C-MI-GAN : Estimation of Conditional Mutual Information Using MinMax Formulation

Arnab Kumar Mondal[*]
anz188380@cse.iitd.ac.in

Arnab Bhattacharjee [*]
arnab.bhattacharjee@uqidar.iitd.ac.in

Sudipto Mukherjee[†]
sudipm@uw.edu

Prathosh AP [*]
prathoshap@ee.iitd.ac.in

Sreeram Kannan [†]
ksreeram@uw.edu

Himanshu Asnani [‡]
himanshu.asnani@tifr.res.in

Abstract

Estimation of information theoretic quantities such as mutual information and its conditional variant has drawn interest in recent times owing to their multifaceted applications. Newly proposed neural estimators for these quantities have overcome severe drawbacks of classical kNN-based estimators in high dimensions. In this work, we focus on conditional mutual information (CMI) estimation by utilizing its formulation as a *minmax* optimization problem. Such a formulation leads to a joint training procedure similar to that of generative adversarial networks. We find that our proposed estimator provides better estimates than the existing approaches on a variety of simulated datasets comprising linear and non-linear relations between variables. As an application of CMI estimation, we deploy our estimator for conditional independence (CI) testing on real data and obtain better results than state-of-the-art CI testers.

1 INTRODUCTION

Quantifying the dependence between random variables is a quintessential problem in data science (Rényi 1959; Joe 1989; Fukumizu et al. 2008). A widely used measure across statistics is the Pearson correlation and partial correlation. Unfortunately, these measures can capture and quantify only linear relation between variables and do not extend to non-linear cases. The field of information theory (Cover and Thomas 2012) gave rise to multiple functionals of data density to capture the dependence between variables even in non-linear cases. Two noteworthy quantities of widespread interest are the mutual information (MI) and conditional mutual information (CMI).

In this work, we focus on estimating CMI, a quantity which provides the degree of dependence between two random variables X and Y given a third variable Z. CMI provides a strong theoretical guarantee that $I(X;Y|Z) = 0 \iff X \perp Y|Z$. So, one motivation for estimating CMI is its use in conditional independence (CI) testing and detecting causal associations. CI tester built using kNN based CMI estimator coupled with a local permutation scheme (Runge (2018)) was found to be better calibrated than the kernel tests. CMI was used for detecting and quantifying causal associations in spike trains data from neuron pairs (Li et al. 2011). Runge et al. (2019) demonstrate how CMI estimator can be combined with a causal recovery algorithm to identify causal links in a network of variables.

Apart from its use in CI testers, CMI has found diverse applications in feature selection, communication, network inference and image processing. Selecting features iteratively so that the information is maximized given already selected features was the basis for Conditional Mutual Information Maximization (CMIM) criterion in Fleuret (2004). This principle was applied by Wang and Lochovsky (2004) for text categorization, where the number of features are quite large. Efficient methods for CMI based feature selection involving more than one conditioning variable was developed by Shishkin et al. (2016). In the field of communications, Yang and Blum (2007) maximized CMI between target and reflected waveforms for optimal radar waveform design. For learning gene regulatory network, in Zhang et al. (2012) CMI was used as a measure of dependence between genes. A similar approach was adapted for protein modulation in Giorgi et al. (2014). Finally, Loeckx et al. (2009) used CMI as a similarity metric for non-rigid image registration. Given the widespread use of CMI as a measure of conditional dependence, there is a pressing

[*]Affiliated with IIT Delhi, India.
[†]Affiliated with University of Washington, USA.
[‡]Affiliated with TIFR Mumbai, India.

Proceedings of the 36[th] Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

need to accurately estimate this quantity, which we seek to achieve in this paper.

2 RELATED WORK

One of the simplest methods for estimating MI (or CMI) could be based on the binning of the continuous random variables, estimating probability densities from the bin frequencies and plugging it in the expression for MI (or CMI). Kernel methods, on the other hand, estimate the densities using suitable kernels. The most widely used estimator of MI, the KSG estimator, is based on k nearest neighbor statistics (Kraskov et al. 2004) and has been shown to outperform binning or kernel-based methods. KSG is based on expressing MI in terms of entropy

$$I(X;Y) = H(X) + H(Y) - H(X;Y) \qquad (1)$$

The entropy estimation follows from $H(X) = -N^{-1}\sum_i \log \widehat{\mu(x_i)}$ (Kozachenko and Leonenko 1987). Distance of the k nearest neighbors of point x_i is used to approximate the density $\mu(x_i)$. The KSG estimator does not estimate each entropy term independently, but accounts for the appropriate bias correction terms in the overall estimation. It ensures that an adaptive k is used for distances in marginal spaces X, Y and for the joint space (X, Y). Several later works studied the theoretic properties of the KSG estimator and sought to improve its accuracy (Gao et al. 2015, 2016, 2018; Póczos and Schneider 2012). Since CMI can be expressed as a difference of two MI estimates, $I(X;Y|Z) = I(X;YZ) - I(X;Z)$, KSG estimator could be used for CMI as well. Even though KSG estimator enjoys the favorable property of consistency, its performance in finite sample regimes suffers from the curse of dimensionality. In fact, KSG estimator requires exponentially many samples for accurate estimation of MI (Gao et al. 2015). This limits its applicability in high dimensions with few samples.

Deviating from the kNN-based estimation paradigm, Belghazi et al. (2018) proposed a neural estimation of MI (referred to as MINE). This estimator is built on optimizing dual representations of the KL divergence, namely the Donsker-Varadhan (Donsker and Varadhan 1975) and the f-divergence representation (Nguyen et al. 2010). MINE is strongly consistent and scales well with dimensionality and sample size. However, recent works found the estimates from MINE to have high variance (Poole et al. 2019; Oord et al. 2018) and the optimization to be unstable in high dimensions (Mukherjee et al. 2019). To counter these issues, variance reduction techniques were explored in Song and Ermon (2020).

While for the estimation of MI we need to perform the trivial task of drawing samples from the marginal distribution, CMI estimation adds another layer of intricacy to the problem. For the above approaches to work for CMI, one needs to obtain samples from the conditional distribution. In Mukherjee et al. (2019), the authors separate the problem of estimating CMI into two stages by first estimating the conditional distribution and then using a divergence estimator. However, being coupled with an initial conditional distribution sampler, this technique is limited by the goodness of the conditional samplers and thus may be sub-optimal. Even when CMI is obtained as a difference of two separate MI estimates (CCMI estimator in Mukherjee et al. (2019)), there is no guarantee that the bias values would be same from both MI terms, thereby leading to incorrect estimates. Based on these observations, in this paper, we attempt to estimate CMI using a joint training procedure involving a min-max formulation devoid of explicit conditional sampling.

The main contributions of our paper are as follows:

- We formulate CMI as a *minimax* optimization problem and illustrate how it can be estimated from joint training. The estimation process has similar flavor to adversarial training (Goodfellow et al. 2014) and so the term C-MI-GAN (read "See-Me-GAN") is coined for the estimator.
- We empirically show that estimates from C-MI-GAN are closer to the ground truth on an average as compared to the estimates of other CMI estimators.
- We apply our estimator for conditional independence testing on a real flow-cytometry dataset and obtain better results than state-of-the-art CI Testers.

3 PROPOSED METHODOLOGY

Information Theoretic Quantities. Let X, Y and Z be three continuous random variables that admit densities. The mutual information between two random variables X and Y measures the amount of dependence between them and is defined as

$$I(X;Y) = \iint P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} dxdy \quad (2)$$

It can also be expressed in terms of the entropies as follows:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (3)$$

Here $H(X)$ is the entropy[1] and is given by $H(X) = -\int p(x)\, log\, p(x)\, dx$. The above expression provides

[1]More precisely, differential entropy in case of continuous random variables.

the intuitive explanation of how the information content changes when the random variable is alone versus when another random variable is given.

The conditional mutual information extends this to the setting where a conditioning variables is present. The analogous expression for CMI is:

$$I(X;Y|Z) = \iiint P_{XYZ} \log \frac{P_{XYZ}}{P_{XZ}P_{Y|Z}} dxdydz \quad (4)$$

In terms of the entropies, it can be expressed as follows:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \quad (5)$$
$$= H(Y|Z) - H(Y|X,Z) \quad (6)$$

Both MI and CMI are special cases of a statistical quantity called KL-divergence, which measures how different one distribution is from another. The KL-divergence between two distributions P_X and Q_X is as follows:

$$D_{KL}(P_X||Q_X) = \int P_X(x) \log \frac{P_X(x)}{Q_X(x)} dx \quad (7)$$

MI and CMI can be defined using KL-divergence as:

$$I(X;Y) = D_{KL}(P_{XY}||P_X P_Y) \quad (8)$$

$$I(X;Y|Z) = D_{KL}(P_{XYZ}||P_{XZ}P_{Y|Z}) \quad (9)$$

This definition of MI (Equation 8) tries to capture how much the given joint distribution is different from X and Y being independent (conditionally independent in case of CMI). Various estimators in the literature aim to utilize a particular expression of MI (or CMI), while avoiding computation of density functions explicitly. While KSG ((Kraskov et al. 2004)) is based on the summation of entropy terms, MINE ((Belghazi et al. 2018)) derives its estimates based on lower bounds of KL-divergence.

Lower bounds of Mutual Information. The following lower bounds of KL-divergence (hence also mutual information) were used in Belghazi et al. (2018) for the MINE estimator.

Donsker-Varadhan bound : This bound is tighter and is given by :

$$D_{KL}(P||Q) = \sup_{R \in \mathcal{R}} (\mathbb{E}_P[R] - \log(\mathbb{E}_Q[e^R])) \quad (10)$$

f-divergence bound : A slightly loose bound is given by the following relation :

$$D_{KL}(P||Q) = \sup_{R \in \mathcal{R}} (\mathbb{E}_P[R] - \mathbb{E}_Q[e^{R-1}]) \quad (11)$$

The supremum in both the bounds (equation 10 and 11) is over all functions $R \in \mathcal{R}$ such that the expectations are finite. MINE uses a neural network as a parameterized function R_β, which is optimized with these bounds.

3.1 MIN-MAX FORMULATION FOR CMI

Building on top of these lower bounds, we further take resort to a variational form of conditional mutual information. Observing closely the expression for $I(X;Y|Z)$ in equation 4, we find that samples need to be drawn from $p(y|z)$ which is not available from given data (X,Y,Z) directly. One approach used in Mukherjee et al. (2019) is to learn $p(y|z)$ using a conditional GAN, kNN or conditional VAE. *Can we combine this step directly with the lower bound maximization ?*

We first note that the CMI expression can be upper bounded as follows :

$$\begin{aligned} I(X;Y|Z) &= D_{KL}(P_{XYZ}||P_{XZ}P_{Y|Z}) \\ &= D_{KL}(P_{XYZ}||P_{XZ}Q_{Y|Z}) \\ &\quad - D_{KL}(P_{Y|Z}||Q_{Y|Z}) \\ &\leq D_{KL}(P_{XYZ}||P_{XZ}Q_{Y|Z}) \end{aligned} \quad (12)$$

since $D_{KL}(P||Q) \geq 0$. In equation 12 the equality is achieved when $Q_{Y|Z} = P_{Y|Z}$ and we can express $I(X;Y|Z)$ as

$$I(X;Y|Z) = \inf_{Q_{Y|Z}} D_{KL}(P_{XYZ}||P_{XZ}Q_{Y|Z}) \quad (13)$$

Equation 13 coupled with the Donsker-Varadhan bound (equation 10) leads to a min-max optimization for MI estimation as follows:

$$\begin{aligned} I(X;Y|Z) = \inf_{Q_{Y|Z}} \sup_{R \in \mathcal{R}} \Big(& \mathop{\mathbb{E}}_{s \sim P_{XYZ}} [R(s)] \\ & - \log \big(\mathop{\mathbb{E}}_{s \sim P_{XZ}Q_{Y|Z}} [e^{R(s)}] \big) \Big) \end{aligned} \quad (14)$$

Equation 14 offers a pragmatic approach for estimating CMI. Since neural nets are universal function approximators, it is a possibility to deploy one such network to approximate the variational distribution $Q_{Y|Z}$ and another for learning the regression function given by R. The following section provides a detailed narration of how to achieve this objective.

3.2 C-MI-GAN

To begin with, we elaborate different components of the proposed estimator - C-MI-GAN. As depicted in Figure 1, the variational distribution, $Q_{Y|Z}$ is parameterized using a network denoted as G_θ. In other words, G_θ is capable of sampling from the distribution $Q_{Y|Z}$, hence it is called the generator network. The regression network, R_ϕ parameterizes the function class on the R.H.S. of the Donsker-Varadhan identity (refer to equation 10). Gaussian noise concatenated with conditioning variable Z is fed as input to G_θ. R_ϕ is trained with samples from

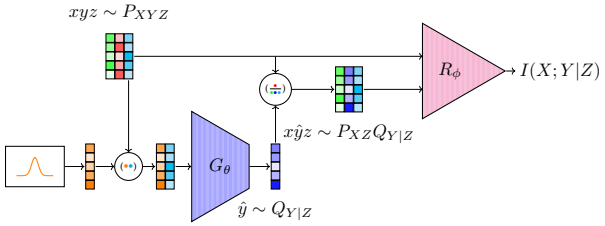


Figure 1: Block Diagram for C-MI-GAN (Best viewed in colour). Samples drawn from any simplistic noise distribution are concatenated with the samples from the marginal P_Z and fed to the generator as input. The generated samples from the variational distribution $Q_{Y|Z}$ are then concatenated with samples from P_{XZ} and given as input to the regression network along with samples from P_{XYZ}. $I(X;Y|Z)$ is obtained by negating the loss of the trained regression network.

P_{XYZ} and $P_{XZ}Q_{Y|Z}$. During training, we optimize the regression network and the generator jointly using the objective function $h(Q_{Y|Z}, R)$ as defined below.

$$h(Q_{Y|Z}, R) = \inf_{Q_{Y|Z}} \sup_{R \in \mathcal{R}} \left(\int_{s \sim P_{XYZ}} P_{XYZ} R(s) ds - \log \left(\int_{s \sim P_{XZ}Q_{Y|Z}} P_{XZ} Q_{Y|Z} e^{R(s)} ds \right) \right) \tag{15}$$

In each training loop, we optimize the parameters of R_ϕ and G_θ, using a learning schedule and RMSProp optimizer. The detailed procedure is described in Algorithm 1. Upon successful completion of training of the joint network, G_θ starts generating samples from the distribution $P_{Y|Z}$ and the output of the regression network converges to $I(X;Y|Z)$.

Next, we formally show that the alternate optimization of R_ϕ and G_θ optimizes the objective function defined in equation 15 and when the global optima is reached, the optimal value of the objective function coincides with CMI. To start with, we derive the expression for the optimal regression network and subsequently show that the optimal value of the objective function coincides with CMI under the optimal regression network.

Theorem 1. *For a given generator, G , the optimal regression network, R^*, is*

$$R^* = \log \frac{P_{XYZ}}{P_{XZ}Q_{Y|Z}} + c \tag{16}$$

Where c is any constant. P_{XYZ}, P_{XZ}, and $Q_{Y|Z}$ denote the data distribution, marginal distribution and generator distribution respectively.

Proof. [2] For a given generator, G, the regression network's objective is to maximize the quantity $h(Q_{Y|Z}, R)$.

$$h(Q_{Y|Z}, R) = \int_{s \sim P_{XYZ}} P_{XYZ} R(s) ds - \log \left(\int_{s \sim P_{XZ}Q_{Y|Z}} P_{XZ} Q_{Y|Z} e^{R(s)} ds \right) \tag{17}$$

For the optimum regression network, R^*,

$$\begin{aligned} & \left. \frac{\partial h}{\partial R} \right|_{R^*} = 0 \\ \implies & P_{XYZ} - \frac{P_{XZ}Q_{Y|Z}e^{R^*}}{\int P_{XZ}Q_{Y|Z}e^{R^*} ds} = 0 \\ \implies & \frac{P_{XZ}Q_{Y|Z}e^{R^*}}{P_{XYZ}} = \int P_{XZ}Q_{Y|Z}e^{R^*} ds = e^c \\ \implies & R^* = \log \frac{P_{XYZ}}{P_{XZ}Q_{Y|Z}} + c \end{aligned} \tag{18}$$

□

Now we show that with the optimal regression network R^*, CMI is obtained when the objective function achieves its minima.

Theorem 2. *$h(Q_{Y|Z}, R^*)$ achieves its minimum value $I(X;Y|Z)$, iff $Q_{Y|Z} = P_{Y|Z}$.*

Proof.

$$\begin{aligned} h(Q_{Y|Z}, R^*) &= \int_s P_{XYZ} \log \frac{P_{XYZ}}{P_{XZ}Q_{Y|Z}} dxdydz - \log \left(\int_s P_{XZ}Q_{Y|Z} e^{\log \frac{P_{XYZ}}{P_{XZ}Q_{Y|Z}}} dxdydz \right) \\ &= \int_s P_{XYZ} \log \frac{P_{XYZ}}{P_{XZ}Q_{Y|Z}} dxdy - \int_s P_{XYZ} dxdydz \\ &= \int_s P_{XYZ} \log \frac{P_{XYZ}}{P_{XZ}P_{Y|Z}} dxdydz + \int_s P_{Y|Z} \log \frac{P_{Y|Z}}{Q_{Y|Z}} dy \\ &= I(X;Y|Z) + D_{KL}(P_{Y|Z} || Q_{Y|Z}) \end{aligned} \tag{19}$$

[2] This proof assumes $P_{XYZ}(s), P_{XZ}Q_{Y|Z}(s) > 0 \; \forall s$.

Since $D_{KL}(P_{Y|Z}||Q_{Y|Z}) \geq 0$, when $P_{Y|Z} = Q_{Y|Z}$, $D_{KL}(P_{Y|Z}||Q_{Y|Z}) = 0$ and the minimum value is $h(P_{Y|Z}, R^*) = I(X;Y|Z)$. □

The alternate optimization of R_ϕ and G_θ is similar to the generative adversarial networks (Goodfellow et al. (2014), Nowozin et al. (2016)), in that both are trained using similar adversarial training procedure. However, C-MI-GAN significantly differs from traditional GAN in the following sense:

- The regression task is completely unsupervised as no target value is used to train the network.
- The proposed loss function to estimate CMI is foreign to traditional GAN literature.
- The binary discriminator in traditional GAN is replaced by a regression network, R_ϕ that estimates the CMI (refer to Figure 1).

Algorithm 1 Pseudo code for C-MI-GAN

Inputs: $\mathcal{D} = \{x^{(i)}, y^{(i)}, z^{(i)}\}_{i=1}^{m} \sim P_{XYZ}$
Outputs: $I(X;Y|Z)$

function CMIGAN
  for $r \leftarrow 1$ to N **do**
    Initialize R_ϕ and G_θ
    for $i \leftarrow 1$ to $training\_steps$ **do**
      for $j \leftarrow 1$ to $reg\_training\_ratio$ **do**
        Shuffle $\mathcal{D} \sim P_{XYZ}$
        $\mathcal{D}_b \leftarrow \{x^{(k)}, y^{(k)}, z^{(k)}\}_{k=1}^{s} \sim P_{XYZ}$
        $noise \leftarrow \{n^{(k)}\}_{k=1}^{s} \sim \mathcal{N}(0, I_{d_n})$
        $\{y_\theta^{(k)}\}_{k=1}^{s} \leftarrow G_\theta(noise, Z_b)$
        $\hat{\mathcal{D}}_b \leftarrow \{x^{(k)}, y_\theta^{(k)}, z^{(k)}\}_{k=1}^{s}$
        $L_{reg} \leftarrow -\mathbb{E}_{\mathcal{D}_b}[R_\phi] + \log(\mathbb{E}_{\hat{\mathcal{D}}_b}[e^{R_\phi}])$
        Minimize L_{reg} and Update ϕ
      end for
      $L_{gen} \leftarrow -\log(\mathbb{E}_{\hat{\mathcal{D}}_b}[e^{R_\phi}])$
      Minimize L_{gen} and Update θ
      Update learning rate as per scheduler.
    end for
    $\hat{I}_n(X;Y|Z) \leftarrow -L_{reg}$
  end for
  $\hat{I}(X;Y|Z) \leftarrow \frac{1}{N} \sum_{n=1}^{N} \hat{I}_n(X;Y|Z)$
end function

4 EXPERIMENTAL RESULTS

In this section we compare the CMI estimates, on different datasets, of our proposed method against the estimations of the state of the art CMI estimators such as f-MINE (Belghazi et al. (2018)) and CCMI (Mukherjee et al. (2019)). We design similar experiments on similar datasets as Mukherjee et al. (2019) to demonstrate the efficacy of our proposed estimator: C-MI-GAN. Unlike the method proposed in this work, the existing methods rely on a separate generator for generating samples from the conditional distribution. Therefore, "Generator"+"Divergence estimator" notation is used to denote the estimators used as baseline. For example, CVAE+f-MINE implies that Conditional VAE (Sohn et al. (2015)) is used for generating samples from $P_{Y|Z}$ and f-MINE (Belghazi et al. (2018)) is used for divergence estimation. MI difference based estimators are represented as MI-Diff.+"Divergence estimator". For baseline models we have used the codes available in the repository of Mukherjee et al. (2019). Architecture for R_ϕ and G_θ and hyper-parameter settings for our proposed method are provided in the supplementary.

To illustrate C-MI-GAN's effectiveness in estimating CMI, we consider two synthetic and one real datasets:

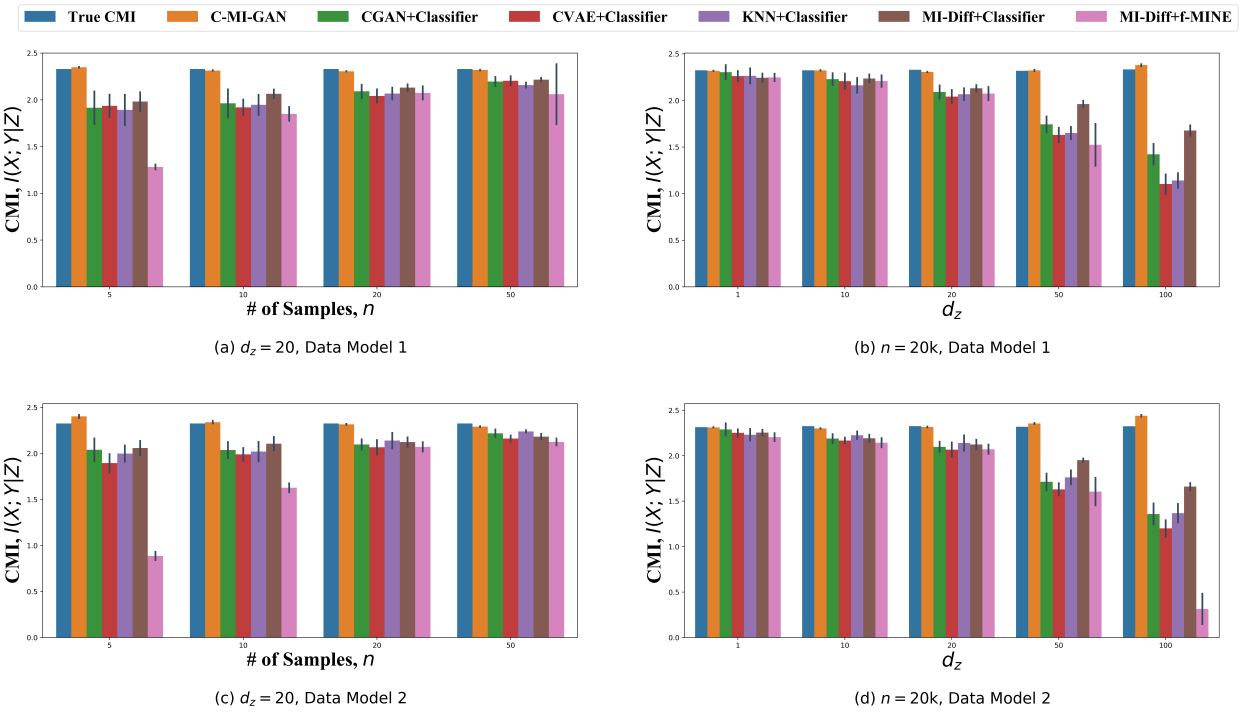
- Synthetically generated datasets having linear dependency.
- Synthetically generated datasets having non-linear dependency.
- Air quality real dataset[3,4] (Vito et al. (2008))

The most severe problem with the existing CMI estimators is that their performance drop significantly with increase in data dimension. To see how well the proposed estimator fares, compared to the existing estimators for high dimensional data, we vary the dimension, d_z of the conditioning variable, Z over a wide range, in both the datasets. For the non-linear dataset $d_x = d_y = 1$, as found commonly in the literature on causal discovery and independence testing (Mukherjee et al. (2019), Sen et al. (2017), Doran et al. (2014)). To validate, how well the proposed estimator performs on a multi-dimensional X and Y, we vary $d_x = d_y \in \{1, 5, 10, 15\}$ for the linear dataset. Besides, we consider datasets having as low as 5k to as high as 50k samples to understand the behaviour of the estimators as sample complexity varies.

Ground Truth CMI: For the datasets with linear dependency, ground truth CMI can be computed by numerical integration. However, to the best of our knowledge, there is no analytical formulation to compute the ground truth CMI for the synthetic datasets with non-linear dependency. As a workaround to this issue, as

[3] `http://archive.ics.uci.edu/ml/datasets/Air+Quality`

[4] For a detailed description of the dataset please refer to the supplementary material.



(a) $d_z = 20$, Data Model 1

(b) $n = 20$k, Data Model 1

(c) $d_z = 20$, Data Model 2

(d) $n = 20$k, Data Model 2

Figure 2: Performance of CMI estimators on the dataset generated using linear models. (a) Model 1: CMI estimates with fixed $d_z = 20$ and variable sample size, (b) Model 1: CMI estimates with fixed sample size, $n = 20$k and variable d_z, (c) Model 2: CMI estimates with fixed $d_z = 20$ and variable sample size, (d) Model 2: CMI estimates with fixed number of samples, $n = 20k$ and variable d_z. Average over 10 runs is plotted. Variation in the estimates are measured using standard deviation and are highlighted in the plots using the thin dark lines on top of the bar plots. The proposed method, C-MI-GAN provides closer estimate of the true CMI, while the state of the art estimators largely underestimate its value. C-MI-GAN outperform the state of the art methods in terms of variation in estimates as well. (Best viewed in color)

proposed by Mukherjee et al. (2019), we transform Z as $U = A_{zy}Z$, where A_{zy} is a random vector with entries drawn independently from $\mathcal{N}(0, 1)$ and then normalized to have unit norm. Following this transformation, $I(X; Y|Z) = I(X; Y|U)$. Since, U has unity dimension $I(X; Y|U)$ can be estimated accurately using KSG given sufficient samples, as shown by Gao et al. (2018) in the asymptotic analysis of KSG. Hence, a set of 50000 samples is generated separately for each data-set to estimate $I(X; Y|U)$ and the estimated value is used as the ground truth for that data-set.

Next, as a practical application of CMI estimation, we test the null hypothesis of conditional independence on a synthetic dataset, and on real flow-cytometry data.

4.1 CMI ESTIMATION

4.1.1 Dataset With Linear Dependence

We consider the following data generative models.

- **Model 1:** $X \sim \mathcal{N}(0, 1); Z \sim \mathcal{U}(-0.5, 0.5)^{d_z}; \epsilon \sim \mathcal{N}(Z_1, 0.01); Y \sim X + \epsilon$.
  In this model X is sampled from standard normal distribution. Value of each dimension of Z is drawn from a uniform distribution with support $[-0.5, 0.5]$. Finally Y is obtained by perturbing X with ϵ, where ϵ comes from a Gaussian distribution having mean Z_1, the first dimension of Z and variance 0.01. Therefore, the dependence between X and Y is through the first dimension of the conditioning variable Z.

- **Model 2:** $X \sim \mathcal{N}(0, 1); Z \sim \mathcal{N}(0, 1)^{d_z}; U = w^T Z; ||w||_1 = 1; \epsilon \sim \mathcal{N}(U, 0.01); Y \sim X + \epsilon$
  Unlike in model 1, here the dimensions of Z comes from standard normal distribution and the mean of ϵ is weighted average of all the dimensions of Z. The weight vector w is constant for a particular dataset and varies across datasets generated by model 2.

- **Model 3:** $X \sim \mathcal{N}(0, 0.25)^{d_x}; Z^{(i)} \sim \mathcal{U}(-0.5, 0.5)^{d_z}; \epsilon \sim \mathcal{N}(Z_1, 0.25)^{d_z}; Y \sim X + \epsilon$.
  This model is very similar to Model 1, except for the fact $d_x = d_y = d_z \in \{5, 10, 15, 20\}$.

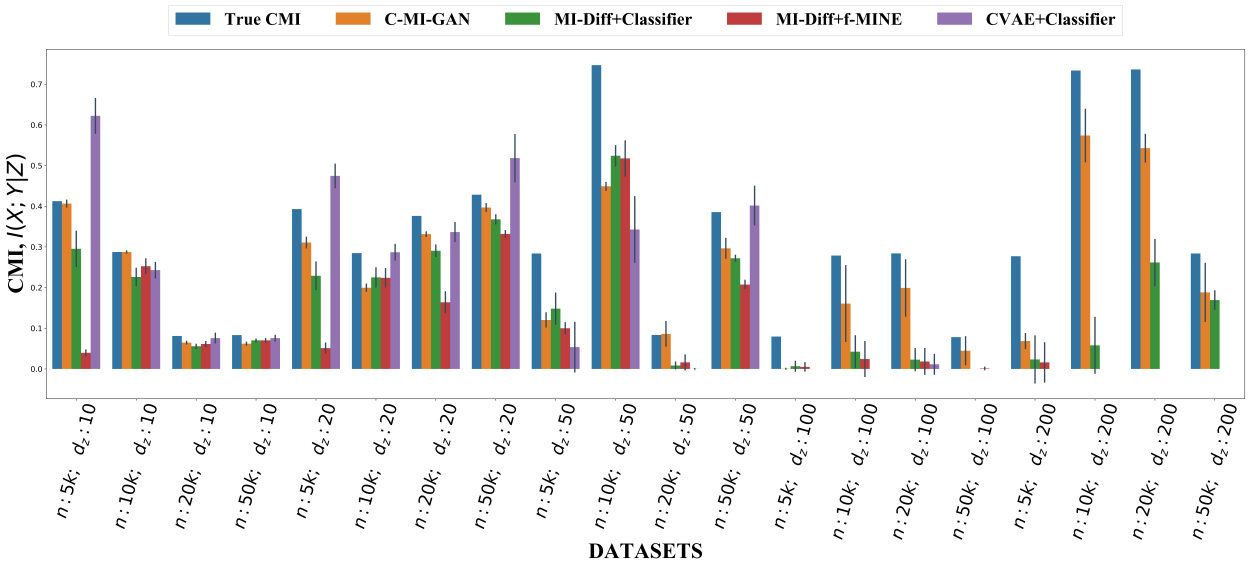


Figure 3: This figure compares the performance of the different CMI estimators on all the 20 non linear datasets. However, due to very poor performance of "Generator"+"Classifier" estimators, we plot the estimates of "CVAE"+"Classifier" only as a representative of that class of estimators. Estimated CMI, averaged over 10 runs is plotted. Standard deviation is indicated with the thin dark lines on top of the bar plot. Like in the linear case, the proposed C-MI-GAN outperforms the state of the art estimators in terms of both average CMI estimation and variation in estimation (Best viewed in color).

> Since, each dimension of the random variables are independent of each other, the ground truth CMI is estimated as sum of CMI of each dimension.

For Models 1 and 2, to study the effect of sample size on the estimate we generate data by fixing $d_z = 20$ and $n \in \{5000, 10000, 20000, 50000\}$. Next, we fix $n = 20000$ and $d_z \in \{1, 10, 20, 50, 100\}$ to observe the effect of dimension on the estimation.

Figure 2 compares the average estimated CMI over 10 runs for linear datasets. C-MI-GAN estimates are usually closer to the ground truth and exhibits less variation as compared to other estimators.

Next, we consider a multidimensional dataset generated using Model 3. We tabulate (refer to Table 1) the average estimate of C-MI-GAN and the current state-of-the-art CCMI[5] (Mukherjee et al. (2019)) over 10 executions.

Table 1: CMI Estimates on Multidimensional Linear Dataset Generated Using Model 3. All datasets have $d_x = d_y = d_z$ (Dim.). C-MI-GAN estimates are closer to ground truth and has less standard deviation.

Dim.	True CMI	CCMI	C-MI-GAN
5	1.75	$1.61 \pm 4.45e-2$	$\mathbf{1.7 \pm 1.6e-5}$
10	3.48	$2.96 \pm 9.87e-2$	$\mathbf{3.34 \pm 1.49e-4}$
15	5.22	$4.2 \pm 2.65e-1$	$\mathbf{5.03 \pm 6.57e-4}$
20	6.99	$4.8 \pm 4.71e-1$	$\mathbf{6.91 \pm 4.32e-2}$

4.1.2 Dataset With Non-Linear Dependence

Data generating model:

$$Z \sim \mathcal{N}(\mathbb{1}, I_{d_z})$$

$$X = f_1(\eta_1)$$

$$Y = f_2(A_{zy}Z + A_{xy}X + \eta_2)$$

Where, $f_1, f_2 \in \{\cos(\cdot), \tanh(\cdot), \exp(-|\cdot|)\}$ and selected randomly; $\eta_1, \eta_2 \sim \mathcal{N}(0, 0.1)$. The elements of the random vector A_{zy} are drawn independently from $\mathcal{N}(0, 1)$. The vector is then normalized to have unit norm. Since, $d_x = d_y = 1$, $A_{xy} = 2$ is a scalar.

To generate the data we consider all possible combinations of $n \in \{5000, 10000, 20000, 50000\}$ and $d_z \in \{10, 20, 50, 100, 200\}$, and obtain a set of 20 data-sets.

Figure 3 plots the average estimate of different estimators over 10 runs for all 20 data-sets. Error bar (standard deviation) is plotted as well on top of the estimation. Only MI-Diff.+Classifier among the existing estimators provides reasonable estimate when d_z is high, while the proposed method tracks the true CMI more closely.

4.1.3 Real Data: Air Quality Data Set

Finally, we estimate CMI on a real dataset of air quality (Vito et al. (2008)) using C-MI-GAN. We tabulate the estimates obtained in Table 2. We consider the causal graph

discovered by Runge (2018) representing the dependencies between three pollutants and three meteorological factors (refer to supplementary material) as the ground truth. The graph captures the strength of these dependencies. However, the ground truth CMI being unknown, we could validate only the ordering of the estimated CMI. As can be seen from the table, our estimates are in coherence with the findings of Runge (2018) and the estimates of CCMI[5] (Mukherjee et al. (2019)) and KSG (Kraskov et al. (2004)).For example, $I(\mathrm{CO}; \mathrm{C_6H_6}|\mathrm{T})$ is higher as compared to $I(\mathrm{C_6H_6}; \mathrm{RH}|\mathrm{T})$, indicating a relatively stronger conditional dependency between the former pair. This is in compliance to the ground truth causal graph (see supplementary material).

Table 2: CMI Estimation: Air Quality Dataset

X	Y	Z	CCMI	KSG	C-MI-GAN
CO	C_6H_6	T	0.61	0.65	0.66
CO	C_6H_6	NO_2	0.33	0.37	0.37
CO	C_6H_6	RH	0.56	0.60	0.59
CO	C_6H_6	AH	0.58	0.63	0.62
NO_2	C_6H_6	AH	0.40	0.45	0.45
NO_2	C_6H_6	T	0.38	0.43	0.44
NO_2	CO	C_6H_6	0.06	0.11	0.12
NO_2	RH	C_6H_6	0.01	0.05	0.05
C_6H_6	AH	T	0.01	0.07	0.07
C_6H_6	RH	T	0.02	0.06	0.06
CO	AH	C_6H_6	0.03	0.06	0.07
RH	CO	C_6H_6	0.04	0.09	0.09

4.2 APPLICATION: CONDITIONAL INDEPENDENCE TESTING (CIT)

4.2.1 Synthetic Dataset

To evaluate the proposed CMI estimator on an application, we consider testing the null hypothesis of conditional independence, used widely in conventional literature (Sen et al. (2017); Mukherjee et al. (2019)). The objective here is to decide, whether X and Y are independent given Z when we have access to samples from the joint distribution. Formally, given samples from the distributions $P(x, y, z)$ and $Q(x, y, z)$ where $Q(x, y, z) = P(x, z)P(y|z)$, we have to test our estimators on the hypothesis testing framework given by the null, $H_0 : X \perp Y|Z$ and the alternative, $H_1 : X \not\perp Y|Z$.

The conditional independence test setting will be used to test our estimator based on the fact that $X \perp Y|Z \iff I(X; Y|Z) = 0$. A simple rule can thus be established: reject the null hypothesis if $I(X; Y|Z)$ is greater than some threshold (to allow some tolerance in the estimation) and accept it otherwise.

[5]MI-diff + Classifier

CIT can be cast as binary classification problem where samples belong to either class-CI or class-CD. Therefore, area under ROC curve (AuROC) is a good metric to compare the performance of different algorithms. Therefore, we consider the AuROC scores of different models for performance comparison.

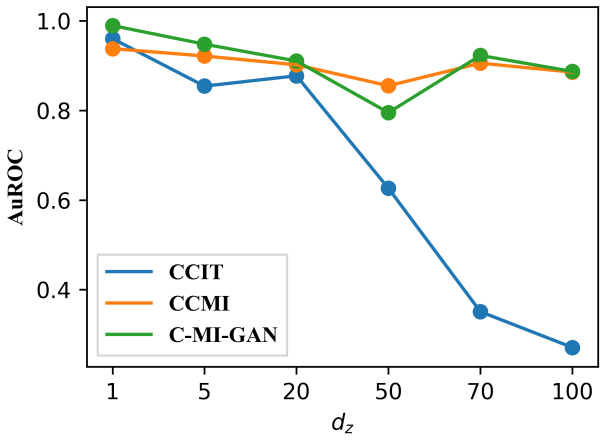


Figure 4: Performace of CCIT degrades with d_z. CCMI and C-MI-GAN are comparable across all d_z. (Best viewed in color).

Synthetic data is generated using the post non-linear noise model as used by Sen et al. (2017) and Mukherjee et al. (2019). The data generation model is as follows:

$$Z \sim \mathcal{N}(\mathbb{1}, I_{d_z})$$

$$X = \cos(a_x Z + \eta_1)$$

$$Y = \begin{cases} \cos(b_y Z + \eta_2), & \text{if } X \perp Y|Z \\ \cos(cX + b_y Z + \eta_2), & \text{otherwise} \end{cases}$$

$\eta_1, \eta_2 \sim \mathcal{N}(0, 0.25)$; $a_x, b_y \sim \mathcal{U}(0, 1)^{d_z}$ and normalized such that $||a_x||_2 = ||b_y||_2 = 1$; $c \sim \mathcal{U}(0, 2)$. As before the model parameters $a_x, b_y,$ and c are kept constant for a particular dataset but varies across datasets. $d_x = d_y = 1$ and $d_z \in \{1, 5, 20, 50, 70, 100\}$. 100 datasets consisting of 50 conditionally independent and 50 conditionally dependent datasets are generated for each d_z. Sample size of each dataset is fixed as $n = 5000$.

Figure 4 compares the performance of C-MI-GAN with CCIT (Sen et al. (2017)) and CCMI (Mukherjee et al. (2019)). Performance of CCIT degrades rapidly as d_z increases. Performance of C-MI-GAN and CCMI remains comparable for all d_z. Performance of C-MI-GAN remains undeterred with increasing dimensions.

4.2.2 Flow-Cytometry: Real Data

To test the efficacy of our proposed method in conditional independence testing on real data, we have

used Flow cytometry dataset introduced by Sachs et al. (2005). This dataset quantifies the availability of 11 biochemical compounds in single human immune system cells under different test conditions. Please refer to the supplementary material for the consensus network, which serves the purpose of ground truth. It depicts the causal relations between the 11 biochemical compounds.

The underlying concept for generating the CI and CD datasets is similar to that used in Sen et al. (2017) and Mukherjee et al. (2019). A node X is conditionally independent of any other unconnected node Y given its Markov blanket i.e. its parents, children and co-parents of children. So given Z consisting of the parents, children and co-parents of children of X, X is conditionally independent of any other node Y. Also, if a direct edge exists between X and Y, then given any Z, X is not conditionally independent of Y. We have used this philosophy to create 70 CI and 54 CD datasets.

Sachs et al. (2005) and Mooij and Heskes (2013) used a subset of 8 of the available 14 original flow cytometry datasets in their experiments to come up with Bayesian networks representing the underlying causal structure. We also used those 8 datasets in our experiments which had a combined total of around 7000 samples. The dimension of Z varies in the range 3 to 8.

Figure 5 compares the AuROC score of C-MI-GAN against the scores of CCMI and CCIT. C-MI-GAN retains its superior performance when compared against CCMI and CCIT. Surprisingly, CCIT outperforms CCMI contradicting the result presented by Mukherjee et al. (2019). This discrepancy might be due to limited capacity of their model architecture. We have created a larger dataset consisting of around 7000 samples. Whereas, the numbers reported by Mukherjee et al. (2019) are based on a smaller subset consisting of 853 data points. A larger network might improve the performance of CCMI.

Although Shah and Peters (2018) argues that domain knowledge is necessary to select an appropriate conditional independence test for a particular dataset, C-MI-GAN outperforms both CCIT and CCMI (as measured using AuROC) consistently across all the experiments. This might be because, in the min-max formulation, the regression network, R_ϕ and the generator network, G_θ are trained jointly using an adversarial scheme. As a result, the two networks together cancel the bias present in each other resulting in performance boost in high sample regime (a regime where GANs excel).

5 DISCUSSION AND CONCLUSION

In this work, we propose a novel CMI estimator, C-MI-GAN. This estimator is based on the formulation of

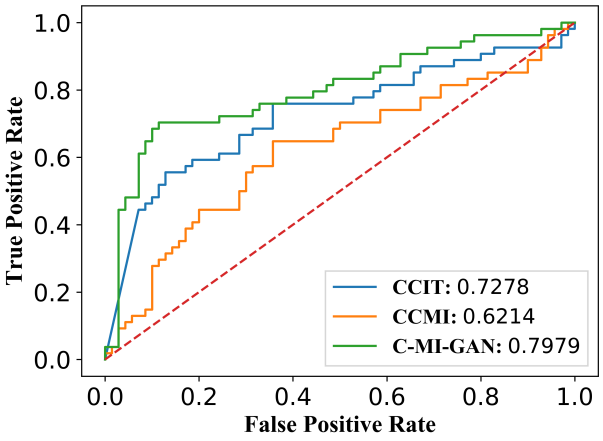


Figure 5: AuROC Curves: Flow-Cytometry Data-set. CCIT obtains a mean AuROC score of 0.728, CCMI obtains a mean of 0.62 while C-MI-GAN outperforms both of them with a mean AuROC score of 0.798 (Best viewed in color).

CMI as a min-max objective that can be optimized using joint training. We refrain from estimating two separate MI terms that could have unequal bias present. As opposed to separately training a conditional sampler and a divergence estimator, which may be sup-optimal, our joint training incorporates both steps into a single training procedure. We find that the estimator obtains improved estimates over a range of linear and non-linear datasets, across a wide range of dimension of the conditioning variable and sample size. Finally, we achieve performance boost in CI testing on simulated and real datasets using our improved estimator.

Acknowledgement

We thank IIT Delhi HPC facility[6] for computational resources. Sreeram Kannan is supported by NSF awards 1651236 and 1703403 and NIH grant 5R01HG008164. Himanshu Asnani acknowledges the support of Department of Atomic Energy, Government of India, under project no. 12-R&D-TFR-5.01-0500. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of the funding agencies.

References

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proc. of ICML*, 2018.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[6] *http://supercomputing.iitd.ac.in*

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 1975.

Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *Proc. of UAI*, 2014.

François Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, 2004.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Proc. of NeurIPS*, 2008.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Proc of AISTATS*, 2015.

Weihao Gao, Sewoong Oh, and Pramod Viswanath. Breaking the bandwidth barrier: Geometrical adaptive entropy estimation. In *Proc. of NeurIPS*, 2016.

Weihao Gao, Sewoong Oh, and Pramod Viswanath. Demystifying fixed k-nearest neighbor information estimators. *IEEE Transactions on Information Theory*, 2018.

Federico M Giorgi, Gonzalo Lopez, Jung H Woo, Brygida Bisikirska, Andrea Califano, and Mukesh Bansal. Inferring protein modulation from gene expression data using conditional mutual information. *PloS one*, 2014.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. of NeurIPS*, 2014.

Harry Joe. Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 84 (405), 1989.

LF Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2), 1987.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6), 2004.

Zhaohui Li, Gaoxiang Ouyang, Duan Li, and Xiaoli Li. Characterization of the causality between spike trains with permutation conditional mutual information. *Physical Review E*, 2011.

Dirk Loeckx, Pieter Slagmolen, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Nonrigid image registration using conditional mutual information. *IEEE transactions on medical imaging*, 2009.

Joris M. Mooij and Tom Heskes. Cyclic causal discovery from continuous equilibrium data. In *Proc. of UAI*, 2013.

Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. CCMI: Classifier based Conditional Mutual Information Estimation. In *Proc. of UAI*, 2019.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11), 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Proc. of NeurIPS*, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Barnabás Póczos and Jeff G Schneider. Nonparametric estimation of conditional information and divergences. In *Proc. of AISTATS*, 2012.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proc. of ICML*, 2019.

Alfréd Rényi. On measures of dependence. *Acta mathematica hungarica*, 10(3-4), 1959.

Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *Proc. of AISTATS*, 2018.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11), 2019.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721), 2005.

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G. Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Proc. of NeurIPS*, 2017.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *arXiv preprint arXiv:1804.07203*, 2018.

Alexander Shishkin, Anastasia Bezzubtseva, Alexey Drutsa, Ilia Shishkov, Ekaterina Gladkikh, Gleb Gusev, and Pavel Serdyukov. Efficient high-order interaction-aware feature selection based on conditional mutual information. In *Proc. of NeurIPS*, 2016.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Proc. of NeurIPS*, 2015.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *Proc. of ICLR*, 2020.

Saverio De Vito, Ettore Massera, Marco Piga, Luca Martinotto, and Girolamo Di Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2), 2008.

Gang Wang and Frederick H Lochovsky. Feature selection with conditional mutual information maximin in text categorization. In *Proc. of ACM CIKM*, 2004.

Yang Yang and Rick S Blum. Mimo radar waveform design based on mutual information and minimum mean-square error estimation. *IEEE Transactions on Aerospace and Electronic Systems*, 43(1), 2007.

Xiujun Zhang, Xing-Ming Zhao, Kun He, Le Lu, Yongwei Cao, Jingdong Liu, Jin-Kao Hao, Zhi-Ping Liu, and Luonan Chen. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, 28(1), 2012.