Diffeomorphic Information Neural Estimation

Bao Duong, Thin Nguyen

Applied Artificial Intelligence Institute, Deakin University, Australia
{duongng,thin.nguyen}@deakin.edu.au

Abstract

Mutual Information (MI) and Conditional Mutual Information (CMI) are multi-purpose tools from information theory that are able to naturally measure the statistical dependencies between random variables, thus they are usually of central interest in several statistical and machine learning tasks, such as conditional independence testing and representation learning. However, estimating CMI, or even MI, is infamously challenging due the intractable formulation. In this study, we introduce **DINE** (**D**iffeomorphic **I**nformation **N**eural **E**stimator)–a novel approach for estimating CMI of continuous random variables, inspired by the invariance of CMI over diffeomorphic maps. We show that the variables of interest can be replaced with appropriate surrogates that follow simpler distributions, allowing the CMI to be efficiently evaluated via analytical solutions. Additionally, we demonstrate the quality of the proposed estimator in comparison with state-of-the-arts in three important tasks, including estimating MI, CMI, as well as its application in conditional independence testing. The empirical evaluations show that **DINE** consistently outperforms competitors in all tasks and is able to adapt very well to complex and high-dimensional relationships.

Introduction

Mutual Information (MI) and Conditional Mutual Information (CMI) are pivotal dependence measures between random variables for general non-linear relationships. In statistics and machine learning, they have been employed in a broad variety of problems, such as conditional independence testing (Runge 2018; Mukherjee, Asnani, and Kannan 2020), unsupervised representation learning (Chen et al. 2016), search engine (Magerman and Marcus 1990), and feature selection (Peng, Long, and Ding 2005).

The MI of two random variables X and Y measures the expected point-wise information, where the expectation is taken over the joint distribution P_{XY}. Due to the expectation, estimating mutual information for continuous variables remains notoriously difficult. Even if one possesses the specification of the joint distribution, i.e., a closed-form of the density, which is most of the time unknown in practice, the expectation may still be intractable. Consequently, exact MI estimation is only possible for discrete random variables. Historically, MI has been estimated by non-paramtric approachs (Kwak and Choi 2002; Paninski 2003; Kraskov, Stögbauer, and Grassberger 2004), which are however not widely applicable due to their unfriendliness with sample size or dimensionality. Recently, variational approaches have been proposed to estimate the lower bound of MI (Belghazi et al. 2018; Oord, Li, and Vinyals 2018). However, a critical limitation of MI lower bound estimators has been studied by (McAllester and Stratos 2020), who show that any distribution-free high-confidence lower bound estimation of mutual information is limited above by $\mathrm{O}\left(\ln n\right)$ where n is the sample size. More recent approaches includes hashing (Noshad, Zeng, and Hero 2019), classifier-based estimator (Mukherjee, Asnani, and Kannan 2020), and inductive maximum-entropy copula approach (Samo 2021).

While estimating MI is hard, estimating CMI is of magnitudes harder due to the presence of the conditioning set. Therefore, CMI estimation methods have seen slower developments than its MI counterparts. Recent developments for CMI estimation include (Runge 2018; Molavipour, Bassi, and Skoglund 2021; Mukherjee, Asnani, and Kannan 2020).

Present work. In this paper, we propose **DINE**[1] (**D**iffeomorphic **I**nformation **N**eural **E**stimator)–a unifying framework that closes the gap between the CMI and MI estimation problems. The approach is advantageous compared with novel variational methods in the way that it can estimate the exact information measure, instead of a lower-bound. Specifically, we harness the observation that CMI is invariant over conditional diffeomorphisms, i.e., differentiable and invertible maps with differentiable inverse parametrized by the conditioning variable.

As a direct consequence, first, we can now build a well-designed conditional diffeomorphic transformation that breaks the statistical dependence between the conditioning variable with the transformed variables, but keeps the information measure unchanged, reducing the CMI to an equivalent MI. Second, the approach offers a complete control over the distribution form of the newly induced MI estimation problem, thus we can easily restrict it to an amenable class of simple distributions with well-established properties and estimate the resultant MI via available analytic forms. Being

[1]Source code and relevant data sets are available at https://github.com/baosws/DINE

aided by the powerful expressivity of neural networks and normalizing flows (Papamakarios et al. 2021), we can define a rich family of diffeomorphic transformations that can handle a wide range of non-linear relationships, but are still efficient in sample size and dimensionality.

Our numerical experiments show that the proposed **DINE** estimator can consistently outperforms the state-of-the-arts in both the MI and CMI estimation tasks. We also apply **DINE** to test for conditional independence (CI)–an important statistical problem where the presence of the conditioning variable is a major obstacle, in which the empirical results indicate that the distinctively accurate CMI estimation of **DINE** allows for a high-accuracy test.

Contributions. The key contributions of our study are summarized as follows:

- We present a reduction of any CMI estimation problem to an equivalent MI estimation problem with the unchanged information measure, which overcomes the central difficulty in CMI estimation compared with MI estimation.
- We introduce **DINE**, a CMI estimator that is flexible, efficient, and trainable via gradient-based optimizers. We also provide some theoretical properties of the method.
- We demonstrate the accuracy of **DINE** in estimating both MI and CMI in comparisons with state-of-the-arts under varying sample sufficiencies, dimensionalities, and non-linear relationships.
- As a follow-up application of CMI estimation, we also use **DINE** to test for conditional independence (CI)–an important statistical problem with a central role in causality, and show that the test performs really well, as well as being able to surpass state-of-the-art baselines by large margins.

Background

In this Section we formalize the CMI estimation problem and explain the characterization of CMI that motivated our method.

Regarding notational interpretations, we use capitalized letters X, Y, etc., for random variables/vectors, with lowercase letters x, y, etc., being their respective realizations; the distribution is denoted by $P_{\cdot}(\cdot)$ with the respective density $p_{\cdot}(\cdot)$.

Conditional Mutual Information

The Conditional Mutual Information between continuous random variables X and Y given Z (with respective compact support sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$) is defined as

$$I(X,Y|Z) = \int_{\mathcal{Z}} \int_{\mathcal{Y}} \int_{\mathcal{X}} p(x,y,z) \ln \frac{p(x,y|z)}{p(x|z)\,p(y|z)} dxdydz \tag{1}$$

$$= \mathbb{E}_{p(x,y,z)} \left[\ln \frac{p(x,y|z)}{p(x|z)\,p(y|z)} \right] \tag{2}$$

where we have assumed that the underlying distributions admit the corresponding densities $p(\cdot)$.

Having CMI defined, our technical research question is to estimate $I(X,Y|Z)$ using the empirical distribution $P^{(n)}_{XYZ}$ of n i.i.d. samples, without having access to the true distribution P_{XYZ}.

Conditional Mutual Information Re-parametrization

Let us first recall that a *diffeomorphism* is defined as

Definition 1. (Diffeomorphism). A map $\tau(\cdot) : \mathcal{X} \to \mathcal{X}'$ is called a diffeomorphism if it is *differentiable* and *invertible*, and its inverse is also *differentiable*.

This kind of transformation is of great interest because it exhibits an important invariance property of MI, which was established by (Kraskov, Stögbauer, and Grassberger 2004):

Lemma 1. *(MI Re-parametrization, (Kraskov, Stögbauer, and Grassberger 2004)). Let $\tau_X : \mathcal{X} \to \mathcal{X}'$ and $\tau_Y : \mathcal{Y} \to \mathcal{Y}'$ be two diffeomorphisms where $x' = \tau_X(x)$ and $y' = \tau_Y(y)$, then we have:*

$$I(X,Y) = I(X',Y') \tag{3}$$

Proof. See the Supplementary Material (Duong and Nguyen 2022b). □

Inspired by this attractive property, CMI can be shown to be also invariant via any *conditional diffeomorphism*, which we define as

Definition 2. (Conditional Diffeomorphism). A differentiable map $\tau(\cdot;\cdot) : \mathcal{X} \times \mathcal{Z} \to \mathcal{X}'$ is called a conditional diffeomorphism if $\tau(\cdot; z) : \mathcal{X} \to \mathcal{X}'$ is a diffeomorphism for any $z \in \mathcal{Z}$.

The following Lemma states that it is possible to reparametrize CMI via some conditional diffeomorphisms:

Lemma 2. *(CMI Re-parametrization). Let $\tau_X : \mathcal{X} \times \mathcal{Z} \to \mathcal{X}'$ and $\tau_Y : \mathcal{Y} \times \mathcal{Z} \to \mathcal{Y}'$ be two conditional diffeomorphisms such that $P_{X'Y'|Z} = P_{X'Y'}$, where $x' = \tau_X(x; z)$ and $y' = \tau_Y(y; z)$, then the following holds:*

$$I(X,Y|Z) = I(X',Y') \tag{4}$$

Proof. See the Supplementary Material (Duong and Nguyen 2022b). □

The Diffeomorphic Information Neural Estimator (DINE)

Our framework can be described using two main components, namely the CMI *approximator* and the CMI *estimator*. While the approximator concerns the hypothesis class of models that are used to approximate the CMI given the access to the true data distribution, the CMI estimator defines how to estimate the CMI using models in the said approximator class, but with only a finite sample size.

CMI Approximation

We start by giving the general CMI approximator based on densities (as a direct solution to Eqn. (2)):

Definition 3. (Density-based CMI approximator). Given a family of density approximators with parameters $\theta \in \Theta$. The density-based CMI approximator $I_\Theta(X, Y|Z)$ is defined as

$$I_\Theta(X, Y|Z) = \mathbb{E}_{p(x,y,z)}\left[\ln \frac{p_{\theta^*}(x, y|z)}{p_{\theta^*}(x|z)\, p_{\theta^*}(y|z)}\right] \tag{5}$$

where the parameter $\theta^* = (\theta^*_X, \theta^*_Y, \theta^*_{XY}) \in \Theta$ are Maximum Likelihood Estimators (MLE) of the true densities $p(x, y|z)$, $p(x|z)$, and $p(y|z)$:

$$\theta^*_X = \arg\max_{\theta_X} \mathbb{E}_{p(x,z)}\left[\ln p_\theta(x|z)\right] \tag{6}$$

$$\theta^*_Y = \arg\max_{\theta_Y} \mathbb{E}_{p(y,z)}\left[\ln p_\theta(y|z)\right] \tag{7}$$

$$\theta^*_{XY} = \arg\max_{\theta_{XY}} \mathbb{E}_{p(x,y,z)}\left[\ln p_\theta(x, y|z)\right] \tag{8}$$

The innovation of **DINE** is fueled by the invariance property of CMI over diffeomorphic transformations as stated in Eqn. (4). To realize this end, the recently emerging Normalizing Flows (NF) technique offers us the exact tool we need to exploit the benefits we have just gained from the CMI reparametrization.

Simply put, NF offers a general framework to model probability distributions (in this case $P_{XY|Z}$) by expressing it in terms of a simple "base" distribution (here $P_{X'Y'}$) and a series of bijective transformations (the diffeomorphisms in our method). For more technical details regarding NFs, see (Kobyzev, Prince, and Brubaker 2020; Papamakarios et al. 2021).

Based on this, our approach involves the design of a class of conditional normalizing flows (in contrast with the unconditional normalizing flows that are not parametrized by Z), referred to as the *Diffeomorphic Information Neural Approximator (DINA),* and formalized as

Definition 4. (Diffeomorphic Information Neural Approximator (DINA)). A DINA $\mathcal{D}_\Theta$ is a density-based CMI approximator characterized by the following elements:

- A compact parameter domain Θ.
- A family of base distributions $\{P_\theta(X', Y')\}_{\theta\in\Theta}$.
- A family of conditional normalizing flows $\{\tau_\theta(\cdot;\cdot) : \mathcal{X} \times \mathcal{Z} \to \mathcal{X}'\}_{\theta\in\Theta}$.

Then, the approximation is defined as

$$I_\Theta(X, Y|Z) = I_\Theta(X', Y') \tag{9}$$

with $x' = \tau_{\theta^*_X}(x; z)$ and $y' = \tau_{\theta^*_Y}(y; z)$.

As mentioned, MI estimation from finite data is still difficult if the underlying distribution function is unknown or the expectation is intractable. Fortunately, the use of normalizing flows allows us to have a complete control over the distribution of the surrogate variables X' and Y'.

However, with an arbitrary base distribution, the finite sample estimation of the CMI as in Eqn. (5) still involves averaging the log-density terms, which results in a possibly very large estimation variance. Therefore, to reduce the estimation variance, we look for distributions with simple closed-form expression of the MI. Towards this end, the Gaussian distribution is an excellent choice thanks to its well-studied information-theoretic properties, especially the availability of a closed-form MI that we can make use of. In more details, when the base distribution $P_\theta(X', Y')$ is jointly Gaussian, we approximate the CMI as follows:

Definition 5. (DINA-Gaussian). If $P_\theta(X', Y')$ is multivariate Gaussian, then the DINA approximator with Gaussian base is defined as

$$I^{\mathcal{N}}_\Theta(X', Y') = \frac{1}{2}\ln \frac{\det \Sigma_{p(x,z)}(X')\det \Sigma_{p(y,z)}(Y')}{\det \Sigma_{p(x,y,z)}(X'Y')} \tag{10}$$

where $\Sigma_{p(x,z)}(X')$ is the covariance matrix of X' evaluated on the true distribution P_{XZ}, and so on.

For the rest of the main Sections we will assume that $P_\theta(X', Y')$ is jointly multivariate Gaussian with standard Gaussian marginals. That being said, the framework is still flexible to adapt to arbitrary base distributions that are independent of Z, as long as it is efficient to evaluate the marginal MI between X' and Y'.

We describe in more details the architecture of the normalizing flows employed for our framework in the next Section.

Learning Conditional Diffeomorphisms

Among a diversely developed literature of normalizing flows (Kobyzev, Prince, and Brubaker 2020; Papamakarios et al. 2021), *autoregressive flows* remain one of the earliest and most widely adopted. The most attractive characteristic of autoregressive flows is their intrinsic expressiveness. More concretely, autoregressive flows are universal approximators of densities (Papamakarios et al. 2021), meaning they can approximate any probability density to an arbitrary accuracy.

Suppose X and U are d-dimensional real-valued random vectors, where we wish to model the true $p(x)$ with respect to the base $p_\theta(u)$. Autoregressive flows transform each dimension i of x using the information of the dimensions $1..i-1$ of itself, hence the name "auto"–regressive:

$$u_i = \tau_\theta(x_i; h_i), \text{ where } h_i = c_i(x_{<i}) \tag{11}$$

Here the diffeomorphism τ is referred to as the *transformer*, and the function c_i is called the *conditioner*, which encodes the information from the dimensions $< i$ of x and defines parameters for the transformer.

Since u_i only depends on $x_{\leq i}$, the Jacobian matrix J_τ of partial derivatives is lower triangular, so the modeled log density is simply inferred using the change of variables rule as

$$\ln p_\theta(x) = \ln p_\theta(u) + \sum_{i=1}^{d} \ln\left|\frac{\partial u_i}{\partial x_i}(u)\right| \tag{12}$$

Furthermore, fitting $p_\theta(x)$ to $p(x)$ involves maximizing the expected likelihood with respect to the parameter θ:

$$\theta^* = \arg\max_\theta \mathbb{E}_{p(x)}[\ln p_\theta(x)] \tag{13}$$

where the expectation can be estimated with the sample mean of empirical data.

Going back to our problem, for example, to build the conditional diffeomorphism $x' = \tau_\theta(x; z)$ with a standard Gaussian distribution $P_{X'}$, we design the base distribution, transformer, and conditioner as follows:

Base distribution. First we choose $P_{X'}$ to be a d-dimensional standard uniform distribution $\mathcal{U}(0,1)^d$, since its density is constant everywhere, so no computation is required for $\ln p_\theta(x')$ nor its derivatives.

Next, to transform $P_{X'}$ to the desired standard Gaussian $\mathcal{N}(\mathbf{0}; \mathbf{1}_{d_X})$, we simply apply the inverse of the cumulative distribution function (CDF) of the standard Gaussian (Φ^{-1}) to x' in an element-wise fashion, where $\Phi(u) = \int_{-\infty}^{u} \mathcal{N}(t; 0, 1)\, dt$.

Transformer. With the base distribution being a standard uniform distribution, a natural choice for the transformer is the CDF of some densities, which is uniformly distributed for any strictly positive density.

To make the transformation more expressive, we compose the transformer as a weighted combination of different CDFs parametrized by the conditioner. More particularly, the transformer for the i-th dimension is given by

$$\tau(x_i; h_i) = \sum_{j=1}^{k} w_{ij}(h_i)\, \Phi\left(x_i; \mu_{ij}(h_i), \sigma_{ij}^2(h_i)\right) \tag{14}$$

where we have used a mixture of k CDF components, with the j-th component being a Gaussian CDF with mean $\mu_{ij}(h_i)$, variance $\sigma_{ij}^2(h_i)$, and positive weight $w_{ij}(h_i)$ such that $\sum_{j=1}^{k} w_{ij}(\cdot) = 1$.

This transformer is also a universal approximator for CDFs because its derivative is essentially a Gaussian mixture model (GMM), a canonical universal approximator of densities (Goodfellow, Bengio, and Courville 2016). Put simply, with a sufficient number of Gaussian components, $\tau(x_i; h_i)$ can express any strictly monotonic $\mathbb{R} \to (0,1)$ map (hence invertible) with an arbitrary accuracy, which is followed by the broad expressiveness of the transformer.

Conditioner. Since we would like to model the conditional diffeomorphism $\tau(x; z)$, the conditioner function c_i must encode both $x_{<i}$ and z, so instead of $h_i = c_i(x_{<i})$, now we let

$$h_i = c_i(x_{<i}, z) \tag{15}$$

In contrary to the transformer, the conditioner needs not to be invertible, so we can freely model it using any family of functions with inputs $x_{<i}$ and z.

Neural Network Parametrization. To maximize the expressivity power of autoregressive flows explained earlier, we parametrize all functional components in Eqn. (14) and Eqn. (15) with neural networks for each dimension, hence the term "Neural" in **DINE**.

More specifically, let d_H be the dimension of H, we model $w_i : \mathbb{R}^{d_H} \to (0,1)^k$ as a Multiple Layer Perceptron (MLP) with Softmax outputs, while $c_i : \mathbb{R}^{i-1+d_Z} \to \mathbb{R}^{d_H}$, $\mu_i : \mathbb{R}^{d_H} \to \mathbb{R}^k$ and $\ln \sigma_i^2 : \mathbb{R}^{d_H} \to \mathbb{R}^k$ are real-valued MLPs for all $i = 1..d_X$.

Since the Jacobian matrix J_τ are now differentiable with respect to the parameter θ, any gradient-based continuous optimization framework can be applied to learn θ^*.

CMI Estimation

Having the ingredients above ready, we can now proceed to define the **DINE** estimator for CMI:

Definition 6. (Diffeomorphic Information Neural Estimator (**DINE**)). Consider a DINA approximator $\mathcal{D}_\Theta$ with parameters in a compact domain Θ. **DINE** is defined as

$$I_n(X, Y|Z) = \mathbb{E}_{p^{(n)}(x,y,z)}\left[\ln \frac{p^{(n)}(x', y')}{p^{(n)}(x')\, p^{(n)}(y')}\right] \tag{16}$$

with $x' = \tau_{\theta_X^*}(x; z)$ and $y' = \tau_{\theta_Y^*}(y; z)$.

where the parameter $\theta^* = (\theta_X^*, \theta_Y^*, \theta_{XY}^*) \in \Theta$ are Maximum Likelihood Estimators (MLE) of the empirical densities $p^{(n)}(x, y|z)$, $p^{(n)}(x|z)$, and $p^{(n)}(y|z)$:

$$\theta_X^* = \arg\max_{\theta_X} \mathbb{E}_{p^{(n)}(x,z)}[\ln p_\theta(x|z)] \tag{17}$$

$$\theta_Y^* = \arg\max_{\theta_Y} \mathbb{E}_{p^{(n)}(y,z)}[\ln p_\theta(y|z)] \tag{18}$$

$$\theta_{XY}^* = \arg\max_{\theta_{XY}} \mathbb{E}_{p^{(n)}(x,y,z)}[\ln p_\theta(x, y|z)] \tag{19}$$

Note that θ_X and θ_Y here denote the parameters of the normalizing flows and the marginal densities $p_\theta(x')$ and $p_\theta(y')$, while θ_{XY} is the parameter of the joint density $p_\theta(x', y')$, which is constrained to have marginals $p_{\theta_X^*}(x')$ and $p_{\theta_Y^*}(y')$.

Specifically, under the case of multivariate Gaussian base, **DINE** can be written as

Definition 7. (**DINE-Gaussian**). If $P_\theta(X', Y')$ is multivariate Gaussian, then the **DINE** estimator with Gaussian base is defined as

$$I_n^{\mathcal{N}}(X, Y|Z) = \frac{1}{2} \ln \frac{\det \Sigma_n(X') \det \Sigma_n(Y')}{\det \Sigma_n(X'Y')} \tag{20}$$

We emphasize that in contrary with the general DINE estimator defined in Eqn. (16), the **DINE-Gaussian** estimator does not require explicit density evaluations but instead leverages the log determinants of sample covariance matrices that offers a lower estimation variance, and thus makes it the preferable estimator in practice.

Theoretical Properties

In this Section we state some important theoretical results regarding the **DINE** estimator, including estimation variance, consistency, and sample complexity. Proof sketches for these results are provided in the Supplementary Material (Duong and Nguyen 2022b).

Variance

Lemma 3. *(Variance of **DINE-Gaussian**). The asymptotic variance of the DINE-Gaussian estimator is given by*

$$\mathrm{Var}\left[I_n^{\mathcal{N}}\right] \xrightarrow{L} \mathrm{O}\left(\frac{d}{n}\right), \text{ as } n \to \infty \tag{21}$$

with d being the dimensionality.

Consistency

The quality of **DINE** depends on the choice of (i) a family of normalizing flows and (ii) n i.i.d. samples from the true distribution P_{XYZ}. The following Lemma states that, given a sufficiently expressive DINA, we can approximate the information measure to arbitrary accuracy.

Lemma 4. *(Approximability of DINA). For any $\epsilon > 0$, there exists a DINA $\mathcal{D}_\Theta$ with some compact domain $\Theta \subset \mathbb{R}^c$ such that*

$$|I(X,Y|Z) - I_\Theta(X,Y|Z)| \le \epsilon, \text{ almost surely} \tag{22}$$

The next Lemma declares that the estimator almost surely converges to the approximator as the sample size approaches infinity.

Lemma 5. *(Estimability of **DINE**). For any $\epsilon > 0$, given a DINA $\mathcal{D}_\Theta$ with parameters in some compact domain $\Theta \subset \mathbb{R}^c$, there exists a $N \in \mathbb{N}$ such that*

$$\forall n \ge N,\ |I_n(X,Y|Z) - I_\Theta(X,Y|Z)| \le \epsilon, \text{ almost surely} \tag{23}$$

Finally, the two Lemmas above together prove the consistency of **DINE**:

Theorem 1. *(Consistency of **DINE**). **DINE** is consistent whenever DINA is sufficiently expressive.*

Sample Complexity

We make the following assumptions: the log-densities are bounded in $[-M, M]$ and L-Lipschitz continuous with respect to the parameters θ, and the parameter domain $\Theta \subset \mathbb{R}^c$ is bounded with $\|\theta\| \le K$.

Theorem 2. *(Sample complexity of general **DINE**). Given any accuracy and confidence parameters $\epsilon, \delta > 0$, the following holds with probability at least $1 - \delta$*

$$|I(X,Y|Z) - I_n(X,Y|Z)| < \epsilon \tag{24}$$

whenever the sample size n suffices at least

$$\frac{72M^2}{\epsilon^2}\left(c \ln\left(\frac{96KL\sqrt{c}}{\epsilon}\right) + \ln\frac{2}{\delta}\right) \tag{25}$$

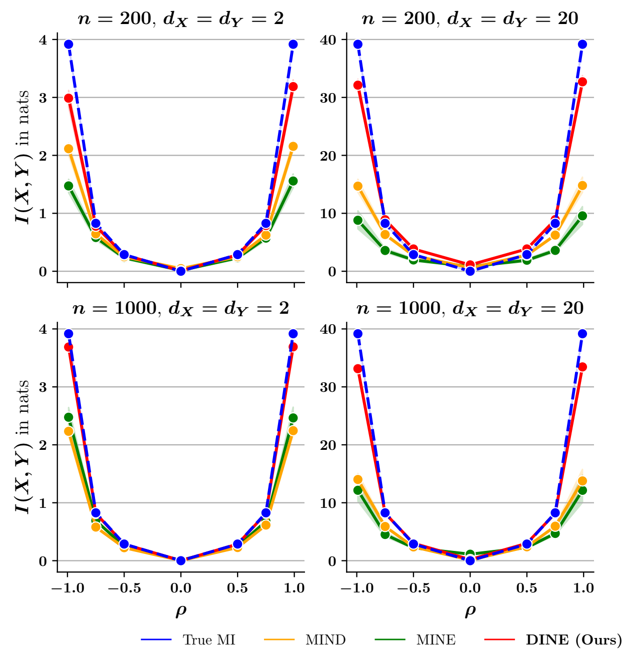

Figure 1: Mutual Information estimation performance. We compare the proposed **DINE** estimator with MINE (Belghazi et al. 2018) and MIND (Samo 2021). Rows: sample sizes, columns: dimensionalities. The dashed line denotes the true MI and the other lines show the averaged estimations for each method over 50 independent runs. The shaded areas show the estimated 95% confidence intervals.

Empirical Evaluations

In what follows, we illustrate that **DINE-Gaussian** (for brevity we refer to it as just **DINE** from now on) is far more effective than the alternative MI and CMI estimators in both sample size and dimensionality, especially when the actual information measure is high. Implementation details and parameters selection of all methods are given in the Supplementary Material (Duong and Nguyen 2022b).

Synthetic Data

We consider a diverse set of simulated scenarios covering different degrees of non-linear dependency, sample size, and dimensionality settings. For each independent simulation, we first generate two jointly multivariate Gaussian variables X', Y' with same dimensions $d_X = d_Y = d$ and shared component-wise correlation, i.e., $(X', Y') \sim \mathcal{N}\left(\mathbf{0}; \begin{bmatrix} \mathbf{I}_d & \rho\mathbf{I}_d \\ \rho\mathbf{I}_d & \mathbf{I}_d \end{bmatrix}\right)$ with a correlation $\rho \in (-1, 1)$. As for Z, we randomly choose one of three distributions $\mathcal{U}(-0.01, 0.01)^{d_Z}$, $\mathcal{N}(\mathbf{0}; 0.01\mathbf{I}_{d_Z})$, and $\mathrm{Laplace}(\mathbf{0}; 0.01\mathbf{I}_{d_Z})$. Then, X and Y are defined as

$$X = f(AZ + X') \tag{26}$$

$$Y = g(BZ + Y') \tag{27}$$

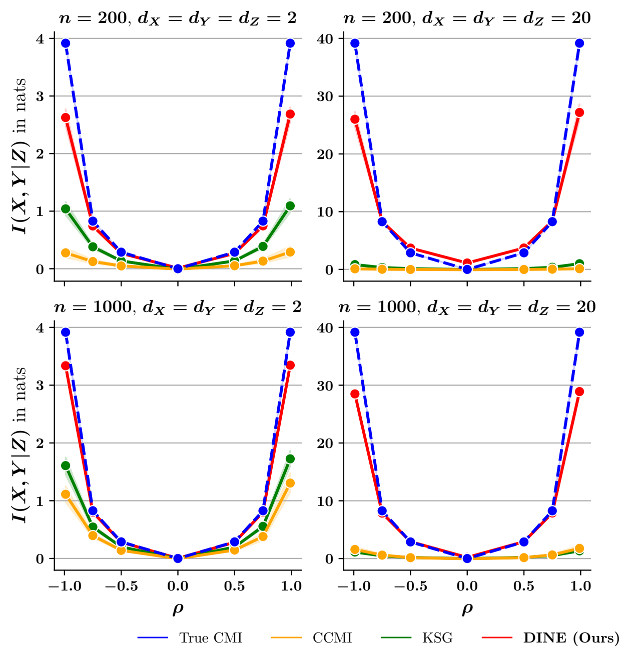

Figure 2: Conditional Mutual Information estimation performance. We compare the proposed **DINE** estimator with CCMI (Mukherjee, Asnani, and Kannan 2020) and KSG (Kraskov, Stögbauer, and Grassberger 2004). Rows: sample sizes, columns: dimensionalities. The dashed line denotes the true CMI and the other lines show the averaged estimations for each method over 50 independent runs. The shaded areas show the estimated 95% confidence intervals.

where $A, B \in \mathbb{R}^{d \times d_Z}$ have independent entries drawn from $\mathcal{N}(0; 1)$, and f, g are randomly chosen from a rich set of mostly non-linear bijective functions $f(x) \in \left\{\alpha x, x^3, e^{-x}, \frac{1}{x}, \ln x, \frac{1}{1+e^{-x}}\right\}$.[2]

By construction, we have the ground truth CMI $I(X, Y|Z) = I(X', Y') = -\frac{d}{2} \ln\left(1 - \rho^2\right)$.

Finally, n i.i.d. samples $\left\{\left(x^{(i)}, y^{(i)}, z^{(i)}\right)\right\}_{i=1}^{n}$ are generated accordingly.

Comparison with MI Estimators

We compare **DINE** with two state-of-the-arts, the variational MI lower bound estimator MINE[3] (Belghazi et al. 2018) and the inductive copula-based MIND[4] method (Samo 2021), which are two of the best approaches focusing solely on MI estimation.

In this setting, we let Z be empty, i.e., $d_Z = 0$, and vary the correlation ρ in $[-0.99, 0.99]$. We consider both the low sample size $n = 200$ and the large sample size $n = 1000$, as well as the low-dimensional $d = 2$ and high-dimensional $d = 20$ settings. For each of the setting combinations, we evaluate the methods using the same 50 independent synthetic data sets according to the described simulation scheme.

The empirical results are recorded in Figure 1. We observe that for all scenarios, our **DINE** method produces nearly identical estimates with the ground truth, regardless of sample size or dimensionality, with clear distinctions from MINE and MIND. Under the most limited setting of low sample size and high dimensionality (top-right), **DINE** estimates are still remarkably close the the ground truth, while MINE and MIND visibly struggles when the ground truth mutual information is high. On the other hand, for the most favorable setting of large sample size and low dimensionality (bottom-left), **DINE** estimates approach the ground truth with a nearly invisible margin, whereas the error gaps of MINE and MIND estimates are clearly distinguishable.

Comparison with CMI Estimators

For this context, we compare **DINE** with the state-of-the-art classifier based estimator CCMI[5] (Mukherjee, Asnani, and Kannan 2020) and the popular k-NN based estimator KSG (Kraskov, Stögbauer, and Grassberger 2004). The experiment setup follows closely to the MI estimation experiment, except that now we let $d_Z = d_X = d_Y$.

Figure 2 captures the results. We can see that, compared to the MI estimation setting, the conditioning variable Z degrades the performance of **DINE**, however only for high ground truth values and not at a considerable magnitude. Meanwhile, the competitors CCMI and KSG do not adapt well to the high dimensional setting when $d_X = d_Y = d_Z = 20$. Yet, for the low dimensional case, they still perform poorly relative to our **DINE** approach, especially when the underlying CMI is high.

Application in Conditional Independence Testing

Among the broad range of applications of CMI estimation, the Conditional Independence (CI) test is perhaps one of the most desired. CI testing greatly benefits the field of Causal Discovery (Spirtes et al. 2000). Therefore, in this Section we illustrate that our approach may be used to construct a CI test that strongly outperforms other competitive baselines solely designed for the same goal, as a down-stream evaluation of **DINE**. Resultantly, the test can be expected to improve Causal Discovery methods significantly.

Context

Formally, CI testing concerns with the statistical hypothesis test with

[2] We scale and translate the inputs before feeding into the functions to ensure numerical stability, e.g. $\frac{1}{x} \rightarrow \frac{1}{x - \min(x) + 1}$, $\ln(x) \rightarrow \ln(x - \min(x) + 1)$, and $e^{-x} \rightarrow e^{\frac{-x}{\mathrm{std}(x)}}$.

[3] We adopt the implementation at https://github.com/karlstratos/doe

[4] We use the author's implementation https://github.com/kxytechnologies/kxy-python/

[5] We use the implementation from the authors at https://github.com/sudiptodip15/CCMI

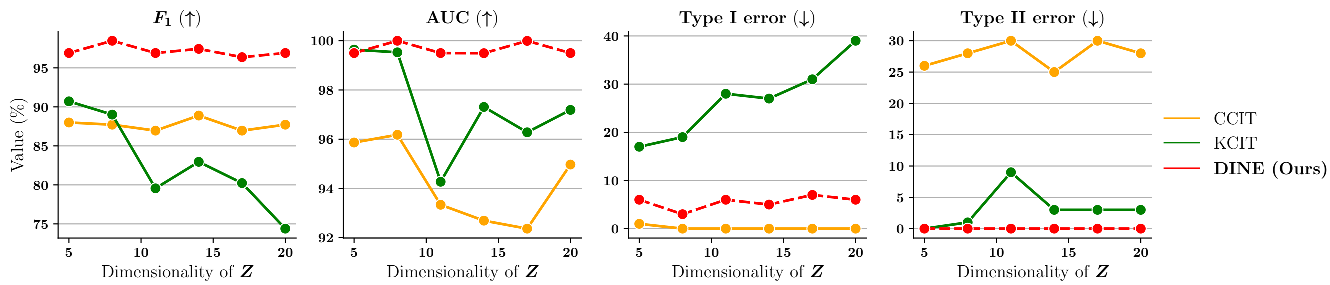

Figure 3: Conditional Independence Testing performance as a function of dimensionality. The evaluation metrics are F_1 score, AUC (higher is better), Type I and Type II error rates (lower is better), evaluated on 200 independent runs. We compare the proposed **DINE**-based CI test with KCIT (Zhang et al. 2012) and CCIT (Sen et al. 2017).

$$\mathcal{H}_0 : X \perp\!\!\!\perp Y \mid Z \quad (28)$$
$$\mathcal{H}_1 : X \not\perp\!\!\!\perp Y \mid Z \quad (29)$$

Related works In the context of CI testing, kernel-based approaches (Zhang et al. 2012), are generally the most popular and powerful methods for CI testing, which adopt kernels to exploit high order statistics that capture the CI structure of the data. Recently, more modern approaches have also been proposed, such as GAN-based (Shi et al. 2021) classification-based (Sen et al. 2017), or latent representation-based (Duong and Nguyen 2022a) methods with promising results.

Description of the test We design a simple **DINE**-based CI test inspired by the observation that $X \perp\!\!\!\perp Y \mid Z \Leftrightarrow I(X', Y') = 0$ and use $I(X', Y')$ as the test statistics. Next, we employ permutation-based boostrapping to simulate the null distribution of the test statistics and estimate the p-value. Finally, given a user-defined significance level α, we reject the null hypothesis $\mathcal{H}_0$ if p-value $< \alpha$ and accept it otherwise. The implementation details and parameters of the test are given in the Supplementary Material (Duong and Nguyen 2022b).

Experiments

To numerically evaluate the quality of the aforementioned **DINE**-based CI test, we compare it with the prominent kernel-based test KCIT[6] (Zhang et al. 2012) and a more recent state-of-the-art classifier-based test CCIT[7] (Sen et al. 2017).

In this experiment, we fix $d_X = d_Y = 1$ and consider d_Z increasing from low to high dimensionalities in $[5, 20]$, with a constant sample size $n = 1000$, and compare the performance of **DINE** against the baselines, assessed under four different criterions, namely the F_1 score, AUC (higher is better), Type I, and Type II error rates (lower is better). These metrics are evaluated using 200 independent runs (100 runs for each label) for each combination of method and dimensionality of Z. Additionally, for F_1 score, Type I, and Type II errors, we adopt the common significance level of $\alpha = 0.05$. Furthermore, for the case of conditional independence we let $\rho = 0$, whereas for the conditional dependence case we randomly draw $\rho \sim \mathcal{U}([-0.99, 0.1] \cup [0.1, 0.99])$. The data is generated according to the CMI experiment in the previous Section.

The numerical comparisons in CI testing are presented in Figure 3, which show that the **DINE**-based CI test obtains very good scores under all performance metrics. More particularly, it nearly never makes any Type II error, meaning when the relationship is actually conditional dependence, the CMI estimate is rarely too low to be misclassified as conditional independence; meanwhile, its Type I errors are roughly proximate to the rejection threshold α, which is expected from the definition of p-value. Moreover, the F_1 and AUC scores of **DINE** are also highest in all cases and closely approach 100%, suggesting the superior adaptability of **DINE** to both non-linearity and higher dimensionalities.

Regarding the baseline methods, KCIT and CCIT show completely opposite behaviors to each other. While KCIT has relatively low Type II errors, its Type I errors are quite high even at lower dimensionalities and increase rapidly in the increment dimensionality. Conversely, CCIT is quite conservative in Type I error as a trade-off for the consistently high Type II errors. However, their AUC scores are still high, indicating that the optimal threshold exists, but their p-value estimates do not reflect accurately the true p-value.

Conclusion

In this paper we propose **DINE**, a novel approach for CMI estimation. Through the use of normalizing flows, we simplify the challenging CMI estimation problem into the easier MI estimation, which can be designed to be efficiently evaluable, overcoming the inherent difficulties in existing approaches. We compare **DINE** with best-in-class methods for MI estimation, CMI estimation, in which **DINE** shows considerably better performance as compared to its counterparts, as well as being friendly in sample size and dimensionality while adapting well to several non-linear relationships. Finally, we show that **DINE** can also be used to define a CI test with an improved effectiveness in comparison with state-of-the-art CI tests, thanks to its accurate CMI estimability.

[6]The implementation from the causal-learn package is adopted https://github.com/cmu-phil/causal-learn

[7]The authors' implementation can be found at https://github.com/rajatsen91/CCIT

References

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *Proceedings of the International Conference on Machine Learning*, 531–540.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*.

Duong, B.; and Nguyen, T. 2022a. Conditional Independence Testing via Latent Representation Learning. *arXiv preprint arXiv:2209.01547*.

Duong, B.; and Nguyen, T. 2022b. Diffeomorphic Information Neural Estimation. *arXiv preprint arXiv:2211.10856*.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT Press.

Kobyzev, I.; Prince, S. J.; and Brubaker, M. A. 2020. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3964–3979.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Physical Review E*.

Kwak, N.; and Choi, C.-H. 2002. Input feature selection by mutual information based on Parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1667–1671.

Magerman, D. M.; and Marcus, M. P. 1990. Parsing a Natural Language Using Mutual Information Statistics. In *AAAI*, 984–989.

McAllester, D.; and Stratos, K. 2020. Formal limitations on the measurement of mutual information. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 875–884.

Molavipour, S.; Bassi, G.; and Skoglund, M. 2021. Neural Estimators for Conditional Mutual Information Using Nearest Neighbors Sampling. *IEEE Transactions on Signal Processing*, 766–780.

Mukherjee, S.; Asnani, H.; and Kannan, S. 2020. CCMI: Classifier based conditional mutual information estimation. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 1083–1093.

Noshad, M.; Zeng, Y.; and Hero, A. O. 2019. Scalable mutual information estimation using dependence graphs. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2962–2966.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Paninski, L. 2003. Estimation of entropy and mutual information. *Neural Computation*, 1191–1253.

Papamakarios, G.; Nalisnick, E. T.; Rezende, D. J.; Mohamed, S.; and Lakshminarayanan, B. 2021. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 1–64.

Peng, H.; Long, F.; and Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1226–1238.

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

Samo, Y.-L. K. 2021. Inductive mutual information estimation: A convex maximum-entropy copula approach. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2242–2250.

Sen, R.; Suresh, A. T.; Shanmugam, K.; Dimakis, A. G.; and Shakkottai, S. 2017. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*.

Shi, C.; Xu, T.; Bergsma, W.; and Li, L. 2021. Double generative adversarial networks for conditional independence testing. *Journal of Machine Learning Research*, 1–32.

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

Zhang, K.; Peters, J.; Janzing, D.; and Schölkopf, B. 2012. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.