 128, 256\}$ bivariate samples. The simulated data are used to compute $ECMI$, JMI, and `fastMI`. The MSEs of all estimators for different models and sample sizes are calculated based on $s = 1000$ replications. Our findings are presented in Fig. 1. To study the behavior of our estimators more closely, in Table 1 we present information on the following percentage decreases:

$$\% \text{ decrease in MSE of fastMI relative to ECMI } = 100 \times \frac{\text{MSE}(ECMI) - \text{MSE}(\texttt{fastMI})}{\text{MSE}(ECMI)}$$

$$\% \text{ decrease in MSE of fastMI relative to JMI } = 100 \times \frac{\text{MSE}(JMI) - \text{MSE}(\texttt{fastMI})}{\text{MSE}(JMI)},$$

with positive values indicating more accurate performance of `fastMI` over $ECMI$ and JMI respectively. Higher values indicate reduced estimation error.

From both the figure as well as the table, we note that `fastMI` has appreciably lower MSE for all models and sample sizes considered, indicating its superior performance over JMI. Further, from Fig. 1 we note that while the overall standard error for `fastMI` is lower than $ECMI$ as well as JMI in almost all the cases considered, the absolute bias performance of `fastMI` and $ECMI$ are comparable while still being lower than that of JMI. Overall, there seems to be a bias–variance trade-off between `fastMI` and $ECMI$, with `fastMI` yielding better MSE performance.

Now, for the multivariate case ($d > 2$) where both X and Y are bivariate random vectors following a joint d–variate normal distribution with zero mean and dispersion matrix Σ of the form

$$\Sigma := \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY}^{\top} & \Sigma_{YY} \end{bmatrix}.$$
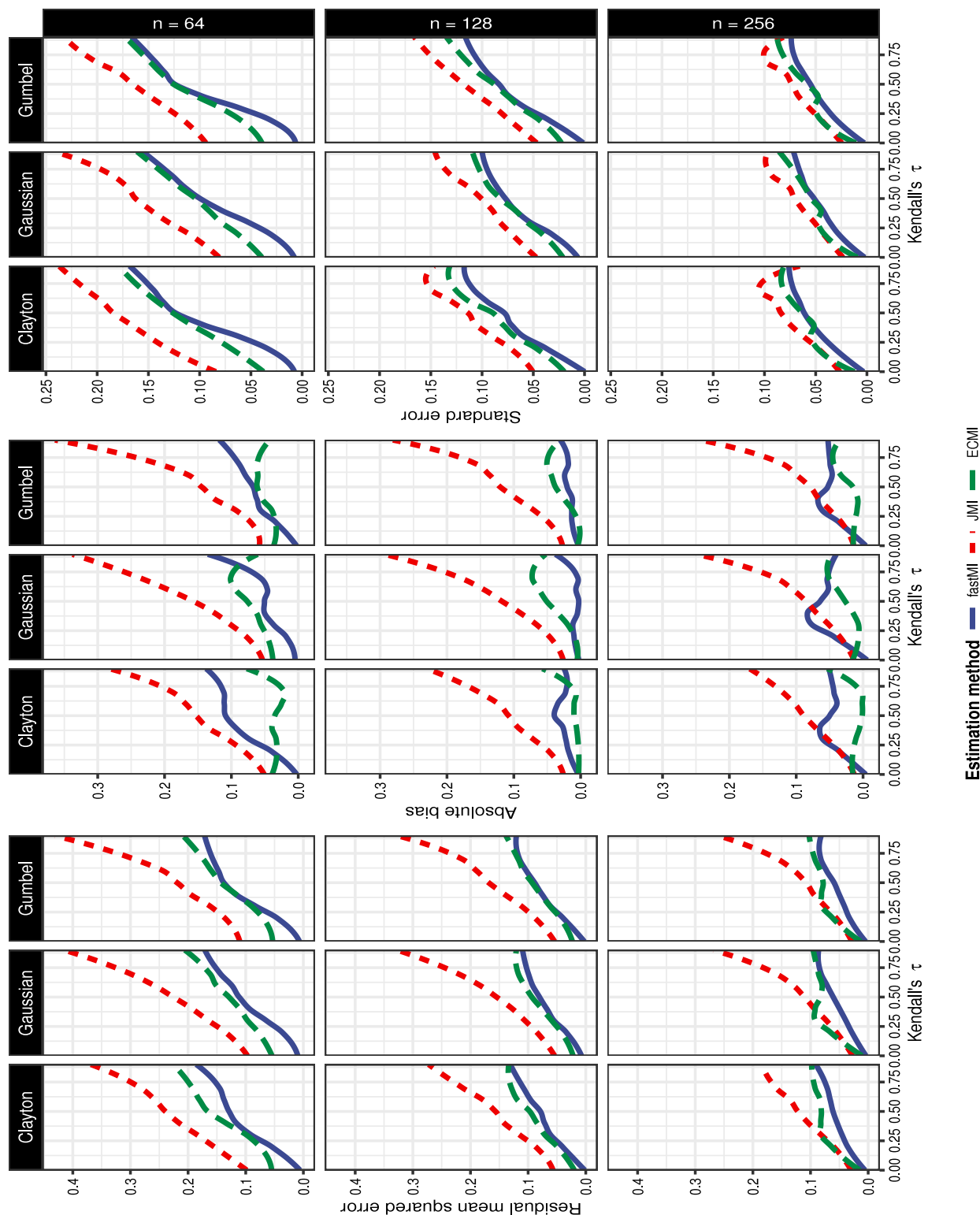
Fig. 1. Comparison of residual mean squared error (left panel), absolute bias (middle panel), and standard error (right panel) of competing MI estimation methods among different configurations of bivariate copula families with varying sample sizes. The competing estimation methods include the empirical copula-based plugin MI (ECMI), the fast Fourier transform based-MI (`fastMI`) and the jackknifed MI (JMI).

Table 2

Percentage decrease in mean squared error values of `fastMI` relative to $ECMI$ (`fastMI` relative to JMI) for sample sizes $n = 256$ for different dispersion structures of a 4-dimensional multivariate normal distribution across various strengths of correlation, controlled by Pearson's ρ.

ρ	Dispersion structure				
	AR-1	*CS*	*Spatial*	*Block ($\rho_w = 1/3$)*	*Block ($\rho_w = 2/3$)*
0.0	83 (94)	80 (95)	75 (96)	63 (97)	47 (97)
0.1	30 (98)	84 (95)	73 (95)	44 (95)	32 (96)
0.2	44 (95)	24 (92)	81 (94)	51 (96)	7 (90)
0.3	7 (88)	19 (87)	26 (89)	36 (88)	65 (95)
0.4	63 (96)	58 (97)	56 (98)	51 (98)	50 (94)
0.5	38 (95)	4 (94)	34 (96)	29 (96)	13 (97)

Table 3

Comparison of empirical type I error of permutation-based test for independence in several bivariate copula families. Using $r = 1000$ permutations, we compare the type I error of $ECMI$, JMI, and `fastMI` for sample sizes $n = 128$ ($n = 256$).

	Copula		
	Clayton	*Gaussian*	*Gumbel*
$ECMI$	0.04 (0.07)	0.04 (0.05)	0.05 (0.04)
JMI	0.06 (0.07)	0.04 (0.06)	0.06 (0.07)
`fastMI`	0.03 (0.04)	0.04 (0.04)	0.03 (0.04)

The reason for choosing the multivariate normal distribution is that it is one of the few multivariate distributions which give rise to a closed-form expression of $MI(X, Y) = 0.5 \ln \left(|\Sigma_{XX}| |\Sigma_{YY}| / |\Sigma| \right)$ where $|\Sigma|$ denotes the determinant of Σ. This enables the comparison of the three competing estimators in a meaningful way. We set all marginal variances to be unity and consider different structures of Σ including the (i) first-order auto-regressive (AR-1) structure, with the generic (i, j)–th element being $\sigma_{ij} = \rho^{|i-j|}$; (ii) compound symmetry (CS) structure, that has off-diagonal elements are set to $\rho \in (0, 1)$; (iii) spatial structure with the generic (i, j)th element being $\sigma_{ij} = \exp(-|i - j|/\rho)$; (iv) two block-correlation matrices that emulate two hierarchies of correlation: a within-block correlation ρ_w and a between-block correlation ρ:

$$\Sigma_{block} := \begin{bmatrix} 1 & \rho_w & \rho & \rho \\ \rho_w & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho_w \\ \rho & \rho & \rho_w & 1 \end{bmatrix}, \quad \rho_w \in \{1/3, 2/3\}.$$

We set $\rho_w = 1/3$ for mild within-block correlation and $\rho_w = 2/3$ for strong within-block correlation.

For each of the five correlation structures described above, we vary $\rho \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and compare the accuracy of the competing estimators by means of MSE. Our findings are presented in Fig. 2. To study the behavior of our estimators more closely, in Table 2 we present information on percentage decrease in MSE of `fastMI` relative to $ECMI$ and JMI for the different models and with sample size set to $n = 256$. As was the case for lower dimensions, from Fig. 2 and Table 2 we note that `fastMI` has appreciably lower MSE than $ECMI$ and JMI for all models and sample sizes considered, indicating its superior performance.

5.2. Test for independence

We compare the permutation-based tests based respectively on our `fastMI` with $ECMI$ and the JMI. For the same bivariate patterns as described in Section 5.1, with $\tau \in \{0, 0.05, \ldots, 0.50\}$ ($\tau = 0$ indicates independence), we plot the empirical power curves for the tests at significance level $\alpha = 0.05$ and present our results under different settings in Fig. 3 for $r = 1000$ permutations. In Table 3 we present a comparison ($ECMI$ and JMI versus `fastMI`) of empirical type I error of the permutation-based test for independence in several bivariate distributions for different sample sizes respectively. Note that the type I error rates of all three methods considered are very close to the nominal level $\alpha = 0.05$.

Similarly, for the same multivariate normal patterns, with $\rho \in \{0, 0.05, \ldots, 0.50\}$ ($\rho = 0$ indicates independence), we plot the empirical power curves for the tests at significance level $\alpha = 0.05$ and present our results under different settings in Fig. 4 for $r = 1000$ permutations. In Table 4 we present a comparison ($ECMI$ and JMI versus `fastMI`) of empirical type I error of the permutation-based test for independence in several multivariate distributions for different sample sizes respectively. Note that the type I error rates of all three methods considered are very close to the nominal level $\alpha = 0.05$.

5.3. Computation time

Since nonparametrically estimating MI is a computationally intensive method, we compare run times of $ECMI$, JMI and `fastMI` for various sample sizes for bivariate data. We report the mean and standard deviation of run time (in seconds) in Table 5.
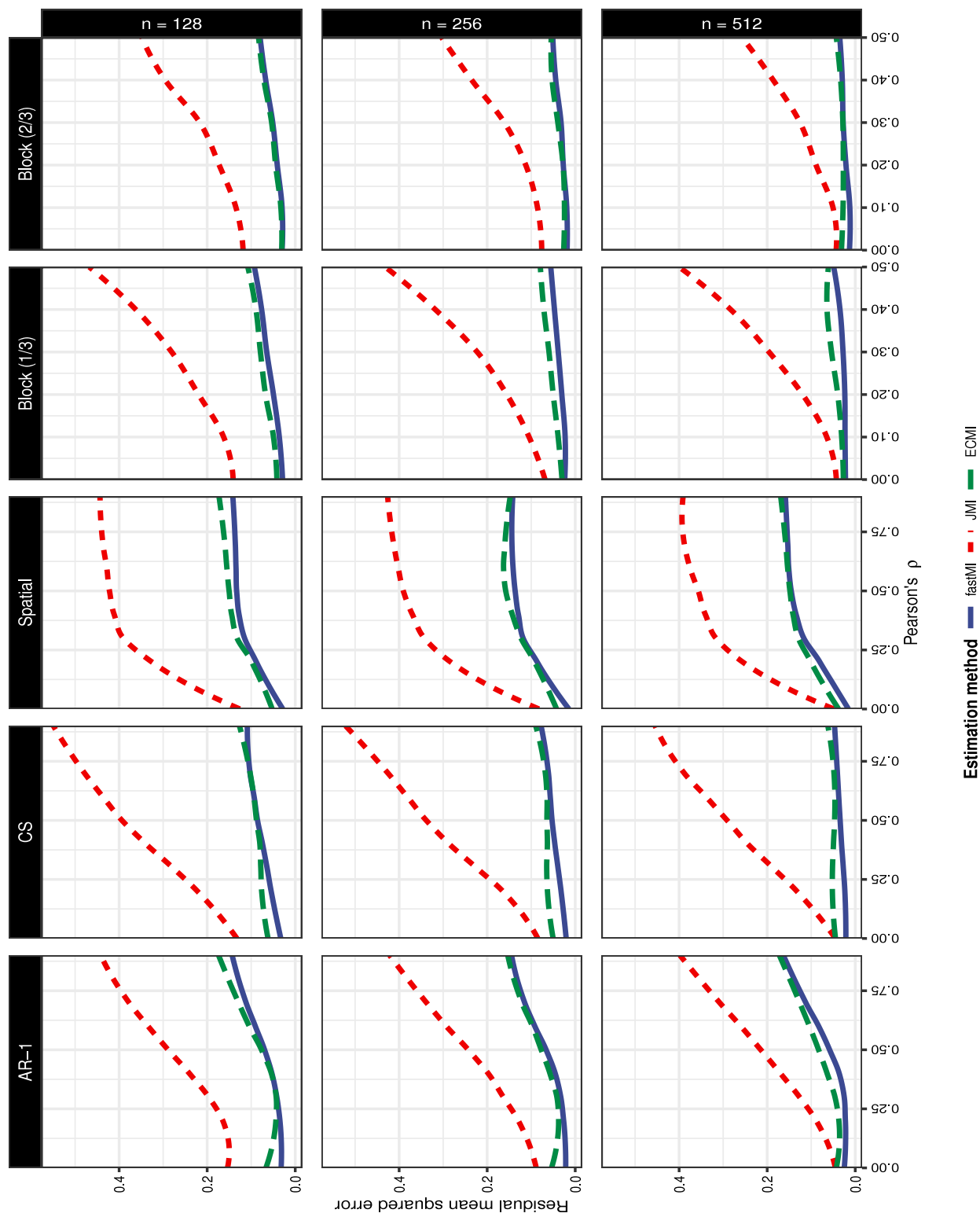
Fig. 2. Comparison of residual mean squared error, absolute bias, and standard error of competing MI estimation methods for different configurations of multivariate copula families with varying sample sizes. The estimation methods in the comparison include the empirical copula-based plugin MI (ECMI), the fast Fourier transform based-MI (fastMI) and the jackknifed MI (JMI).
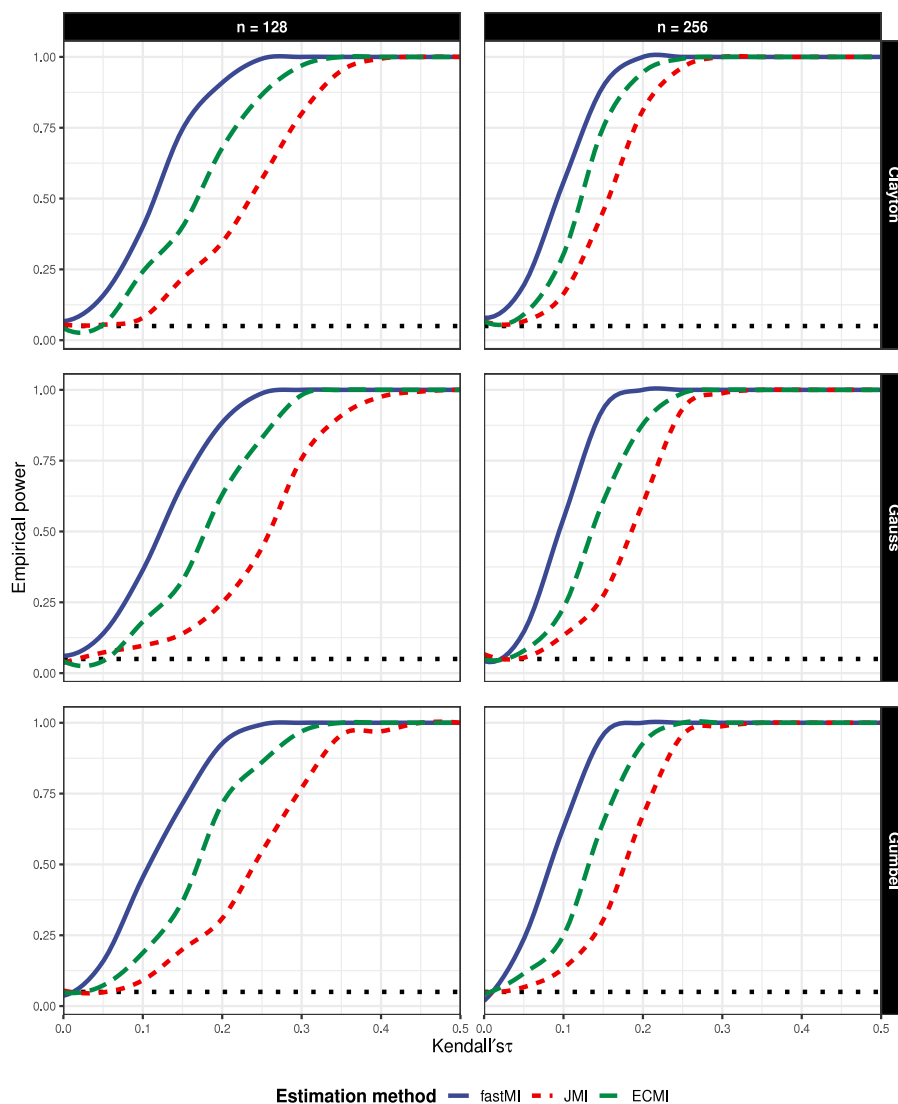
Fig. 3. Comparison on empirical power of competing tests for independence via mutual information (MI) among various configurations of bivariate copula families with varying sample sizes. The dotted black line parallel to the *x*-axis denotes the specified level of significance $\alpha = 0.05$. The estimation methods considered in the comparison include the empirical copula-based plugin MI (ECMI), the fast Fourier transform based-MI (fastMI) and the jackknifed MI (JMI).

Table 4
Comparison of empirical type I error of permutation-based test for independence in several multivariate normal families. Using $r = 1000$ permutations, we compare the type I error of $ECMI$, JMI, and fastMI for sample sizes $n = 128$ ($n = 256$).

	Dispersion structure				
	AR-1	*CS*	*Spatial*	*Block* $(\rho_w = 1/3)$	*Block* $(\rho_w = 2/3)$
$ECMI$	0.05 (0.05)	0.08 (0.02)	0.06 (0.10)	0.07 (0.06)	0.03 (0.07)
JMI	0.07 (0.04)	0.04 (0.07)	0.07 (0.02)	0.08 (0.08)	0.08 (0.06)
fastMI	0.08 (0.07)	0.04 (0.04)	0.04 (0.03)	0.07 (0.07)	0.04 (0.03)

Of note, for smaller sample sizes, JMI is fastest; in contrast, as sample size increases, fastMI becomes fast with run times being many fold smaller than both $ECMI$ and JMI. This improvement in run time establishes fastMI as an attractive method for studying association in practice, because it is the case of large sample size that matters in real-world computations.
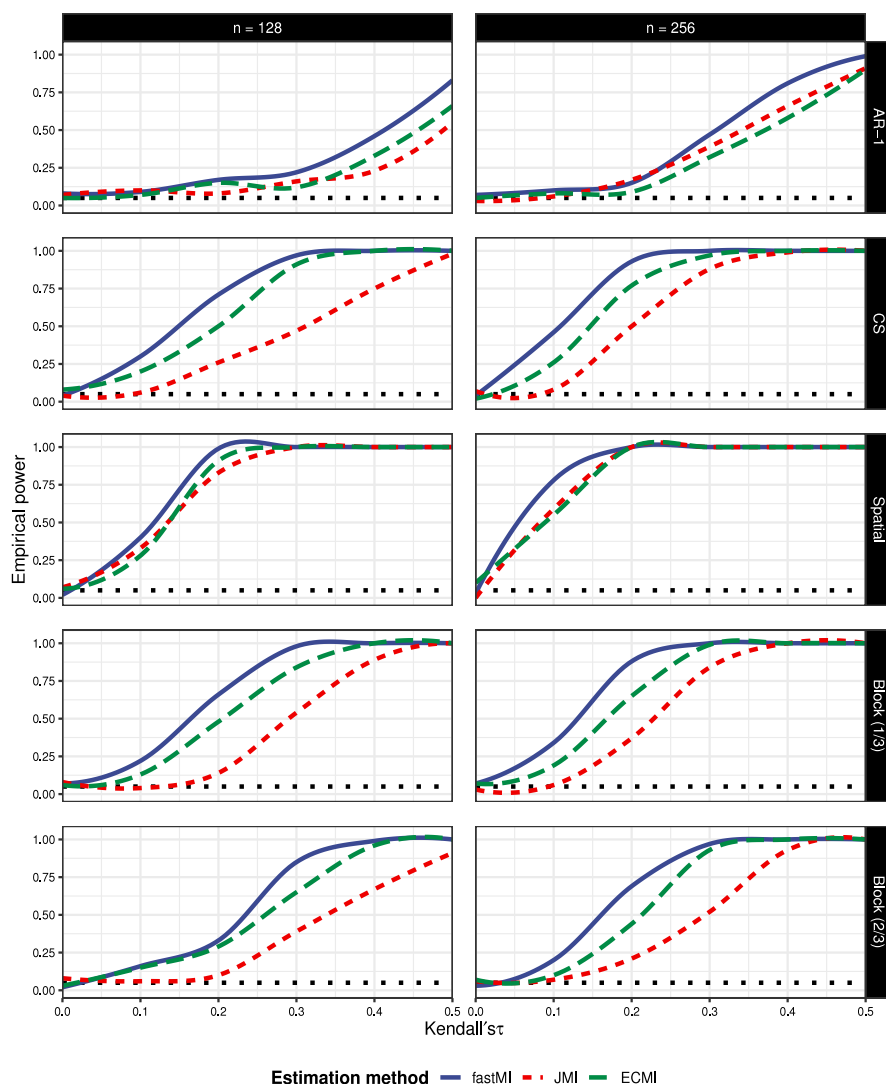
Fig. 4. Comparison on empirical power of competing tests for independence via mutual information (MI) for different configurations of multivariate copula families with varying sample sizes. The dotted black line parallel to the *x*-axis denotes the specified level of significance $\alpha = 0.05$. The estimation methods compared here are the empirical copula-based plugin MI (ECMI), the fast Fourier transform based-MI (fastMI) and the jackknifed MI (JMI).

Table 5
Mean (standard deviation) computation time (in seconds) of *JMI* and fastMI for bivariate data of varying sample size (n) for $s = 100$ iterations.

	Sample size (n)				
	250	500	1000	2500	5000
ECMI	0.695 (0.141)	1.599 (0.281)	4.360 (0.356)	5.368 (0.308)	6.404 (0.254)
JMI	**0.114 (0.043)**	**0.659 (0.200)**	3.150 (0.107)	18.446 (0.116)	62.454 (4.601)
fastMI	0.355 (0.088)	0.750 (0.213)	**1.199 (0.135)**	**2.964 (0.081)**	**5.952 (0.125)**

6. Data analysis

To illustrate an application of our method in practice, we re-analyze the dependence between X = 'death rate' and Y = 'birth rate' in 229 countries and territories around the world in the first trimester of 2020 [11]. This dependence was previously found in [11] to be complex with a departure from linearity and monotonicity. (X, Y) denote the number of deaths and births per year per 1000 individuals in the country. Fig. 5 presents a scatterplot of the data (X_i, Y_i) denoting the number of deaths and births per year per 1000 individuals in the *i*th country, indicated by one point on the plot. It demonstrates an interesting 'C' shape. For ease of
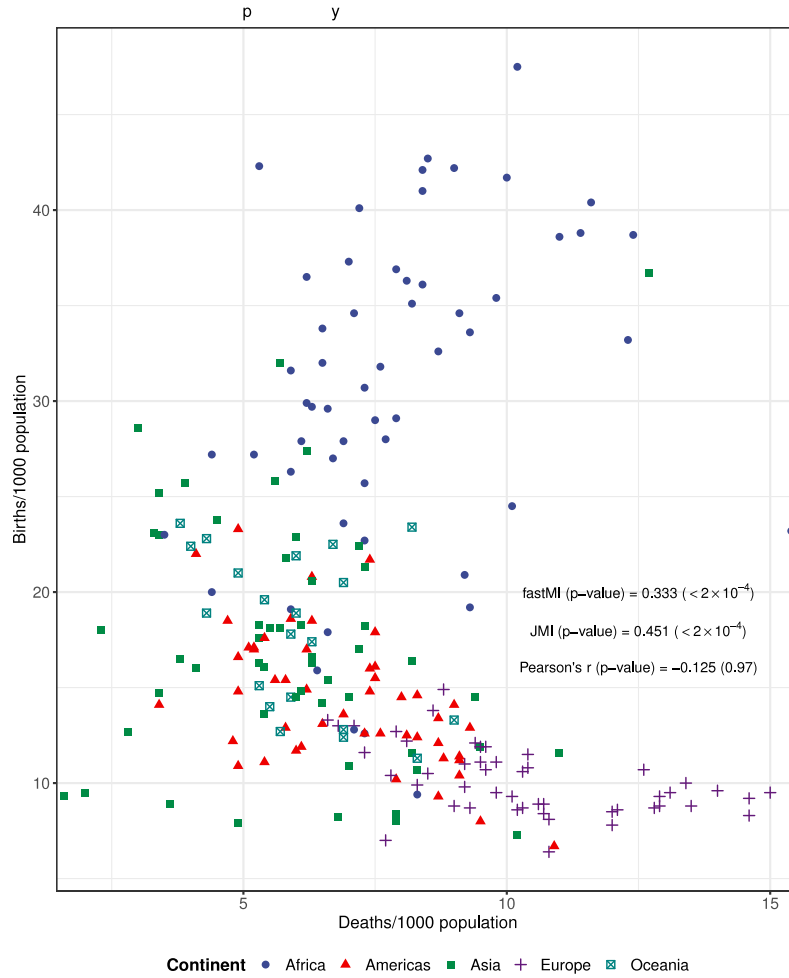
Fig. 5. Birth rate versus death rate (per thousand individuals) around the world in 2020, where countries are color-/shape-coded by their continent. The data are available in the R package `HellCor`.

exposition, we stratified countries by the continent they belong to, marked by different symbols in the figure. Note that we clubbed North American and South American countries together in Fig. 5; this clustering has no role to play in our analysis and is used only for ease of visualization. We may, indeed by simple data visualization, expect a strong association between these two variables whose data points are not randomly distributed but show a clear C-shape.

A closer investigation by [11] reveals the presence of two possible and opposite relationship patterns. First, it shows an 'decreasing' trend from 'moderate' birth rate to 'low' birth rate as death rate increases, for the industrialized countries (mostly Europe, Oceania and the Americas). Second, an 'increasing' trend is present from 'moderate' birth rate to 'high' birth rate as death rate increases, mostly for African nations. The downward trend is more pronounced, while the increasing trend is more diffused.

The strength of such nonlinear dependence and non-monotonic patterns is hardly captured by Pearson's correlation whose estimate is $\hat{r} = -0.125$, implying an insignificant association with p-value equal to 0.97 at level $\alpha = 0.05$. In contrast, both `fastMI` (estimate of $MI = 0.333$) and JMI (estimated $MI = 0.451$) report significant association with $r = 5000$ permutations. We find the p-values for both statistic to be less than 2×10^{-4} via the permutation testing approach at level $\alpha = 0.05$.

We further note the disparity between estimated values of MI using `fastMI` and JMI estimators, respectively. Based on our knowledge learned from the simulation studies in Section 5.1, namely `fastMI`-based estimate having reduced estimation error, it is rational to deduce that the JMI method overestimated the strength of association between death rate and birth rate from the data, when compared with our `fastMI` method. This may lead to erroneous inference in other applications. This example illustrates the advantage of `fastMI` as being a useful and meaningful measure of dependence to capture complex nonlinear relationships.

7. Concluding remarks

This paper develops a fast and consistent MI estimator through a powerful nonparametric copula estimation approach. Through extensive simulation studies, we have demonstrated that the proposed nonparametric method has several desirable properties,

outperforming the current gold standard, the JMI as well as the benchmark $ECMI$. Since the proposed method is nonparametric, it does not assume any parametric copula models and works under minimal model assumptions about relationships. An advantage of the MI-based approach is that the quantification of dependence is little dependent on marginal distributions, in which by virtue of its focus on the copula, the proposed method only takes into account the intrinsic association between variables without suffering from potential irregularities in the marginal distributions. An appealing technical advance pertains to that we overcome issues that plague bandwidth-based MI estimators, such as bandwidth selection and slow computation times for large data sets. `fastMI` relies on a data-adaptive Fast Fourier transformation-based approach to nonparametrically estimating the underlying copula structures. Its run time is many-fold faster than both $ECMI$ as well as JMI for large data sets. `fastMI` exhibits reduced estimation error and provides a more powerful test of independence than both $ECMI$ and JMI.

While typically used to study pairwise association, MI may be extended in a multivariate setting. However, as evidenced by the discussion below, MI may lose some of its original interpretability, which affects its legitimate use in testing for independence. Let consider a d-dimensional random vector $\boldsymbol{X} = \left(X_1, \ldots, X_d\right)^\top$ with joint density function $f_{\boldsymbol{X}}$ with marginal density functions $\{f_1, \ldots, f_d\}$. It follows from Sklar's theorem that there exists a unique associated copula density given by

$$c\left(u_1, \ldots, u_d\right) = \frac{f_{\boldsymbol{X}}\left(F_1^{-1}\left(u_1\right), \ldots, F_d^{-1}\left(u_d\right)\right)}{\prod_{i=1}^d f_i\left(F_i^{-1}\left(u_i\right)\right)}.$$

Here copula density $c(u_1, \ldots, u_d)$ can in turn, be used to define $MI(X_1, \ldots, X_d)$, known as the multivariate mutual information (MMI). Other names of MMI include total correlation [35] or multi-information [30]. Basically, MMI quantifies the amount of information shared among the different random variables, which characterizes the relatedness of the random variables in a group. [32] proved the possible negativity of MMI when $d \geq 3$. Thus, MMI loses one key property of non-negativity, which makes MMI illegitimate to quantify the size of information (which is believed to be non-negative). MMI is deemed undesirable for its role as a metric of information volume.

Nevertheless, mathematically, MMI may still be used as a tool to test for independence. From Theorem 2 of [1], we know that d variables $X_1, \ldots, X_d$ are mutually independent if and only if all MMIs of $2^d - d - 1$ sub-vectors vanish, i.e., $MI(X_{i_1}, \ldots, X_{i_k}) = 0$ with $d \geq k \geq 2$, where the indices $\{i_1, \ldots, i_k\}$ are a subset of size k of $\{1, \ldots, d\}$. This would lead to an unduly tedious testing procedure if MMIs were used to investigate independence. Some alternative forms of MI are worth future exploration with no sacrifice of non-negativity and avoidance of exhaustive tests.

In conclusion, due to its adaptability to complex associations and robustness to sample size, `fastMI` may be potentially applicable in a wide range of practical problems where complex non-linearities [26] are of particular interest. To encourage use of our proposed toolkit for estimation and testing purposes, we developed an R package `fastMI` that is publicly available on Github.

8. Technical details

Proof of Theorem 1. *Large sample behavior of the optimal density estimator.* Here, we prove Theorem 1 which establishes strong consistency of the estimator $\hat{f}_{SC}$ to the true density $f_{\boldsymbol{Z}}$ as $n \to \infty$ at all points on the support of $f_{\boldsymbol{Z}}$.

Note the frequency filter A_n, its complement $\bar{A}_n$ and its volume $\mathcal{V}(A_n)$ are described in Remarks 1 and 2 in Section 3.2. Since the true density $f_{\boldsymbol{z}}$ and the estimator $\hat{f}_{SC}$ are both square-integrable, we can express them in terms of their corresponding Fourier transforms ϕ and $\hat{\phi}$ respectively. Since the characteristic function is integrable, we have, $\int |\phi(t)| \, dt < \infty$. Through the following sequence of inequalities, we are able to establish an upper bound for the absolute error $\left|\hat{f}_{SC}(z) - f_{\boldsymbol{z}}(z)\right|$ for any $z \in \mathcal{Z}$. By definition, note that $\hat{\phi}(t) = 0$ for $t \notin A_n$. To establish Theorem 1, it is sufficient to show that the upper bound of the absolute error given below tends to zero as $n \to \infty$. We have:

$$\left|\hat{f}_{SC}(z) - f_{\boldsymbol{z}}(z)\right| = \left|(2\pi)^{-d} \int_{\mathbb{R}^d} \exp(-\mathrm{i}t'z) \left\{\hat{\phi}(t) - \phi(t)\right\} \mathrm{d}t\right| \leq (2\pi)^{-d} \int_{\mathbb{R}^d} |\exp(-\mathrm{i}t'z)||\hat{\phi}(t) - \phi(t)|\mathrm{d}t$$

$$= (2\pi)^{-d} \int_{\mathbb{R}^d} |\hat{\phi}(t) - \phi(t)|\mathrm{d}t = (2\pi)^{-d} \int_{A_{n_1}} |\hat{\phi}(t) - \phi(t)|\mathrm{d}t + (2\pi)^{-d} \int_{\bar{A}_n} |\phi(t)|\mathrm{d}t$$

$$\leq (2\pi)^{-d} \int_{A_n} |\hat{\phi}(t) - C(t)|\mathrm{d}t + (2\pi)^{-d} \int_{A_n} |C(t) - \phi(t)|\mathrm{d}t + (2\pi)^{-d} \int_{\bar{A}_n} |\phi(t)|\mathrm{d}t := D_1 + D_2 + D_3. \tag{17}$$

Under the assumptions, $\lim_{n\to\infty} \mathcal{V}\left(A_n\right) = \infty$ and $\lim_{n\to\infty} \mathcal{V}\left(\bar{A}_n\right) = 0$. Consequently, the second term in (17), $D_2 \to 0$ as $n \to \infty$ due to Theorem 1 of [6]. Further, $D_3 \leq \mathcal{V}\left(\bar{A}_n\right)/(2\pi)^d$, since $|\phi(t)| \leq 1$. Consequently, $D_3 \to 0$ as $n \to \infty$. To prove $D_1 \to 0$ as $n \to \infty$, we first consider the two following disjoint sets,

$$B_n^+ = \{t : |C(t)|^2 \geq 4(n-1)/n^2\}, \quad B_n^- = \{t : |C(t)|^2 < 4(n-1)/n^2\}.$$

Using (14), we rewrite the first integral D_1 as follows

$$D_1 = (2\pi)^{-d} \int_{A_n \cap B_n^+} |C(t)| \left(1 - \frac{n}{2(n-1)}\left[1 + \sqrt{1 - \frac{4(n-1)}{|nC(t)|^2}}\right] \mathrm{d}t\right) + (2\pi)^{-d} \int_{A_n \cap B_n^-} |C(t)|\mathrm{d}t := D_4 + D_5.$$

The first term D_4 may be simplified by noting that for $t \in B_n^+$, we have $|C(t)|^2 \geq 4(n-1)/n^2$. This ensures a non-negative argument under the square root operation. Using the inequality $\sqrt{1-x} + \sqrt{x} \geq 1$ for $0 \leq x \leq 1$ for D_4, and using the inequality

$|C(t)| \leq \sqrt{4(n-1)/n^2}$ for $t \in B_n^-$, we establish that D_1 is bounded as follows:

$$
\begin{aligned}
D_1 = D_4 + D_5 &\leq (2\pi)^{-d} \int_{A_n \cap B_n^+} \left\{ \frac{1}{\sqrt{n-1}} - \frac{|C(t)|}{n-1} \right\} dt + (2\pi)^{-d} \frac{\sqrt{4(n-1)}}{n} \int_{A_n \cap B_n^-} dt \\
&\leq (2\pi)^{-d} \left\{ \frac{1}{\sqrt{n-1}} + \frac{1}{n-1} \right\} \int_{A_n \cap B_n^+} dt + (2\pi)^{-d} \frac{\sqrt{4(n-1)}}{n} \int_{A_n} dt \\
&= (2\pi)^{-d} \left\{ \frac{1}{\sqrt{n-1}} + \frac{1}{n-1} \right\} \mathcal{V}(A_n \cap B_n^+) + (2\pi)^{-d} \frac{\sqrt{4(n-1)}}{n} \mathcal{V}(A_n \cap B_n^-) \\
&\leq (2\pi)^{-d} \left\{ \frac{1}{\sqrt{n-1}} + \frac{1}{n-1} + \frac{\sqrt{4(n-1)}}{n} \right\} \mathcal{V}(A_n) \leq (2\pi)^{-d} \left\{ \frac{1}{\sqrt{n-1}} + \frac{1}{n-1} + \frac{2}{\sqrt{n-1}} \right\} \mathcal{V}(A_n). \quad (18)
\end{aligned}
$$

The assumptions in Theorem 1 include $\mathcal{V}(A_n)/\sqrt{n} \to 0$ as $n \to \infty$, which ensures that the upper bound in (18) tends to zero for large n. In summary, assuming $\mathcal{V}(A_n) \to \infty$, $\mathcal{V}(\bar{A}_n) \to 0$, and $\mathcal{V}(A_n)/\sqrt{n} \to 0$ as $n \to \infty$, we have $\left| \hat{f}_{SC}(z) - f_z(z) \right| \to 0$ for every $z \in \mathcal{Z}$. In fact, we can claim $\sup_{z \in \mathcal{Z}} \left| \hat{f}_{SC}(z) - f_z(z) \right| \to 0$ as $n \to \infty$. Further, assuming the true density f_z is continuous on $\mathcal{Z}$ we have uniform strong convergence of $\hat{f}_{SC}$ to f_Z as $n \to \infty$. $\square$

Proof of Theorem 2. *Consistency of* $\widehat{MI}_{fast}$. Here, we prove Theorem 2 which establishes that the estimator $\widehat{MI}_{fast}$ given by (3) converges in probability to the true MI in (1). For this proof we define the oracle estimator

$$
MI_{oracle} = n^{-1} \sum_{i=1}^{n} \ln \left\{ \frac{c_{U_X U_Y}(U_{Xi}, U_{Yi})}{c_{U_X}(U_{Xi}) c_{U_Y}(U_{Yi})} \right\}.
$$

Note that by the law of large numbers MI_{oracle} is consistent for MI. From Theorem 1 we know that for any small ϵ, there exists a large enough N such that $\sup_{u \in (0,1)^d} |\hat{c}(u) - c(u)| < \epsilon$. Using Taylor series expansion to $\ln \{\hat{c}(u_i)/c(u_i)\}$, we get $\ln \{\hat{c}(u_i)\} = \ln \{c(u_i)\} + \frac{\hat{c}(u_i)-c(u_i)}{c(u_i)} + o(\epsilon)$, where we ignore the last term. Summing the expression above over N observations, we get the following chain of inequalities:

$$
\left| \frac{1}{N} \sum_{i=1}^{N} \ln \{\hat{c}(u_i)\} - \frac{1}{N} \sum_{i=1}^{N} \ln \{c(u_i)\} \right| \leq \frac{1}{N} \sum_{i=1}^{n} \left| \frac{\hat{c}(u_i) - c(u_i)}{c(u_i)} \right| \leq \frac{\sup_i |\hat{c}(u_i) - c(u_i)|}{c_0} \leq \frac{\epsilon}{c_0},
$$

where the true copula density is bounded below by c_0. The argument above can be repeated individually for all three of the underlying copula density functions $c_{U_X U_Y}$, c_{U_X}, and c_{U_Y} and their respective estimates $\hat{c}_{U_X U_Y}$, $\hat{c}_{U_X}$, and $\hat{c}_{U_Y}$. Finally, we can write

$$
\begin{aligned}
\left| \widehat{MI}_{fast} - MI_{oracle} \right| &\leq \left| \frac{1}{N} \sum_{i=1}^{N} \ln \left\{ \hat{c}_{U_X U_Y}(U_{Xi}, U_{Yi}) \right\} - \frac{1}{N} \sum_{i=1}^{N} \ln \left\{ c_{U_X U_Y}(U_{Xi}, U_{Yi}) \right\} \right| \\
&+ \left| \frac{1}{N} \sum_{i=1}^{N} \ln \left\{ \hat{c}_{U_X}(U_{Xi}) \right\} - \frac{1}{N} \sum_{i=1}^{N} \ln \left\{ c_{U_X}(U_{Xi}) \right\} \right| + \left| \frac{1}{N} \sum_{i=1}^{N} \ln \left\{ \hat{c}_{U_Y}(U_{Yi}) \right\} - \frac{1}{N} \sum_{i=1}^{N} \ln \left\{ c_{U_Y}(U_{Yi}) \right\} \right| \leq \epsilon^*,
\end{aligned}
$$

where the last step follows from the intermediate inequality obtained above. Hence, as MI_{oracle} is consistent for MI, we can claim $\widehat{MI}_{fast}$ is consistent for MI as well. $\square$

CRediT authorship contribution statement

Soumik Purkayastha: Literature review, Technical proofs, Paper writing, Numerical analyses. **Peter X.-K. Song:** Literature review, Technical proofs, Paper writing, Conceptual framework.

Acknowledgments

References

[1] P. Baudot, M. Tapia, D. Bennequin, J.-M. Goaillard, Topological information data analysis, Entropy 21 (9) (2019) 869.
[2] A. Bernacchia, S. Pigolotti, Self-consistent method for density estimation, J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (3) (2011) 407–422.
[3] P. Billingsley, Probability and Measure, third ed., in: A Wiley-Interscience publication, Wiley, New York [u.a.], 1995.
[4] M. Bolbolian Ghalibaf, Relationship between Kendall's tau correlation and mutual information, Rev. Colombiana Estadística 43 (1) (2020) 3–20.
[5] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 2005.
[6] S. Csörgő, V. Totik, On how long interval is the empirical characteristic function uniformly consistent? Acta Sci. Math. (Szeged) 45 (1–4) (1983) 141–149.

[7] C. Czado, Analyzing Dependent Data with Vine Copulas: A Practical Guide with R, Paperback ed., in: Lecture Notes in Statistics, vol. 222, Springer, New York, 2019, p. 271.

[8] C.O. Daub, R. Steuer, J. Selbig, S. Kloska, Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data, BMC Bioinformatics 5 (1) (2004) 1–12.

[9] B.S. Everitt, A. Skrondal, The Cambridge Dictionary of Statistics, fourth ed., Cambridge University Press, Cambridge, England, 2010.

[10] G. Geenens, A. Charpentier, D. Paindaveine, Probit transformation for nonparametric kernel estimation of the copula density, Bernoulli (2017).

[11] G. Geenens, P.L. de Micheaux, The Hellinger correlation, J. Amer. Statist. Assoc. 117 (538) (2020) 639–653.

[12] N.-B. Heidenreich, A. Schindler, S. Sperlich, Bandwidth selection for kernel density estimation: a review of fully automatic selectors, AStA Adv. Stat. Anal. 97 (4) (2013) 403–433.

[13] R. Heller, Y. Heller, M. Gorfine, A consistent multivariate test of association based on ranks of distances, Biometrika 100 (2) (2012) 503–510.

[14] H. Joe, Dependence Modeling with Copulas, Hardcover ed., Chapman & Hall/CRC Monographs on Statistics and Applied Probability, Chapman and Hall/CRC, New York, 2014, p. 480.

[15] J.B. Kinney, G.S. Atwal, Equitability, mutual information, and the maximal information coefficient, Proc. Natl. Acad. Sci. 111 (9) (2014) 3354–3359.

[16] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (6) (2004).

[17] E. Kreyszig, Advanced Engineering Mathematics, Loose Leaf ed., Wiley, New York, 2020, p. 1280.

[18] B.F. Manly, Randomization, Bootstrap and Monte Carlo Methods in Biology, Chapman and Hall/CRC, 2018.

[19] J.S. Marron, D. Ruppert, Transformations to reduce boundary bias in kernel density estimation, J. R. Stat. Soc. Ser. B Stat. Methodol. 56 (4) (1994) 653–671.

[20] Y.-I. Moon, B. Rajagopalan, U. Lall, Estimation of mutual information using kernel density estimators, Phys. Rev. E 52 (1995) 2318–2321.

[21] T.A. O'Brien, K. Kashinath, N.R. Cavanaugh, W.D. Collins, J.P. O'Brien, A fast and objective multidimensional kernel density estimation method: fastKDE, Comput. Statist. Data Anal. 101 (2016) 148–160.

[22] L. Paninski, Estimation of entropy and mutual information, Neural Comput. 15 (6) (2003) 1191–1253.

[23] A.M. Peter, A. Rangarajan, Maximum likelihood wavelet density estimation with applications to image and shape matching, IEEE Trans. Image Process. 17 (4) (2008) 458–468.

[24] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, P.C. Sabeti, Detecting novel associations in large data sets, Science 334 (6062) (2011) 1518–1524.

[25] C.P. Robert, Monte Carlo Statistical Methods, Paperback ed., Springer Texts in Statistics, Springer, New York, 2010, p. 679.

[26] H. Safaai, A. Onken, C.D. Harvey, S. Panzeri, Information estimation using nonparametric copulas, Phys. Rev. E 98 (5) (2018).

[27] C. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.

[28] B. Silverman, Density Estimation for Statistics and Data Analysis, Routledge, New York, 2018.

[29] S.P. Strong, R. Koberle, R.R. de Ruyter van Steveninck, W. Bialek, Entropy and information in neural spike trains, Phys. Rev. Lett. 80 (1998) 197–200.

[30] M. Studený, J. Vejnarová, The multiinformation function as a tool for measuring stochastic dependence, in: M.I. Jordan (Ed.), Learning in Graphical Models, Springer Netherlands, Dordrecht, 1998, pp. 261–297, URL.

[31] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, Ann. Statist. 35 (6) (2007).

[32] H.K. Ting, On the amount of information, Theory Probab. Appl. 7 (4) (1962) 439–447.

[33] N.G. Ushakov, Selected Topics in Characteristic Functions, De Gruyter, Berlin, Boston, 1999.

[34] Q. Wang, S.R. Kulkarni, S. Verdu, Divergence estimation for multidimensional densities via k-nearest-neighbor distances, IEEE Trans. Inf. Theory 55 (5) (2009) 2392–2405.

[35] S. Watanabe, Information theoretical analysis of multivariate correlation, IBM J. Res. Dev. 4 (1) (1960) 66–82.

[36] G.S. Watson, M.R. Leadbetter, On the estimation of the probability density, I, Ann. Math. Stat. 34 (2) (1963) 480–491.

[37] N. Wiener, Extrapolation, Interpolation, and Smoothing of Stationary Time Series, The MIT Press, 1964.

[38] X. Zeng, Y. Xia, H. Tong, Jackknife approach to the estimation of mutual information, Proc. Natl. Acad. Sci. 115 (40) (2018) 9956–9961.