Information estimation using nonparametric copulas

Houman Safaai, Arno Onken, Christopher D. Harvey, and Stefano Panzeri

Information Estimation Using Non-Parametric Copulas

Houman Safaai,[1, 2, *] Arno Onken,[3] Christopher D. Harvey,[1] and Stefano Panzeri[2]

[1] *Department of Neurobiology, Harvard Medical School, Boston, MA*
[2] *Istituto Italiano di Tecnologia, Rovereto, Italy*
[3] *School of Informatics, University of Edinburgh, Edinburgh, UK*
(Dated: October 9, 2018)

Estimation of mutual information between random variables has become crucial in a range of fields, from physics to neuroscience to finance. Estimating information accurately over a wide range of conditions relies on the development of flexible methods to describe statistical dependencies among variables, without imposing potentially invalid assumptions on the data. Such methods are needed in cases that lack prior knowledge of their statistical properties and that have limited sample numbers. Here we propose a powerful and generally applicable information estimator based on non-parametric copulas. This estimator, called the non-parametric copula-based estimator (NPC), is tailored to take into account detailed stochastic relationships in the data independently of the data's marginal distributions. The NPC estimator can be used both for continuous and discrete numerical variables and thus provides a single framework for the mutual information estimation of both continuous and discrete data. By extensive validation on artificial samples drawn from various statistical distributions, we found that the NPC estimator compares well against commonly used alternatives. Unlike methods not based on copulas, it allows an estimation of information that is robust to changes of the details of the marginal distributions. Unlike parametric copula methods, it remains accurate regardless of the precise form of the interactions between the variables. In addition, the NPC estimator had accurate information estimates even at low sample numbers, in comparison to alternative estimators. The NPC estimator therefore provides a good balance between general applicability to arbitrarily shaped statistical dependencies in the data and shows accurate and robust performance when working with small sample sizes. We anticipate that the non-parametric copula information estimator will be a powerful tool in estimating mutual information between a broad range of data.

I. INTRODUCTION

Mutual Information, the fundamental mathematical quantity of information theory, provides a universal way to quantify dependencies, transmission rates, and representations of data [1]. It has become an indispensable tool in many domains such as signal processing, data compression, finance, dynamical systems and neuroscience [2–7]. Mutual information quantifies the information that one random variable carries about another by measuring the reduction in uncertainty about a given variable from knowing another variable. Uncertainty in turn is quantified by means of entropy. Shannon's entropy therefore is at the core of virtually all applications of information theory.

Quantifying entropy and information of a random variable poses a difficult problem because it requires knowledge about its probability distribution. In most practical applications, the exact shape of the distribution of a random variable is unknown and thus needs to be estimated from data. This requires either strong parametric assumptions, such as assuming for instance that data follow a normal distribution, or large amounts of data to estimate the distribution directly from the samples.

In an ideal case, information estimators would estimate variable distributions directly from the data and would

not require parametric assumptions that could impose invalid structures on the data. In addition, ideal estimators would be accurate also in situations with limited sample numbers. Furthermore, given that mutual information quantifies only the dependencies between the variables [8–11], ideal estimators should be sensitive only to the dependencies between the random variables of interest, which fully define mutual information, and should be insensitive to other aspects of the data, such as the marginal distributions of the individual random variables. To date, it has been challenging to develop information estimators that have all these properties. It is clear that developing such estimators would greatly increase the range of applicability and the accuracy of information measures over a wide range of important empirical problems.

For continuous random variables, powerful estimators have been developed that estimate mutual information directly from the samples in a non-parametric way. One popular class is based on the k-nearest neighbor (kNN) estimators [12–14], which in their original form assume local uniformity in the vicinity of each point. For accurate information estimation with these approaches, the required number of samples scales exponentially with the value of mutual information [15]. This has limited the effectiveness of these estimators in cases with strong dependencies and thus high mutual information, or in situations with smaller numbers of samples. The performance of these estimators, especially for strong dependency cases, has been improved through the introduction of a

---
* houman_safaai@hms.harvard.edu.

correction term for local non-uniformity (LNC)[15]. The LNC method assumes a particular non-uniformity structure of the distribution in the kNN ball or max-norm rectangle. These assumptions, however, can produce inaccuracies in information estimates for data with marginal distributions with long tails, such as the gamma distribution, or distributions with sharp boundaries. Thus, assumptions about local non-uniformity could lead to different estimates of mutual information for two sets of variables that have similar dependency structures, and hence similar mutual information values, but different marginal distributions. These methods therefore encounter a significant trade-off between assumptions imposed on the distribution of the data and the number of samples required for accurate information estimation.

For discrete variables, estimation methods have been proposed based on either subtracting out an analytical approximation to the limited sampling bias [16, 17], or in using a Bayesian approach. In the latter, instead of estimating the probability mass function, a prior, in the form of Dirichlet distributions, is placed over the space of discrete probability distributions. The entropy then is estimated using the inferred posterior distribution over entropy values [18, 19]. A more complete set of priors has been recently proposed in [20], using a mixture of Pitman-Yor priors (PYM), which is a two-parameter generalization of the Dirichlet process and parameterized to be flat over entropy values. It has been shown that such a flat prior provides better estimates of entropy and mutual information with low sample numbers compared to analytical bias subtraction methods [18, 20]. However, like the LNC estimator, the PYM estimator is sensitive to the form of the marginal distributions. In particular, Gerlach et al. [21] confirmed that the PYM estimator reduces the estimation bias but that the bias scales in the same way with the number of samples as for other type of estimators. Moreover, the PYM estimator performs worse on heavy-tailed distributions [21, 22].

The previously proposed estimators considered above have in common that in one way or another they make use of the full joint distribution of the random variables of interest, which includes contributions from both the marginal distributions and the dependencies between the variables of interest. However, because mutual information is determined only by the dependencies between variables [8–10], information estimators only need to focus on correctly capturing the dependency structure. Such dependency structures are best isolated using the mathematical construct known as the copula. Formally, any joint distribution can be decomposed into its marginal distributions and a copula. The latter quantifies the dependency structure irrespective of the marginals, and the negative of the copula entropy exactly equals the mutual information that one random variable carries about the other [8, 23]. Copula-based methods are therefore well-suited for isolating the dependencies and are insensitive to the form of marginal distributions. Previous copula based information estimators have been proposed both

in the continuous domain [9–11] and mixtures between discrete and continuous domains [24]. All such copula based information estimators have made use of copulas selected from parametric families. The parametric copula estimators have the advantage of simplicity, but the disadvantage that they make systematic assumptions on the dependency structure of the data [25]. These assumptions might differ greatly from the real data structure, leading to large estimation errors when used on datasets with complex and non-linear dependency structures that are difficult to fit with simple parametric copula families. However, recently some non-parametric copula estimation methods have been proposed in [26–28], and their properties in density estimation have been studied. Yet, a systematic study of the application of such methods in mutual information estimation is lacking.

Here, we propose information estimators based on non-parametric copulas (NPC). These NPC estimators first identify the copula that characterizes the relationship between the random variables of interest and then calculate the entropy of the copula to obtain an estimate of the mutual information. Contrary to parametric copula families, non-parametric copulas do not impose strong assumptions on the shape of the stochastic relationship between the variables of interest and thereby avoid systematic biases in the information estimates. We present methods to identify the copula non-parametrically, both for continuous and discrete data. We show that, compared to previously reported information estimators (in particular the LNC and PYM estimators), NPC mutual information estimators are robust to the parameters of the marginal distribution and perform well in cases with low sample numbers. NPC-based estimators are therefore some of the first information estimators that simultaneously do not impose strong parametric assumptions, can work with relatively small sample sizes, and isolate the dependencies in the data that matter for mutual information both in the continuous, discrete or mixed domains.

II. THEORY AND METHODOLOGY

We estimate information by means of copulas and their entropy. Copulas mathematically formalize the concept of statistical dependencies: a given copula quantifies a particular relationship between a set of random variables. Here we give a brief summary of the basics of the copula and its relation to mutual information. We then continue by presenting the non-parametric copula and how it can be computed empirically from given data.

A. Formal copula definition

A d-dimensional copula is the cumulative distribution function $C(u_1, \ldots, u_d) : [0,1]^d \to [0,1]$ of a random vector defined on the unit hypercube $[0,1]^d$ with uniform

marginals $\mathcal{U}_{[0,1]}$ over $[0,1]$.

$$C(u_1, \ldots, u_d) = P(U_1 \leq u_1, \ldots, U_d \leq u_d), * \quad (1)$$

where $U_i \sim \mathcal{U}_{[0,1]}$.

The great strength of copulas is their utility for representing the statistical relationship between multiple random variables. Copulas can be used to couple arbitrary marginal cumulative distribution functions (CDFs) to form a joint CDF. Sklar's theorem [23, 29] lays out the theoretical foundations for this construction:

Theorem 1 *Sklar's theorem*: *For a d-dimensional random vector* $\boldsymbol{X} = (X_1, \ldots, X_d)$, *let* $F_{\boldsymbol{X}}$ *be its CDF with marginals* $F_1, \ldots, F_d$. *Then there exists a copula* C *such that* $\forall \boldsymbol{x} \in \mathbb{R}^d$:

$$F_{\boldsymbol{X}}(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)), \quad x_i \in \mathbb{R}, \quad (2)$$

C *is unique, if the marginals* F_i *are continuous. Conversely, if* C *is a copula and* $F_1, \ldots, F_d$ *are CDFs, then the function* $F_{\boldsymbol{X}}$ *defined by* $F_{\boldsymbol{X}}(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$ *is a d-dimensional CDF with marginals* $F_1, \ldots, F_d$.

Sklar's theorem relates the copula C of Eq.(1) to the joint distribution function of the variables $U_i = F_i(X_i)$,

$$C(u_1, \ldots, u_d) = F_{\boldsymbol{X}}(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)), \quad u_i \in [0,1] \quad (3)$$

where F_i^{-1} are the inverse cumulative distribution functions. For a differentiable copula C, we can define the copula probability density function (PDF) $c(u_1, \ldots, u_d) = \frac{\partial^d}{\partial u_1 \ldots \partial u_d} C(u_1, \ldots, u_d)$. For $u_i := F_i(x_i)$ and $f_i(\cdot)$ as PDFs corresponding to the CDFs $F_i(\cdot)$, we can write the copula density as

$$c(u_1, \ldots, u_d) = \frac{f_{\boldsymbol{X}}\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right)}{\prod_{i=1}^d f_i(F_i^{-1}(u_i))}. \quad (4)$$

This means that the multivariate PDF can be decomposed into the copula density and the product of the marginal densities. The copula can be interpreted as the part of the density function that is independent from the single variable marginals and rather captures the dependencies between the variables. This decomposition is useful to estimate the joint density function and also to estimate the likelihood which is needed in statistical inference, but here in this work we only focus on the copula density as a tool to compute entropy and mutual information.

An example bivariate density function is shown in Fig.1 which consists of a gamma marginal distribution (x_1), a Gaussian marginal distribution (x_2), and a particular parametric copula density (student-t copula) as its dependency structure. The decomposition of the full density function into the dependency structure (copula) and the marginal distributions makes it possible to study any measure which is independent from the marginal distributions by considering only the copula structure. Here
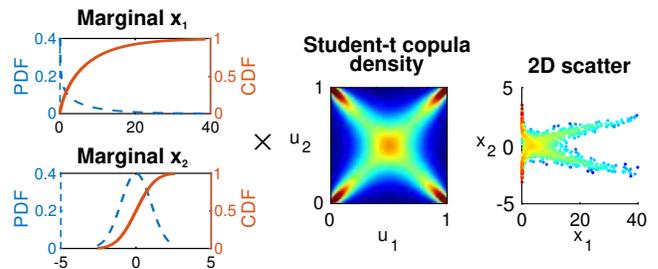
FIG. 1. A bivariate dataset (right panel) is generated from mixing two marginal distributions (left panels, gamma top and Gaussian bottom). The PDF (dashed line) and CDF (solid line) for the two marginals are shown. They merge with the copula density (middle panel) to generate the joint density function.

the gamma marginal distribution has a sharp boundary at $x_1 = 0$ which makes it difficult for conventional density estimation methods to compute the full bivariate density function. The copula, on the other hand, can easily cope with the density behavior at $x_1 = 0$.

B. Entropy and mutual information

Entropy quantifies the uncertainty associated with a given random variable and lays the foundation for mutual information. For a continuous multivariate distribution, the differential entropy $h(\boldsymbol{X})$ is defined as

$$h(\boldsymbol{X}) = - \int f_{\boldsymbol{X}}(\boldsymbol{x}) \log_2 f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}, \quad (5)$$

where $f_{\boldsymbol{X}}$ denotes the multivariate probability density function [1, 3]. With this, the mutual information $I(\boldsymbol{X}; \boldsymbol{Y})$ between two continuous multivariate random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by

$$I(\boldsymbol{X}; \boldsymbol{Y}) = h(\boldsymbol{X}) + h(\boldsymbol{Y}) - h(\boldsymbol{X}, \boldsymbol{Y}), \quad (6)$$

where $h(\boldsymbol{X}, \boldsymbol{Y})$ is the joint differential entropy of the joint distribution $(\boldsymbol{X}, \boldsymbol{Y})$ with joint PDF $f_{\boldsymbol{X}, \boldsymbol{Y}}$ [1, 3].

Using Eq.(4), one can show that the mutual information equals the negative of the entropy of the copula density between $\boldsymbol{X}$ and $\boldsymbol{Y}$ [8–11]:

$$I(\boldsymbol{X}; \boldsymbol{Y}) = -h(c) = \int_{[0,1]^d} c(\boldsymbol{u}) \log_2 c(\boldsymbol{u}) \, d\boldsymbol{u}, \quad (7)$$

where $\boldsymbol{u} = (u_1, \cdots, u_d)$. This makes the computation of mutual information independent from the marginal distributions and reduces the computational error in estimating the MI for two reasons. First, the irrelevance of the marginals removes the need for faithfully capturing their properties in the information estimation procedure. Thus, copula-based estimators separate the relevant entropy from the irrelevant entropies and thereby effectively reduce the number of implicit quantities contributing to the final mutual information estimate, thereby reducing

the estimation error. Second, the independence of copula from the marginals makes copula based methods robust to any irregularity which might exist in the marginals. This is in contrast to density dependent methods, such as kNN-based estimators [12–15] which might struggle with marginal irregularities.

We can estimate the integral Eq.(7) using classical Monte Carlo (MC) sampling [24, 30]. The entropy can be expressed as an expectation over the copula density c

$$h(c) = -\mathbb{E}_c \left[\log_2 c(\boldsymbol{U})\right], \quad (8)$$

where $\boldsymbol{U} = (U_1, \ldots, U_d)$ denotes a random vector from the copula space. This expectation can then be approximated by the empirical average over a large number of d-dimensional samples $\boldsymbol{u}_j = ((\boldsymbol{u}_j)_1, \cdots, (\boldsymbol{u}_j)_d)$ from the random vector $\boldsymbol{U}$:

$$-\mathbb{E}_c \left[\log_2 c(\boldsymbol{U})\right] \approx \widehat{h_k} := -\frac{1}{k} \sum_{j=1}^{k} \log_2(c(\boldsymbol{u}_j)) \quad (9)$$

By the strong law of large numbers, $\widehat{h_k}$ converges almost surely to $h(c)$. Moreover, we can assess the convergence of $\widehat{h_k}$ by estimating the sample variance of $\widehat{h_k}$:

$$\text{Var}\left[\widehat{h_k}\right] \approx \frac{1}{k+1} \sum_{j=1}^{k} \left(\log_2(c(\boldsymbol{u}_j)) - \widehat{h_k}\right)^2, \quad (10)$$

With this estimate, the term $\frac{\widehat{h_k} - h(c)}{\sqrt{\text{Var}\left[\widehat{h_k}\right]}}$ is approximately standard normal distributed, allowing us to obtain confidence intervals for our differential entropy estimates [30].

1. Sampling from the copula

To sample from a d-dimensional copula, we use the Rosenblatt transform [31, 32]. This approach applies a sequence of conditional distributions and makes use of the fact that the marginal distributions of a copula are always uniform. First, we draw independent uniform samples $v_1, \ldots, v_d$ from $[0, 1]$. Then, we sequentially transform these samples by means of the inverse conditional CDFs of the copula:

$$\begin{aligned}
u_1 &= v_1 \\
u_2 &= C_{2|1}^{-1}(v_2|u_1) \\
u_3 &= C_{3|1,2}^{-1}(v_3|u_1, u_2) \\
&\vdots \\
u_d &= C_{d|1,\ldots,d-1}^{-1}(v_d|u_1, \ldots, u_{d-1})
\end{aligned} \quad (11)$$

where $C_{i|1,\ldots,i-1}^{-1}$ denotes the inverse of the copula CDF of element i conditioned on the elements $1, \ldots, i-1$. The resulting vector $(u_1, \ldots, u_d)$ is a sample from the copula.

The conditional CDFs can be obtained from the copula CDF by calculating [29, 33]

$$\begin{aligned}
&C_{i|1,\ldots,i-1}(u_i|u_1, \ldots, u_{i-1}) \\
&= \frac{\partial^{i-1} C_{1,\ldots,i}(u_1, \ldots, u_i) / \prod_{k=1}^{i-1} \partial u_k}{c_{1,\ldots,i-1}(u_1, \ldots, u_{i-1})}, \quad (12)
\end{aligned}$$

where $C_{1,\ldots,i}$ denotes the copula CDF with the elements $i+1, \ldots, d$ marginalized out and $c_{1,\ldots,i}$ denotes its PDF.

For the special case $d = 2$, computation of the conditional CDF reduces to a partial derivative of the original copula CDF with respect to one variable.

C. Parametric copulas

Many parametric families of copulas have been proposed, representing various relationship shapes with different tail dependencies and symmetries [29, 33, 34]. These families are appropriate for fitting data with corresponding features. However, such parametric families make strong assumptions about the shapes of the relationships. This may in turn introduce considerable biases in information estimates when the shape of the dependencies in the real data does not match those that can be described by the copula family.

In this work we will use the parametric copulas for two different purposes. The first is to test the performance of information estimation methods based on parametric copulas. The second is to use particular parametric families to generate data with a known ground-truth information value in order to test the accuracy of our non-parametric copula-based information estimators. For this purpose, the most convenient parametric families are those for which we can analytically calculate mutual information. Two particular parametric families with known closed-form solutions for calculating mutual information are given by the Gaussian and student-t copula families. We describe their properties in this section. For our simulations, we consider only bivariate copulas. However, these copulas can be readily extended to large-dimensional copulas by means of pair-copula-constructions [35].

- **Gaussian copula family** One of the most commonly applied parametric copula is the Gaussian copula with CDF defined as $C_g(u, v) = \Phi_{\boldsymbol{\Sigma}}(\Phi^{-1}(u), \Phi^{-1}(v))$ where $u, v \in [0, 1]$ and Φ and $\Phi_{\boldsymbol{\Sigma}}$ are the univariate standard normal CDF and multivariate normal CDF with zero mean and correlation matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$, respectively. The copula PDF can be written as

$$c_g(u, v) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} X^\top (\boldsymbol{\Sigma}^{-1} - \mathbf{I}_2) X\right), \quad (13)$$

where $\boldsymbol{X} = (x, y)$, $(x, y) = (\Phi^{-1}(u), \Phi^{-1}(v))$ and $\mathbf{I}_2$ denotes the 2×2 identity matrix.

The Gaussian copula entropy has the following analytical form

$$h(c_g) = -\frac{1}{2}\log_2\left(1 - r^2\right),\qquad(14)$$

- **Student-t copula family** The student-t copula is another well established parametric copula family which can be used to model elliptical dependency structures. Contrary to Gaussian copulas, copulas from the student-t family have tail dependency and hence can be used to generate datasets with heavy tails. The bivariate student-t copula is defined by means of the standardized bivariate student-t CDF $t_{\boldsymbol{\Sigma},\nu}$ as $C_t(u,v) = t_{\boldsymbol{\Sigma},\nu}(t_\nu^{-1}(u), t_\nu^{-1}(v))$, where $\boldsymbol{\Sigma}$ is the correlation matrix and ν is the degrees of freedom. The PDF of the bivariate student-t copula is

$$c_t(u,v) = \frac{\Gamma(\frac{\nu+2}{2})\Gamma(\frac{\nu}{2})}{\sqrt{|\boldsymbol{\Sigma}|}\Gamma(\frac{\nu+1}{2})}\frac{(1 + \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}/\nu)^{-(\nu+2)/2}}{((1 + x^2/\nu)(1 + y^2/\nu))^{-(\nu+1)/2}},\qquad(15)$$

where $\boldsymbol{X} = (x,y)$, $(x,y) = (t_\nu^{-1}(u), t_\nu^{-1}(v))$ and $\Gamma(\cdot)$ denotes the gamma function.

The student-t copula has the following analytical entropy [9]:

$$h(c_t) = \frac{\Omega}{\log(2)} - \frac{1}{2}\log_2\left(1 - r^2\right),\qquad(16)$$

where

$$\Omega = 2\log\left(\sqrt{\frac{\nu}{2\pi}}\beta\left(\frac{\nu}{2}, \frac{1}{2}\right)\right)\qquad(17)$$
$$- \frac{2+\nu}{\nu} + (1+\nu)\left[\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right]$$

is a constant and $\beta(\cdot)$ and $\psi(\cdot)$ are the beta and digamma function, respectively.

D. Non-parametric copulas

Our information estimator is based on a recently developed non-parametric version of the copula, which can be used to model any general dependency structure and does not involve making assumptions on the structure of data [7, 11, 24]. One challenge in using non-parametric copula estimators is to deal with the close support of the copula: the support of a bivariate copula is restricted to the unit square $[0,1]^2$. 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. To address this problem, we apply a transformation such that the support of the density in the transformed space is unbounded [26, 27, 36].
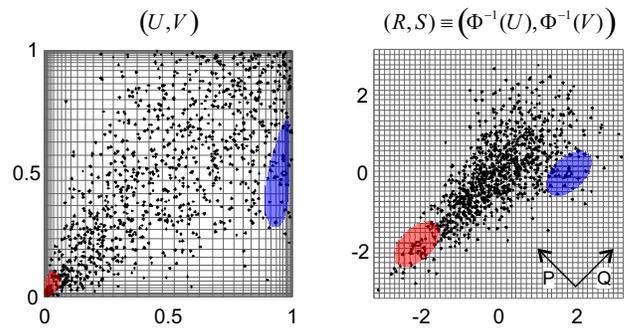
FIG. 2. (u_i, v_i) samples and their probit transformations (r_i, s_i) are shown. Grids are equally spaced in (R,S) space. The direction of the rotated (P,Q) are shown as insets. The red and blue areas are two example kernel functions.

Let us assume that we want to estimate a copula density c given n bivariate random samples (u^i, v^i), $i = 1...n$ from the random vector (U,V). Let Φ be the standard normal CDF and ϕ its density. Then the random vectors $(R,S) = (\Phi^{-1}(U), \Phi^{-1}(V))$ have normal distributed marginals with support on the full $\mathbb{R}^2$ (Fig.2). In this domain, kernel density estimators work well and have less asymptotic and boundary problems since the density slowly converges to zero on the edges. This transformation is known as the probit transformation.

By sklar's theorem for densities, Eq.(1), the density f of (r,s) will be decomposed into

$$f(r,s) = c(u,v)\phi(r)\phi(s).\qquad(18)$$

After change of variables, we get the copula density

$$c(u,v) = \frac{f(\Phi^{-1}(u), \Phi^{-1}(v))}{\phi(\Phi^{-1}(u))\phi(\Phi^{-1}(v))}.\qquad(19)$$

The non parametric copula can be estimated in several ways, described in what follows.

1. Naive kernel estimation

The naive kernel estimate of the density function can be written as

$$c_{\text{naive}}(u,v) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_{\vec{b}_n}(r - r_i, s - s_i)}{\phi(\Phi^{-1}(u))\phi(\Phi^{-1}(v))},\qquad(20)$$

where the sum is over the n samples $(r_i, s_i) \equiv (u_i, v_i)$ and (r,s) is related to (u,v) through Eq.(21). For the density kernel $K(\cdot)$ we consider a symmetric bounded probability density function with bandwidth vector $\vec{b}$. Furthermore, we can make another transformation to the principle component coordinates

$$\begin{pmatrix} p \\ q \end{pmatrix} \equiv \mathbf{W}\begin{pmatrix} r \\ s \end{pmatrix} = \mathbf{W}\begin{pmatrix} \Phi^{-1}(u) \\ \Phi^{-1}(v) \end{pmatrix},\qquad(21)$$

where the matrix $\mathbf{W}$ is the rotation matrix to the principle component coordinates. In this coordinate space, since the covariance matrix is diagonal, we can approximate the kernel function as the product of the two kernels for each of the coordinates $K_{\vec{b}}(p - p_i, q - q_i) \approx K_{b_P}(p - p_i)K_{b_Q}(q - q_i)$ where b_Q and b_P are the corresponding bandwidths of each coordinate. An example bivariate data is shown both in the (p, q) and (u, v) spaces in Fig.2.

2. Local-likelihood density estimation

When used for non-parametric copula estimation, the naive kernel estimator has asymptotic problems at the edges of the distribution support. In particular, it might find false peaks and troughs when there is an asymmetry in the tails of the distribution. This happens because small fluctuations in unbalanced tails are greatly magnified when transformed back to the copula space [26]. To remedy this problem, we can make use of a similar approach as in [37], where it was shown that the the local likelihood density estimation gives a much better behavior on the boundaries [26]. We adapted this approach by assuming that the density function can be written locally for any point (p', q') around each point (p, q) as a continuous function $f(p', q') = \psi_{\boldsymbol{\theta}(p,q)}(p - p', q - q')$ for some parameters $\boldsymbol{\theta}(p, q)$ and a continuous parametric function $\psi_{\boldsymbol{\theta}(p,q)}(p - p', q - q')$.

The log-likelihood of such an estimate can be written as follows [26, 36]

$$\mathcal{L}(p, q) = \frac{1}{n}\sum_{i=1}^{n} K_{b_p}(p_i - p)K_{b_q}(q_i - q)\log\psi_{\boldsymbol{\theta}(p,q)}(p - p_i, q - q_i) - \iint_{\mathbb{R}^2} K_{b_p}(p - \tilde{p})K_{b_q}(q - \tilde{q})\psi_{\boldsymbol{\theta}(p,q)}(p - \tilde{p}, q - \tilde{q})d\tilde{p}\,d\tilde{q}. \quad (22)$$

After fixing the functional form for ψ, the parameters $\boldsymbol{\theta}$ can be obtained by maximizing the log-likelihood

$$\boldsymbol{\theta}(p, q) = \arg\max_{a_1,\dots,a_F}\mathcal{L}(p, q), \quad (23)$$

where we considered F degrees of freedom for $\boldsymbol{\theta}$. A possible choice for the functional form of the ψ studied in [26, 36–38] is to assume that its logarithm is a polynomial. For a polynomial of order two, the $\psi_{\boldsymbol{\theta}}$ around each point (p, q) can be written as

$$\psi_{\boldsymbol{\theta}}(p - p', q - q') = a_1 e^{a_2(p-p')+a_3(q-q')+a_4(p-p')^2+a_5(q-q')^2}, \quad (24)$$

where $\boldsymbol{\theta} = (a_1, a_2, a_3, a_4, a_5)$ are the parameters to be defined at each point (p, q). Note that the local likelihood density function is equal to $f_{\text{LL}}(p, q) = a_1(p, q)$. This particular functional form simply means that locally and not globally, around each point (p, q), the log-likelihood function has a Gaussian form. The choice of the kernel functions $K(\cdot)$ are of lower importance, since they will be weighted with the local function ψ. Given that the data in the probit coordinates is normal, and has diagonal covariance matrix, the Gaussian kernel function seems to be a natural choice

$$K_b(u) = \frac{1}{\sqrt{2\pi}b}e^{-\frac{u^2}{2b^2}}, \quad (25)$$

We can now solve Eq.(23) by imposing $\frac{\delta\mathcal{L}(p,q)}{\delta\boldsymbol{\theta}} = 0$ and solving the following set of equation which we get after using Eq.(23) and Eq.(22) at each point (p, q):

$$\begin{pmatrix} f_{\text{naive}} \\ f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} := \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} 1 \\ (p_i - p) \\ (q_i - q) \\ (p_i - p)^2 \\ (q_i - q)^2 \end{pmatrix} K_{b_p}(p_i - p)K_{b_q}(q_i - q)$$

$$= \iint_{\mathbb{R}^2}\begin{pmatrix} 1 \\ \tilde{p} \\ \tilde{q} \\ \tilde{p}^2 \\ \tilde{q}^2 \end{pmatrix} K_{b_p}(\tilde{p})K_{b_s}(\tilde{q})\psi_{\boldsymbol{\theta}(p,q)}(\tilde{p}, \tilde{q})d\tilde{p}d\tilde{q} \quad (26)$$

The set of equations Eqs.(26) can be solved analytically for the Gaussian kernel as follows:

$$\begin{pmatrix} f_{\text{naive}} \\ f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} = \frac{a_1}{e_p e_q}\begin{pmatrix} 1 \\ \frac{a_2 b_p^2}{e_p^2} \\ \frac{a_3 b_q^2}{e_q^2} \\ \frac{b_p^4 a_2^2 + b_p^2 e_p^2}{e_p^4} \\ \frac{b_q^4 a_3^2 + b_q^2 e_q^2}{e_q^4} \end{pmatrix}\exp\left(\frac{a_2^2 b_p^2}{2e_p^2} + \frac{a_3^2 b_q^2}{2e_q^2}\right) \quad (27)$$

where e_p and e_q are defined as $e_x := \sqrt{1 - 2b_p^2 a_4}$ and $e_y := \sqrt{1 - 2b_q^2 a_5}$.

The functions $f_{\text{naive}}, f_1, f_2, f_3, f_4$ can be computed empirically for given bandwidths b_p and b_q and from the summation over the data points (p_i, q_i), with $i = 1, \cdots, n$, in Eq.(26). We can then solve for the likelihood-estimated copula density $f_{\text{LL}}(p, q)$ as $f_{\text{LL}}(p, q) = a_1(p, q)$

using the following identities which can be extracted from Eqs.(27):

$$a_2 = \frac{e_p^2}{b_p^2} \frac{f_1(p,q)}{f_{\text{naive}}(p,q)}, a_3 = \frac{e_q^2}{b_q^2} \frac{f_2(p,q)}{f_{\text{naive}}(p,q)}$$

$$e_p = b_p \left(\frac{f_3(p,q)}{f_{\text{naive}}(p,q)} - \left(\frac{f_1(p,q)}{f_{\text{naive}}(p,q)} \right)^2 \right)^{\frac{1}{2}} \quad (28)$$

$$e_q = b_q \left(\frac{f_4(p,q)}{f_{\text{naive}}(p,q)} - \left(\frac{f_2(p,q)}{f_{\text{naive}}(p,q)} \right)^2 \right)^{\frac{1}{2}},$$

which can be used to compute the local-likelihood copula density at each point (p,q) as

$$f_{\text{LL}}(p,q) = \quad (29)$$

$$f_{\text{naive}} e_p e_q \exp \left(-\frac{e_p^2}{2b_p^2} \left(\frac{f_1}{f_{\text{naive}}} \right)^2 - \frac{e_q^2}{2b_q^2} \left(\frac{f_2}{f_{\text{naive}}} \right)^2 \right).$$

The copula density function Eq.(29) can be computed at any point using Eqs.(27). This equation gives an analytic correction to the naive density estimate f_{naive} for the local-likelihood density f_{LL}. The only unknowns at this point are the kernel bandwidths b_s and b_r which will be discussed in the next section.

After computing the density in the (p,q) space, we can transform back to the probit dimensions (r,s) and then to the original (u,v) in the CDF domain $u,v \in [0,1]$

$$\begin{pmatrix} r \\ s \end{pmatrix} = \mathbf{W}^{-1} \begin{pmatrix} p \\ q \end{pmatrix}, \quad \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \Phi(r) \\ \Phi(s) \end{pmatrix}, \quad (30)$$

where $\mathbf{W}$ is the transformation matrix to the PCA coordinates. The transformation from (p,q) to (r,s) is an isometry, hence $f_{\text{LL}}(r(p,q), s(p,q)) = f_{\text{LL}}(p,q)$. The copula density is then computed using Eq.(19).

Selecting proper bandwidths is crucial to get well behaved and precise kernel density estimates specially on the borders. This is a sensitive issue which can drastically affect the local and asymptotic properties of the density estimation. The transformation of the data to probit coordinates and then to the principal components makes it natural to consider a diagonal bandwidth matrix as we did in the previous section with two diagonal components b_r and b_s as the only remaining parameters which should be estimated in Eq.(29).

There are two main approaches for estimating the bandwidths. In the first one, we consider a constant bandwidth for all the points on the plane while in the second one, we define the bandwidth according to the local distribution of the data, for example to be proportional to the distance of each point to its kth-nearest neighbor point. Since we want to take advantage of the analytical solution for the local-likelihood copula, we here use a fixed bandwidth.

As discussed in [26, 27], a good choice of the bandwidth should balance the integrated asymptotic squared bias and the variance of the considered estimator. We do this by minimizing the mean integrated squared error (MISE). However, popular data-driven selection strategies are based on cross-validation. The most popular instances are least-squares cross-validation [39] and biased cross-validation [40]. Here, the MISE takes the form

$$\text{MISE}[f_{\text{LL}}] = \iint_{\mathbb{R}^2} \mathbb{E} \left[(f_{\text{LL}}(p,q) - f_T(p,q))^2 \right] dpdq$$

$$\propto \iint_{\mathbb{R}^2} f_{\text{LL}}^2(p,q) dpdq - \frac{2}{n} \sum_{i=1}^{n} f_{\text{LL}}^{\{CV\}}(p_i, q_i),$$

$$(31)$$

where f_T is the true density of the data points (p_i, q_i) and f_{LL} is the local likelihood approximate of the density. The $\sum_{i=1}^{n} f_{\text{LL}}^{\{CV\}}$ term is the cross-validated sum of the copula density over the data points. For instance, a leave-one-out or k-fold procedure can be used to split the data into training and test subsets. The density at each test set can be estimated using the density function which is estimated using only the corresponding training set. By having a cross-validated copula density for each point, we can estimate the sum in the second term. The integral part is computed numerically using the equally spaced binning of the (p,q) space as it is shown in Fig.1. The possible effect of the number of bins on the mutual information estimation will be shown in the next sections.

The bandwidth parameters $\boldsymbol{b} = (b_p, b_q)$ can then be estimated numerically by minimizing the MISE$[f_{\text{LL}}]$

$$\boldsymbol{b} = \arg\min_{b_p, b_q} \left\{ \iint_{\mathbb{R}^2} f_{\text{LL}}^2(p,q) dpdq - \frac{2}{n} \sum_{i=1}^{n} f_{\text{LL}}^{\{CV\}}(p_i, q_i) \right\}.$$

$$(32)$$

In the results presented in this paper we used 5-fold cross-validation to estimate MISE. We did not observe any significant difference in results when using values of k as large as 20, for a dataset with $n = 1000$ samples.

3. Bandwidth selection

One possible simplification for the bandwidth selection is used in [27, 38] where it was shown that a *rule of thumb* way of defining the bandwidth can perform well. In this rule, the bandwidth will be proportional to the square root of the covariance matrix $\boldsymbol{b} \propto \Sigma^{1/2}$ where Σ is the empirical covariance matrix in the (p,q) coordinates (so it is diagonal). We can then use this approach and instead of optimizing Eq.(32) for two free parameters, we can solve the problem for one parameter α after defining the bandwidth as $\boldsymbol{b} = \alpha \Sigma^{1/2}$. We will refer to the density function obtained from the simplified one-parameter bandwidth as LL1 and the density function obtained from two-parameter bandwidth as LL2.

4. Copula density normalization

One important property of the copula density is that it has uniform marginals. It is important to ensure that the

estimated empirical copula density satisfies this property as well. This means that we should have

$$\int_0^1 c(u,x)dx = \int_0^1 c(x,v)dx = 1, \quad u,v \in [0,1]. \ (33)$$

Because of numerical imperfections and approximations of the kernel estimation, these constraints might be violated. In order to impose these constraints, we follow the iterative normalization suggested in Nagler et al. [38] by repeatedly dividing the copula density by its marginals

$$c(u,v) \to \frac{c(u,v)}{\int_0^1 c(u,x)dx \int_0^1 c(x,v)dx}. \quad (34)$$

A relatively small number of iterations ($\sim$1000) is sufficient to get copula densities with almost uniform marginals. Finally, in order to be sure that we get a proper density function, we normalize the resulting copula density with its integral over the two-dimensional domain $(u,v) \in [0,1]^2$. This numerical normalization assures that the resulting density satisfies the properties of a copula density.

Also note that numerical computation of the integrals required for the estimation of bandwidths in Eq.(32) as well as for the estimation of the density in Eq.(29) and in the normalization procedure of Eq.(34), we use a grid of the (p,q) domain as shown in Fig.2. In principle, it would be possible to use different grid sizes for bandwidth optimization and for density estimation. For example, it may be useful to use a coarser grid to estimate the bandwidth (since it will be more efficient in terms of computational time) and a finer grid size to estimate the density (to have a higher resolution density estimation) or in the sampling procedure. For the simulations presented in this paper, we used equal grids for bandwidth optimization and density estimation both to get a lower bound of the information estimation error and to simplify the procedure.

III. RESULTS

Our approach is to compare the NPC-based estimator with current information estimators using numerical simulations in both continuous and discrete domains. We focus on the performance of the NPC-based estimator in terms of optimizing its parameter selection, and evaluating its accuracy, sensitivity to sample size, and robustness to the form of the marginal distributions. We also focus on comparing the properties of NPC-based estimator to those of the best performing estimators among those currently available.

A. Continuous variables

We first consider the case of estimating information between continuous valued variables. These cases are rele-
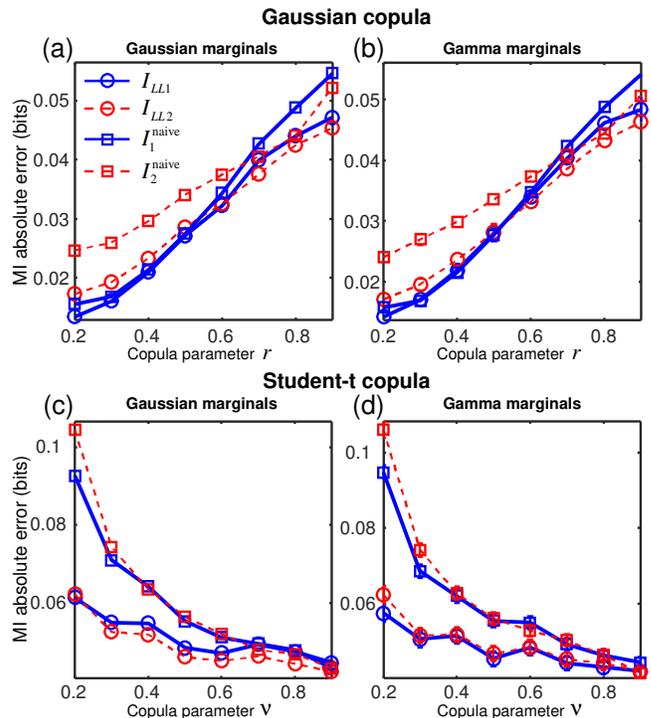
FIG. 3. Mutual information absolute error is computed for four different datasets using one-parameter bandwidth (I_{LL1} and I_1^{naive}) and two-parameter bandwidth (I_{LL2} and I_2^{naive}) non-parametric copula using both the naive and local likelihood estimators. Data are generated by means of the copula method with $k = 100$ grid size and $N = 1024$ sample size. The data in each panel are combinations of two similar marginal distributions and a parametric copula. a) Gaussian copula with parameter r, Gaussian marginals. b) Same Gaussian copula, gamma marginals. c) Student-t copula with parameters $r = 0$ and ν, Gaussian marginals. d) Same Student-t copula, gamma marginals.

vant for many important applications, ranging from analysis of gene networks [41] to the analysis of neuroimaging data such as electro- and magneto-encephalograms [11] and to the analysis of continuous valued dynamical systems [7, 42].

In the continuous domain, we tested the NPC-based information estimators in four different simulated conditions. We generated the datasets so that we had the ground-truth theoretical values of the mutual information for those probability distributions. We quantified estimation accuracy by computing the mutual information absolute error $E[|I_{\text{estimate}} - I_{\text{theory}}|]$, the normalized bias $E[I_{\text{estimate}} - I_{\text{theory}}]/I_{\text{theory}}$ and the normalized variance of the estimator $E[(I_{\text{estimate}} - E[I_{\text{estimate}}])^2]/I_{\text{theory}}$ over a number of simulations (1000 simulations for each condition). For each condition, we generated simulated data using a known parametric copula dependency structure and known marginal structures. For the dependency structure between variables, we considered two families of parametric copulas: the Gaussian copula family and the student-t copula family, each of which has closed-form

solutions for calculating the associated entropies (see Section II C). For the Gaussian copula, r was varied from 0.2 to 0.9. For the Student-t copula, r was set to 0 and ν was varied between 0.2 and 0.9, forming entirely nonlinear dependencies and zero linear correlation (see the copula in Fig.1).

Note that the mutual information is positively correlated with r and negatively correlated with ν. We combined copulas from each of these families with marginal distributions that were either Gaussian $\mathcal{N}(\mu = 0, \sigma = 1)$ or gamma-exponential $\Gamma(\alpha = 0.1, \beta = 10)$. The selected parameters for the gamma-exponential marginal distribution formed a sharp boundary peak at zero (similar to the gamma distribution shown in Fig.1), which is difficult to capture with methods that operate on the properties of the density function. We therefore generated bivariate distributions with selected marginals and a relationship structure specified by the selected copula (see Sklar's theorem 1). In each case, we simulated the data with the sampling approach explained in Sec.II B 1.

1. Optimization of the non-parametric copula

Given that the use of non-parametric copulas has been introduced only recently [26, 43, 44] and that they have not been used for information estimation before, we first investigated how to optimize the performance of various possible implementations of the non-parametric copula (Fig.3). We considered versions with a two-parameter bandwidth and a simpler version with a one-parameter bandwidth local-likelihood method (see Sec.II D 3). For all the simulated dataset conditions, we did not find a significant difference between the two- and one-parameter bandwidth versions of the NPC estimator in terms of information estimate accuracy (Fig.3). This result suggests that, in the (p, q) space, the covariance of the distribution was enough to capture the local variations in the density and hence the optimal shape of the kernel function. We also compared the absolute mutual information error obtained with the local-likelihood copula with that obtained with the naive copula. It has been already shown that the local-likelihood density copula describes better data with sharp variations, edges or other types of local nonuniformity [26]. In Fig.3 we tested whether these properties lead to a more accurate mutual information estimation. As expected, the naive estimation of the copula density was accurate in simple situations, such as the Gaussian copula. However, for the case of high non-linear correlations in a student-t copula (smaller ν values), the naive method failed to capture the sharp corners of the copula and had double the estimation error of the local-likelihood methods. We therefore chose to use the $LL1$ as the copula estimator for the comparison with other methods.

This version had the advantage of having fewer parameters for the bandwidth parameterization, which made the optimization of Eq.(32) faster and easier to converge,
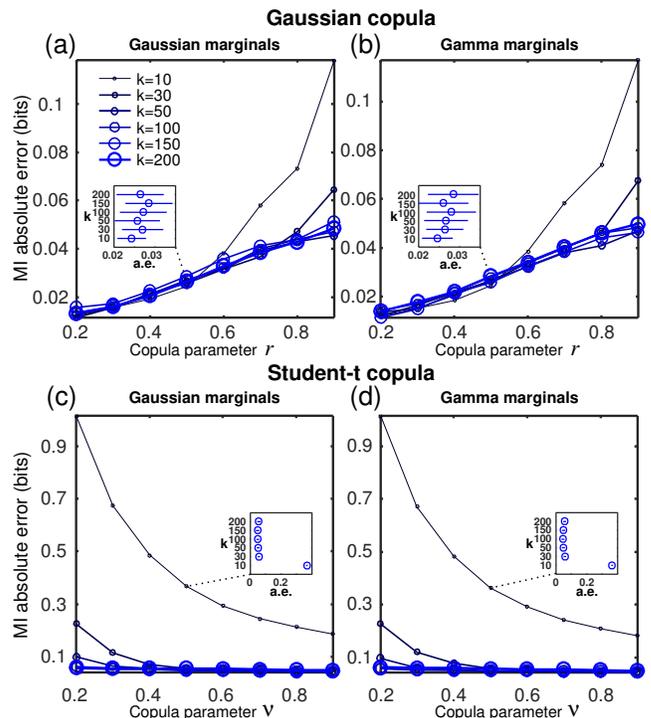
FIG. 4. Mutual information absolute error for a range of grid sizes is shown for the same data as of Fig.3 using the $LL1$ local-likelihood non-parametric copula method. Insets show the 95% confidence interval of the mean of MI absolute error for each k value after correction for multiple comparison for the $r, \nu = 0.5$ cases.

without much cost to the accuracy of the density function and the mutual information estimations. The only free parameter of the non-parametric copula that needs to be selected a priori is the number of grids k that are used to quantize the (p, q) space for estimation of the bandwidths in Eq.(32) and normalization of the copula density in Eq.(34). To test how this parameter affected the estimated mutual information, in Fig.4 we tested the NPC estimator, on the same simulated data used in the previous figure, varying k from 10 to 200 (in the previous figure a value of $k = 100$ was used). For $k \geq 50$, there was little improvement in the information error with increasing values of k, both in strongly correlated and less correlated copulas. We thus selected $k = 100$ for the remaining analysis. For smaller k's, e.g. $k = 10$, the resolution of the binning of the copula space was not sufficient to capture the sharp corners of the student-t copula, even though it performed well for the Gaussian copula (Fig.4). In the practical implementation of the above procedures, we found that, for strongly correlated data (e.g. large r values of the Gaussian copula or small ν values of the student-t copula), the MI absolute error decreased monotonically with k until reaching a constant value at larger k, and a small number of iterations was enough to optimize bandwidth. For weak correlation

cases, we still observed a decrease of the estimation error when increasing k, although in such cases the copula bandwidths were usually larger and so the bandwidth optimization needed more iterations for large k values. In the results presented in this paper, we used a bounded optimization function since the size of the bandwidth is bounded by the extension of the data in the (p, q) domain[1].

2. Comparing the NPC estimator in the continuous domain with existing established estimators

We next compared the NPC method with two other alternative established methods. First, we tested our non-parametric copula estimator against a parametric copula-based estimator based on the Gaussian copula (GC) whose parameters were estimated by maximum likelihood [11]. This estimator was selected for comparison because it is a popular method for estimation of

FIG. 5. Mutual information absolute error is shown for data similar to Fig.3 using different information estimation methods.

FIG. 6. The difference of mutual information estimated with normal and gamma marginal distributions is shown for LNC and NPC methods for (a) the Gaussian copula and (b) the student-t copula, with similar parameters as in Fig.3. In all the simulations we used $k = 100$ and $N = 1024$ samples.

information in continuous brain signals [11]. Second, we also compared our NPC estimator against the mutual information estimates obtained with the LNC method [15]. This comparison was chosen because, as we also confirmed in our experience on our simulated data, the LNC method is considered to be the best performing among those not based on copulas such as those based on nearest neighbors [12, 14, 15]. The results for all four simulation conditions and for a range of copula parameters are shown in Fig.5.

The GC gave the most accurate results in the case of data generated using a Gaussian copula Fig.5, as expected because in this case the parametric copula used for generating the data matched the one used for estimating information, but it gave the largest error in estimating the mutual information on data generated by the student-t copula, which lacked linear correlations in the data.

The LNC method worked well for both copula families when we used normal marginal distributions to generate the data but it was highly sensitive to the change of marginal distribution to gamma distribution[2]. The absolute mutual information error obtained using the LNC method was nearly an order of magnitude larger for the gamma function marginal distribution compared to the Gaussian marginals, for the same copula function. This result shows that the LNC method was strongly affected by the form of the marginal distribution, especially in the strongly correlated situations, e.g. large r for Gaussian copula and small ν for student-t copula. In contrast to the LNC and the GC methods, the NPC had both desirable properties expected by an ideal estimator.

First, it worked well for all the types of dependencies

---

[1] The MISE is a convex function which can be optimized easily and reliably. We used the MATLAB function `fminbnd` with maximum 500 number of iterations for the optimization. Using smaller number of iterations as low as 100 will have minimal effect on the results specially in the more correlated cases. The bandwidths are bounded to zero from below and to the rule-of-thumb bandwidth value used in [38] from the above.

[2] The numerical estimations from LNC are computed with $k = 5$ (k being the number of nearest neighbors) and the default value of α parameter, using the toolbox available online at https://github.com/BiuBiuBiLL/MIE.

used to generate the data, giving low absolute errors for both data generated with the Gaussian and the student-t copula. Second, further quantification of the difference in the error in mutual information estimates when using either Gaussian or Gamma marginal distributions (Fig.6) showed that the NPC-based estimator was not affected by the marginal distributions used to generate the data. The mutual information depends only on the copula, thus an ideal estimator should give equal results regardless of the marginal distribution. In sum, unlike previous methods the NPC-based estimator had the double advantage that it both functioned accurately for both types of copula families, including both linear and nonlinear dependency structures, and was insensitive to the marginal distributions. To further investigate the performance properties of the mutual information estimators, we computed the normalized bias and standard deviation of each of them, for the same data used in the above figures. The results (Fig.7) show that the better performance of the NPC estimator is largely due to a decrease in bias, but that the NPC estimator has also the additional desirable property of having in general less variance. Given that, in practical applications, data available for information estimation are often scarce, it is important that an estimator is accurate also when small datasets are available. We thus investigated in Figs.(8,9) how the performance of the NPC-based estimator varied with the sample size. We computed, for the four simulated data conditions and
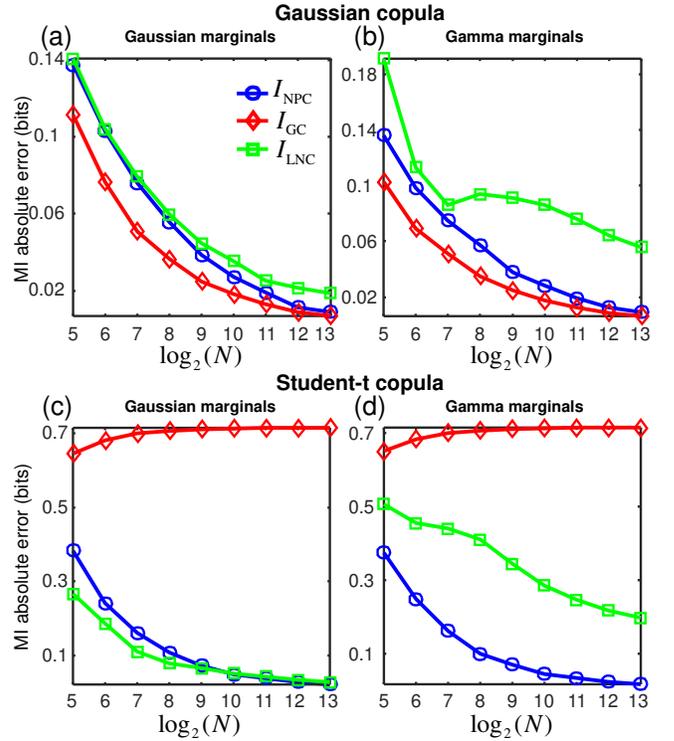
FIG. 8. Mutual information absolute error similar to Fig.3 but for different sample sizes (N), and the fixed values of $r = 0.5$ and $\nu = 0.5$.

across a range of sample sizes ($N = 2^5, \cdots, 2^{13}$), the mutual information absolute error (Fig.8) and the mutual information normalized bias (Fig.9). In these cases, we fixed the parameters of the copulas as $r = 0.5$ for the Gaussian copula and $\nu = 0.5$ for the student-t copula. The NPC method rapidly converged to a low error level with increasing sample size and had low error even at the smallest sample size. At most of the cases and sample sizes, the NPC method outperformed the LNC method, including for sample sizes as small as 64, for which there was an order of magnitude difference in the estimation error between the NPC and LNC methods for the simulated data with gamma function marginal distributions.

B. Discrete variables

We next considered the problem of estimating the mutual information between two random variables taking integer numerical variables. Having efficient information estimators in such cases is important for many applications. For example, in neuroscience experiments it is often important to estimate the information that the number of spikes emitted by neurons carry about sensory or behavioral variables taking integer values. Note that, any discrete set of discrete variables can in principle be one-to-one mapped to a set of integer variables, with similar

FIG. 7. Mutual information bias ratio is shown for similar data as in Fig.3 using different information estimation methods. The errorbars represent the standard deviation over the 1000 simulations of the data.
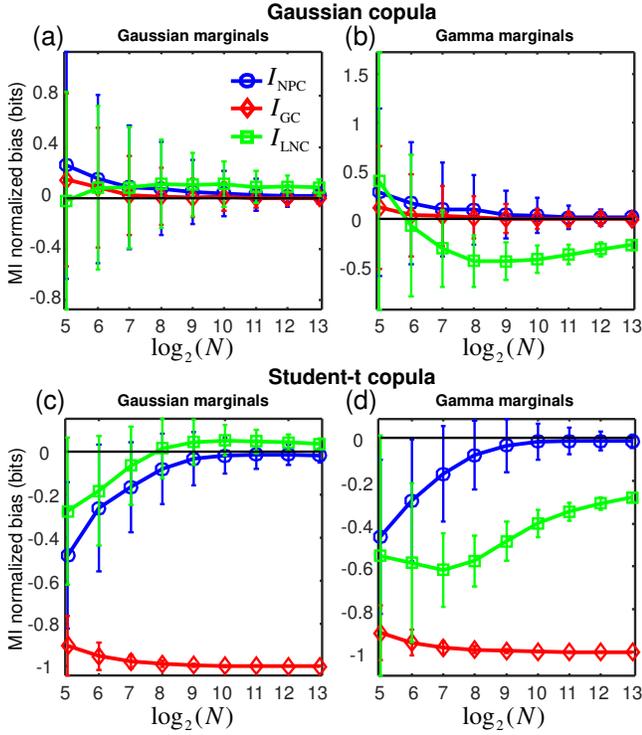
FIG. 9. Mutual information bias is plotted for data similar to Fig.3 but for different sample sizes (N) and the fixed values of $r = 0.5$ and $\nu = 0.5$. The errorbars represent standard deviation over data simulations.

probability mass function of the original discrete variables; this makes the current setting quite general.

The local-likelihood kernel method requires a continuous, smooth, and integrable copula density, which is not the case for integer variables. We therefore used a simple approach to transform discrete data into the continuous domain, without affecting the information content, by adding appropriate noise to the data. This approach provided a single framework for computing mutual information between continuous and integer variables and their mixtures.

1. Adapting the NPC estimator to discrete numerical variables

We first examined how to transfer integer variables into the continuous domain without affecting the information content. Consider a bivariate set of integer variables (n_X, n_Y). We can show that there exists proper noise variables ϵ_X and ϵ_Y independent from (n_X, n_Y) such that

$$I(n_X + \epsilon_X; n_Y + \epsilon_Y) = I(n_X; n_Y). \qquad (35)$$

One possible noise distribution satisfying 35 is a union of uniform distributions filling the gaps between consecutive integer variables. Consider $\{n_i\}$ as the sorted set of integer variables ($n_i > n_{i+1}$ for all $i = 1 \cdots N_{\max} - 1$) according to their indices. We then add the following uniform noise

$$\epsilon_i \sim \mathcal{U}_{[n_i, n_{i+1}]} \qquad (36)$$

to each integer n_i which transforms each integer variable n_i to their corresponding $\tilde{n}_i$ in real domain satisfying $\tilde{n}_i > \tilde{n}_{i+1}$ for all $i = 1 \cdots N_{\max} - 1$. We can then write the probability of the noised variable $\tilde{n}_i$ as

$$p(\tilde{n}_i) = p(n_i + \epsilon_i) = \sum_{n=1}^{N_{\max}} P(n)p(\epsilon_i = \tilde{n}_i - n) \qquad (37)$$

Since, based on the definition of the noise ϵ_i, we have $p(\epsilon_i = \tilde{n}_i - n) = 0$ for $n \neq n_i$ we will have

$$p(\tilde{n}_i) = P(n_i)p(\epsilon_i). \qquad (38)$$

Similarly, the joint density can be decomposed as the product of the mass function of the integer variables n_X and n_Y and the noise densities

$$p(\tilde{n}_X, \tilde{n}_Y) = P(n_X, n_Y)p(\epsilon_{n_X})p(\epsilon_{n_Y}). \qquad (39)$$

We then write the mutual information between the continuous variables $\tilde{n}_X$ and $\tilde{n}_Y$ as:

$$I(\tilde{n}_X; \tilde{n}_Y) = \iint_{\tilde{n}_X \tilde{n}_Y} p(\tilde{n}_X, \tilde{n}_Y) \log \frac{p(\tilde{n}_X, \tilde{n}_Y)}{p(\tilde{n}_X)p(\tilde{n}_Y)} d\tilde{n}_X d\tilde{n}_Y$$

$$= \sum_{n_X, n_Y} P(n_X, n_Y) \log \frac{P(n_X, n_Y)}{P(n_X)P(n_Y)} \iint_{\epsilon_{n_X} \epsilon_{n_Y}} p(\epsilon_{n_X})p(\epsilon_{n_Y}) d\epsilon_{n_X} d\epsilon_{n_Y} = I(n_X; n_Y) \qquad (40)$$

Which means that adding this noise and transforming the integer data to the real domain does not change the information between the variables. We can then use the variables $(\tilde{n}_X; \tilde{n}_Y)$ in the continuous domain together with the kernel copula to estimate their mutual informa-

tion. An example simulation of such continuation of integer bivariate data into the continuous domain is shown in Fig.10. We note that a similar approach for continuation of the discrete domain into mixed variables for density estimation has been proposed in [45], to which we refer
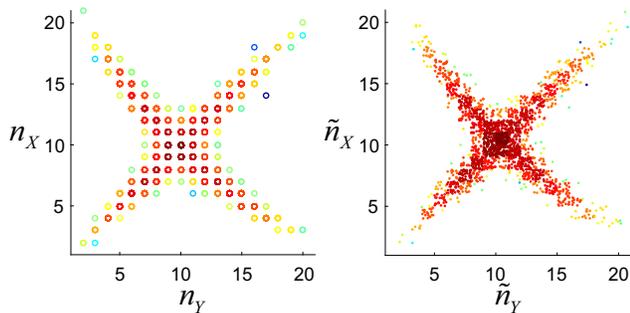
FIG. 10. A set of bivariate integer data points (left) become continuous (right) after adding variables with proper noise distributions to each point.

for further details.

2. Testing the performance of the discrete NPC estimator

To test the performance of the NPC method, we simulated data using Gaussian and student-t copulas with $r = 0.5$ and $\nu = 0.5$, respectively. Here, for the marginal distributions, we used Poisson distributions with a variable range of Poisson rates $\lambda = 20, \cdots, 70$ to see how changing the properties of the marginal distribution affects the mutual information estimation. Poisson distributions fit well many empirical data of relevance, such as the distribution of spike count of cortical neurons[46]. We added noise to the data using Eq.(36) and computed the corresponding copula and its entropy. We compared the NPC method with direct fitting using a Gaussian copula, because this comparison is useful to illustrate the specific advantages of a non-parametric copula. We also tested the NPC against the Pitman-Yor Mixture (PYM)[3] information estimation method [20]. We selected the PYM method for comparison because, as also confirmed by our experience on these simulated data, it has been shown [20, 47] to further improve the performance of previous pioneering Bayesian estimators [18, 19], and the latter compare favorably to other bias subtraction methods [19, 48].

We first focused on how to optimize the computation of the NPC estimator. As we did for the continuous case, we tested various values of k (the binning parameter), compared models across simulation conditions, and analyzed estimation errors and biases as a function of sample size. In the discrete cases, we used the method used in [24, 49] to compute the ground truth mutual information.

The NPC-based estimator had low and flat error across a wide range of k values as is shown in Fig.11, with similar levels of error for $k > 10$. Also, the performance of the NPC estimator was insensitive to the properties of the marginal distributions and had similar levels of error across all tested values of λ. Further, the NPC estimator performed similarly well on both the Gaussian copula and student-t copula datasets. These results indicate that the NPC-based estimator performed similarly on integer variables as it did on continuous data. They also show flat normalized bias over the change of the Poisson rates (Fig.11). We then compared the NPC to other approaches over a range Poisson rates. As shown in Fig.11, the NPC estimator had significant advantages. Direct fitting with the GC approach worked well on the data generated from the Gaussian copula, but performed poorly on the data simulated with the student-t copula, as expected. The PYM approach performed worse than the NPC estimator on both cases, especially on the Gaussian copula case. The PYM method showed a strong dependency on the form of the marginals and had an order of magnitude larger errors for the largest values of λ. The NPC-based method was the only approach that generalized well across values of the marginal distribu-
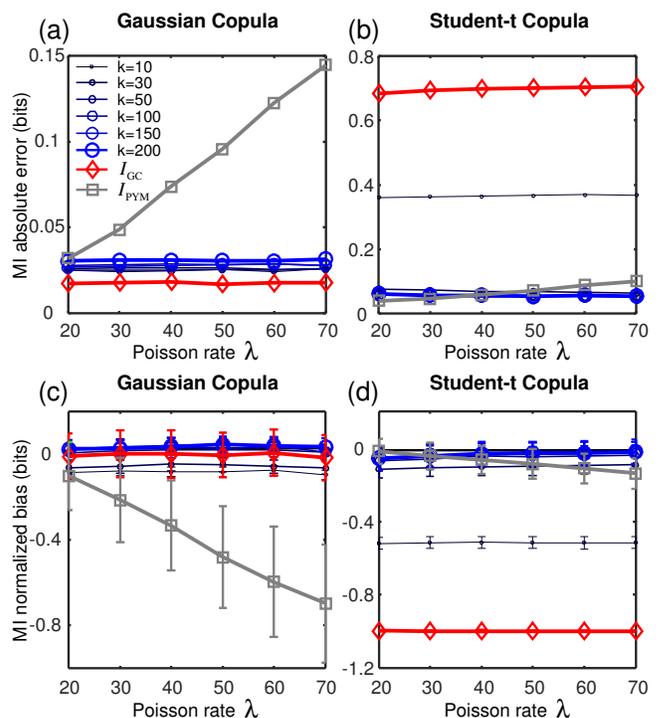
FIG. 11. The MI absolute error (top) and MI bias (bottom) are shown for the NPC model with a range of $k = 10, \cdots, 200$ number of bins, parametric Gaussian copula and PYM methods. To simulate the dataset, we used Poisson marginal distributions with $\lambda = 50$ and the Gaussian (left) and student-t copulas (right) and generate $N = 1024$ samples. The error bars of the MI bias plots are the standard deviation over 1000 data simulations.

---

[3] For all the comparisons with PYM, we used the default setting of the codes available online at https://github.com/pillowlab/CDMentropy. We computed the joint entropy H(X,Y) from the multiplicities of all the unique pairs of integers in the data.

tions and across the type of dependency structure in the data. Furthermore, the sample size dependency of different methods are shown in Fig.12. The performance of the NPC-based method, with a fixed Poisson rate at $\lambda = 50$, had a weaker dependence on the sample size than the PYM method and had significantly lower estimation absolute error than the PYM method for sample sizes $N < 2^{10}$. Furthermore, the NPC shows small and flat normalized biases and variances over the same range of sample sizes, contrary to the PYM estimator which shows large negative normalized biases and large variances for small samples sizes.

These results further demonstrate that the NPC method has an important property of mutual information estimators, namely that they estimate similar mutual information values for a fixed dependency structure over a wide range of marginal distributions and sample sizes. In order to quantify the degree of the dependency of each estimator to the parameters of the marginal distributions, after fixing the dependency structure, we computed the variability in the estimated mutual information, measured as the standard deviation of the information values estimated over the a range of Poisson rates λ (Fig.13). Across a wide range of sample sizes, the variability in the information estimate with varied λ was flat

FIG. 13. The standard deviation of MI is computed over a set of marginal distributions with firing rates $\lambda = 20, \cdots, 70$. The Poisson marginals are combined with (a) the Gaussian copula and (b) the student-t copula to generate the samples.

for NPC and GC methods and low relative to that of the PYM method. The PYM shows strong marginal distribution dependency especially for smaller sample sizes. The NPC-based estimator therefore appeared unaffected by large changes in sample size or marginal distributions, consistent with what was observed in the continuous case.

IV. CONCLUSIONS

Here we developed a new mutual information estimator based on non-parametric copulas. We have demonstrated that the method has several desirable features of a high-performance information estimator. First, the method is non-parametric, which means that assumptions about relationships in the data are not imposed. Second, the method is not sensitive to the distributions of individual variables (marginal distributions); rather, by virtue of its focus on the copula, it only takes into account the dependencies between variables. We were able to extend this advantage even to the discrete case, forming a single framework for the study of continuous, discrete, and mixed combinations of variables. Third, the NPC-based estimator worked well at low sample numbers, which has commonly been challenging for non-parametric approaches. We additionally demonstrated that this approach performed and generalized better than state-of-the-art mutual information estimators in many cases.

Many currently used mutual information estimators have made important progress in being able to estimate information accurately and from limited samples, also in cases when the underlying probability distributions do not necessarily fit traditional parametric families of probabilities. However, these existing non-parametric methods do not explicitly single out the copula as the only part of the joint distribution that is taken into consideration for mutual information estimates [12, 15, 20]. We showed that estimators such as the kNN-based estimators and the PYM estimator were sensitive to the properties of the marginal distributions and can thus lead to inac-

FIG. 12. The MI absolute error (top) and normalized bias (bottom) are shown for different sample sizes for the NPC method, parametric Gaussian copula and PYM methods for Poisson marginal distributions with $\lambda = 50$ and the Gaussian (left) and student-t copulas (right). The errorbars of the bottom panels are the normalized standard deviation computed over a set of data simulations.

curate information estimates. For example, even with the same dependency structure and thus identical mutual information, these methods could erroneously estimate different levels of mutual information due to differences in the properties of the marginal distributions. By making use of copulas, we isolated the part of the joint distribution that is relevant for the mutual information and avoided contamination of the information estimates from irregularities in the marginal distributions. Both in the continuous and integer domains, the NPC estimator provided a stable information estimate across values of the marginal distributions and across sample sizes, and it shows less performance degradation at small sample numbers. These results indicate that the NPC approach is able to identify the dependency structure, which is exactly the property critical for the mutual information between the variables of interest, and the method was correctly not affected by changes in other aspects of the data.

'To model the copula, we made use of non-parametric methods. Contrary to parametric methods, non-parametric methods do not make strong assumptions with respect to the shape of the distribution and the dependency structure of the data. Here we showed that the use of non-parametric approaches allowed for successful information estimation both in data generated from Gaussian dependencies with linear correlations and from student-t copulas with only nonlinear relationships. In particular, we used the probit transformation in conjunction with principal component analysis to transform the data samples in the copula domain into a space that lends itself well to kernel density estimators. We made progress in kernel-based methods for copula density estimation. In such methods, the selection of the appropriate kernel bandwidth is a crucial factor for achieving faithful density estimates [28]. We derived analytical solutions for the likelihood-estimated copula density with Gaussian kernels, making possible quick calculations of the density and the associated mean integrated square error. This allowed us to apply efficient methods for selecting the right kernel bandwidth. While other non-parametric copula methods such as splines smoothing [50] and Bernstein polynomials [51] have been put forward, a recent comparison suggests that probit-transformation-based methods tend to outperform alternative non-parametric estimators over a wide range of used cases [28] when combined with the local-likelihood density estimation [26].

Thus, the advantages of the NPC estimators result from being able to combine, for the first time into a single formalism, the best of two complementary approaches: the advantage of copula to focus specifically on the parts of the probability distribution that are important for information and the advantage of non-parametric methods in being able to adapt to a wide range of situations.

We tested the NPC-based estimator only in the bivariate case. The extension of copulas to multivariate cases has been developed through the vine-copula structures, showing that density estimators based on the vine-copula have better bias and variance scaling properties in terms of sampling size with respect to conventional non-copula based methods [26, 28, 33, 35, 52, 53]. Because the multivariate d-dimensional structures can be built using $d(d-1)/2$ bivariate copulas, the performance of the bivariate NPC suggests that similar trends are expected in higher dimensions. Investigation of the vine-copula as a mutual information estimator in higher dimensions is an important area of focus for future work.

We anticipate that, due to their adaptability to complex structures and their robustness to sample size, the NPC-based information estimator will be generally applicable in a wide range of fields and will advance and enhance the impact of information theory in many domains. In particular, application of information theory especially to biological problems in which data collection is constrained by insurmountable practical reasons and is both limited by the difficulty of estimating information accurately from limited samples [6, 54] and by the presence of complex nonlinearities [55].

As an important example, in neuroscience, hypotheses about how neurons encode information about certain behavioral variables (such as the parameters quantifying the nature of sensory stimuli or of behavioral choices) have thus far been limited to testing simple quantifications of the neural response, such as the number of action potentials fired in a given time window. Yet, evidence suggests that information may be encoded by more complex neural variables that include, for example, the pattern of firing of single neurons [56] or of neuronal populations [57], or the interactions between the timing of action potentials and of continuous neural response variables such as the power or phase of brain oscillations [58]. The nature of the interactions between such neural variables and potentially complex external variables of ethological interest (such as the value of sensory stimuli of naturalistic complexity) is largely unknown and cannot be safely described by parametric methods. Yet, the number of samples that can be collected is limited by factors such as the small length of time in which a subject can perform a cognitive task. Our NPC information estimator can be used to measure accurately relationships between such neural and behavioral variables, helping researchers to crack the code used by neurons to mediate complex behaviors. The Matlab package implementing the pairwise local-likelihood copula and the NPC information estimation algorithm is available at `github.com/houman1359/NPC_Info`.

[1] C. E. Shannon, Bell System Technical Journal **27**, 379 (1948).

[2] D. J. C. MacKay, *Information theory, inference and learning algorithms* (Cambridge University Press, 2003).

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (New York: Wiley, 2006).

[4] Y. K. Goh, H. M. Hasim, and C. G. Antonopoulos, PloS one **13**, e0192160 (2018).

[5] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, MA, USA, 1999).

[6] R. Quiroga and S. Panzeri, Nature Reviews Neuroscience **10**, 173 (2009).

[7] C. J. Cellucci, A. M. Albano, and P. E. Rapp, Physical Review E: Statistical, Nonlinear, and Soft Matter Physics **71**, 066208 (2005).

[8] R. L. Jenison and R. A. Reale, Neural Computation **16**, 665 (2004).

[9] R. S. Calsaverini and R. Vicente, EPL (Europhysics Letters) **88**, 68003 (2009).

[10] X. Zeng and T. Durrani, Electronics letters **47**, 493 (2011).

[11] R. A. Ince, B. L. Giordano, C. Kayser, G. A. Rousselet, J. Gross, and P. G. Schyns, Human Brain Mapping **38**, 1541.

[12] A. Kraskov, H. Stögbauer, and P. Grassberger, Physical Review E **69**, 066138 (2004).

[13] D. Pál, B. Póczos, and C. Szepesvári, in *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10 (Curran Associates Inc., USA, 2010) pp. 1849–1857.

[14] J. D. Victor, Physical Review E **66**, 051903 (2002).

[15] S. Gao, G. Ver Steeg, and A. Galstyan, in *Artificial Intelligence and Statistics* (2015) pp. 277–286.

[16] L. Paninski, Neural Computation **15**, 1191 (2003).

[17] S. Panzeri and A. Treves, Network: Computation in Neural Systems **7**, 87 (1996).

[18] I. Nemenman, F. Shafee, and W. Bialek, in *Advances in Neural Information Processing Systems 14*, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani (MIT; 1998, 2002) pp. 471–478.

[19] I. Nemenman, W. Bialek, and R. d. R. van Steveninck, Physical Review E **69**, 056111 (2004).

[20] E. Archer, I. M. Park, and J. W. Pillow, The Journal of Machine Learning Research **15**, 2833 (2014).

[21] M. Gerlach, F. Font-Clos, and E. G. Altmann, Physical Review X **6**, 021009 (2016).

[22] P. Wollstadt, K. K. Sellers, L. Rudelt, V. Priesemann, A. Hutt, F. Fröhlich, and M. Wibral, PLoS Computational Biology **13**, e1005511 (2017).

[23] A. Sklar, Publications de l'Institut de Statistique de L'Université de Paris **8**, 229 (1959).

[24] A. Onken and S. Panzeri, in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (2016) pp. 1325–1333.

[25] P. Jaworski, F. Durante, and W. K. Härdle, Lecture Notes in Statistics, Proceedings **213** (2013).

[26] G. Geenens, Journal of the American Statistical Association **109**, 346 (2014).

[27] G. Geenens, A. Charpentier, D. Paindaveine, *et al.*, Bernoulli **23**, 1848 (2017).

[28] T. Nagler, C. Schellhase, and C. Czado, Dependence Modeling **5**, 99 (2017).

[29] R. B. Nelsen, *An Introduction to Copulas*, 2nd ed. (Springer, New York, 2006).

[30] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. (New York: Springer, 2004).

[31] M. Rosenblatt, The Annals of Mathematical Statistics **23**, 470 (1952).

[32] L. Devroye, in *Proceedings of the 18th conference on Winter simulation* (ACM, 1986) pp. 260–265.

[33] H. Joe, *Dependence modeling with copulas*, Monographs on Statistics and Applied Probability No. 134 (CRC Press, Boca Raton, FL, 2015).

[34] H. Joe and J. J. Xu, *The estimation method of inference functions for margins for multivariate models*, Tech. Rep. 166 (Department of Statistics, University of British Colombia, 1996).

[35] K. Aas, C. Czado, A. Frigessi, and H. Bakken, Insurance: Mathematics and Economics **44**, 182 (2009).

[36] C. R. Loader *et al.*, The Annals of Statistics **24**, 1602 (1996).

[37] N. L. Hjort and M. C. Jones, The Annals of Statistics **24**, 1619 (1996).

[38] T. Nagler, arXiv preprint arXiv:1603.04229 (2016).

[39] M. Rudemo, Scandinavian Journal of Statistics , 65 (1982).

[40] D. W. Scott and G. R. Terrell, Journal of the American Statistical Association **82**, 1131 (1987).

[41] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, in *BMC bioinformatics*, Vol. 7 (BioMed Central, 2006) p. S7.

[42] M. Paluš and A. Stefanovska, Physical Review E **67**, 055201 (2003).

[43] S. X. Chen and T.-M. Huang, Canadian Journal of Statistics **35**, 265 (2007).

[44] J. S. Racine, Empirical Economics **48**, 37 (2015).

[45] T. Nagler, Statistics & Probability Letters **137**, 326 (2018).

[46] A. Amarasingham, T.-L. Chen, S. Geman, M. T. Harrison, and D. L. Sheinberg, Journal of Neuroscience **26**, 801 (2006).

[47] E. Archer, I. M. Park, and J. W. Pillow, Entropy **15**, 1738 (2013).

[48] S. Panzeri, R. Senatore, M. A. Montemurro, and R. S. Petersen, Journal of Neurophysiology **98**, 1064 (2007).

[49] A. Onken, S. Grünewälder, M. Munk, and K. Obermayer, in *Advances in Neural Information Processing Systems 21*, edited by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (2008) pp. 1233–1240.

[50] G. Kauermann, C. Schellhase, and D. Ruppert, Scandinavian Journal of Statistics **40**, 685 (2013).

[51] P. Janssen, J. Swanepoel, and N. Veraverbeke, Journal

of Multivariate Analysis **124**, 480 (2014).

[52] E. F. Acar, C. Genest, and J. NešLehová, Journal of Multivariate Analysis **110**, 74 (2012).

[53] T. Nagler and C. Czado, Journal of Multivariate Analysis **151**, 69 (2016).

[54] E. N. Brown, R. E. Kass, and P. P. Mitra, Nat Neurosci **7**, 456 (2004).

[55] F. Franke, M. Fiscella, M. Sevelev, B. Roska, A. Hierle-

mann, and R. A. da Silveira, Neuron **89**, 409 (2016).

[56] Y. Zuo, H. Safaai, G. Notaro, A. Mazzoni, S. Panzeri, and M. E. Diamond, Current Biology **25**, 357 (2015).

[57] C. A. Runyan, E. Piasini, S. Panzeri, and C. D. Harvey, Nature **548**, 92 (2017).

[58] C. Kayser, M. A. Montemurro, N. K. Logothetis, and S. Panzeri, Neuron **61**, 597 (2009).