Causal Discovery with Flow-based Conditional Density Estimation

Shaogang Ren, Haiyan Yin, Mingming Sun, Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, Washington 98004, USA
No.10 Xibeiwang East Road, Beijing 100193, China
{shaogangren, haiyanyin, sunmingming01, liping11}@baidu.com

***Abstract*—Cause-effect discovery plays an essential role in many disciplines of science and real-world applications. In this paper, we introduce a new causal discovery method to solve the classic problem of inferring the causal direction under a bivariate setting. In particular, our proposed method first leverages a flow model to estimate the joint probability density of the variables. Then we formulate a novel evaluation metric to infer the scores for each potential causal direction based on the variance of the conditional density estimation. By leveraging the flow-based conditional density estimation metric, our causal discovery approach alleviates the restrictive assumptions made by the conventional methods, such as assuming the linearity relationship between the two variables. Therefore, it could potentially be able to better capture the complex causal relationship among data in various problem domains that comes in arbitrary forms. We conduct extensive evaluations to compare our method with decent causal discovery approaches. Empirical results show that our method could promisingly outperform the baseline methods with noticeable margins on both synthetic and real-world datasets.**

I. Introduction

The study of causal inference aims to identify the complex traits underlying the system via predicting the effect of taking actions on the system [17]. Many contemporary causal inference methods focus on identifying causal relationship from observatory data only, where such task is also referred to as causal discovery [3, 16]. There are two main categories of methods have been developed to tackle the causal discovery problem, constraint-based and score-based approaches . The constraint-based approaches perform statistical conditional independence tests to determine the existence of a shared edge between each pair of variables and output a DAG that entails all Markov equivalent classes [17, 19]. The score-based methods, on the contrary, adopt carefully specified score functions to evaluate the quality of candidate causal models and thus being able to generate such confidence indicator [2, 4, 8].

In this paper, we propose a method for bivariate causal discovery that exploits the asymmetry between the cause and effect variables in the expected variance of their conditional density estimations. We formulate a novel score function constructed upon estimating the expected variance of conditional density at each potential causal direction. Our proposed method is mostly related to the previous approaches [2] which exploit the asymmetry in the from of mean-squared error (MSE) derived from regression tasks of predicting cause from effect and effect from cause, respectively. An apparent difference between our method and the aforementioned works is that our method intends to use conditional data density values, rather than considering the regression errors. Using variance of conditional density values allows us to make less and weaker assumptions about the causal model. Meanwhile, the proposed method can be naturally integrated with a broad family of density estimation methods such as normalizing flows [5, 6, 12] and kernel based density estimation [14, 15].

Overall, the contributions of this paper are three-fold.

- We formulate a principled score function for bivariate causal discovery with the causal directions evaluated by the expected variance of their conditional density values. Our method extends causal discovery to density estimation-based methods which effectively reinforces the stability and robustness of the model performance.
- We provide theoretical proof that justifies our proposed method under clear and weak formal assumptions.
- We present evaluation results that validate the reliability of our proposed density estimation-based score function. Our method could outperform several decent causal discovery baselines in both synthetic and real-world datasets.

II. Preliminaries

We present the preliminary notations for the flow-based model and the task of bivariate causal discovery in this section.

A. Flow-based Model

Compare with other deep generative models, one appealing property for the flow model is that it could perform exact likelihood inference [5, 6, 12]. In our work, we leverage such property of a normalizing flow model [6] to accomplish the task of estimating the conditional density of the variables for causal discovery. With the flow model $\mathbf{f}$, the density of sample $\mathbf{x}$, $p(\mathbf{x})$ could be straightforwardly computed as

$$\log p(\mathbf{x}) = \log p(\mathbf{z}) + \sum_{i}^{n} \log \left| \det \left(\frac{\partial \mathbf{f}_{\theta_i}}{\partial \mathbf{h}_{i-1}} \right) \right|, \qquad (1)$$

where $\frac{\partial \mathbf{f}_{\theta_i}}{\partial \mathbf{h}_{i-1}}$ is the Jacobian of the mapping function $\mathbf{f}_{\theta_i}$ at the i-th layer, and $\mathbf{z} = \mathbf{f}(\mathbf{x})$.

B. Bivariate Causal Discovery

Formally, given a pair of random variables X and Y, the task is to identify whether the causal DAG should take the direction of $X \to Y$ or $X \leftarrow Y$. By slightly abusing terminology, we will not further distinguish between a distribution and its density since the Lebesgue measure as a reference is implicitly understood. The notations $p(X)$, $p(Y)$, and $p(Y|X)$ are used for denoting the corresponding marginal and conditional densities. Alternatively, we also use p_X, $p_{X,Y}$, and $p_{Y|X}$ to represent $p(X)$, $p(X,Y)$, and $p(Y|X)$, respectively. With a score-based approach, the bivariate causal discovery evaluates the score over each causal direction [9, 11], i.e., $X \to Y$ or $Y \to X$, and compare the scores as follows, $R = L_{X\to Y} - L_{Y\to X}$, where $L_{X\to Y}$ represents the score for evaluating the relation when X is assumed to be the *cause* variable and Y is the *effect* variable, and vice versa.

Among the various ways to derive the score functions, one intimate approach to ours is to infer it from the variance over the regression error [2]. Specifically, the method takes the assumption that the variance of regression error would be lower at the correct causal direction, where the conditional variance of the regression error for predicting Y given $X = x$ could be evaluated as follows:

$$\text{Var}[Y|x] := \mathbb{E}\big[\big(Y - \mathbb{E}[Y|x]\big)^2 \big| x\big]. \tag{2}$$

Thus the score function could be evaluated as the following expectation: $L_{X\to Y} = \mathbb{E}\big[\text{Var}[Y|X]\big] := \int \text{Var}[Y|x] p_X(dx)$.

In our paper, we propose a method to evaluate the conditional density without relying on the linear assumption over the relationship between the variables or a known density function. We leverage a flow-based model to effectively perform the conditional density estimation.

III. Proposed Method

Let $p(Y|x) = p(Y|X = x)$ and $\mathbb{E}_{p(Y|x)}[f(Y)] = \mathbb{E}_{y\sim p(Y|x)}[f(y)]$. We present a score function which is defined upon the estimation over the expected variance of the conditional density values as follows.

Definition 1: The **v**ariance of **c**onditional **d**ensity values (Vcd) for a variable Y (the potential effect variable) given $X = x$ (the potential cause variable) is defined as:

$$\text{Vcd}\,[Y|x] := \mathbb{E}_{p(Y|x)}\big[\big(p(Y|x) - \mathbb{E}_{p(Y|x)}\,[p(Y|x)]\big)^2\big]. \tag{3}$$

Note that the term $\mathbb{E}_{p(Y)}\,[p(Y|x)]$ in definition (3) is a constant. Thus the emphasis of the method would be to devise an efficient estimation for the conditional density $p(Y|x)$. Also note that compared to our closest counterpart [2], our method constructs the score function based on the conditional density estimate without using a regression task.

Given $X = x$, $\text{Vcd}[Y|x]$ is evaluated to be a constant value. Thus when different values for the random variable $X \sim p(X)$ is substituted to the equation, $\text{Vcd}[Y|X]$ becomes a variable. To evaluate the score function for causal discovery, we define the following score function, where the score for the causal direction $X \to Y$ is evaluated as the expectation of the variable $\text{Vcd}[Y|X]$, as follows.

Definition 2:

$$\mathbb{E}\big[\text{Vcd}[Y|X]\big] := \mathbb{E}_{x\sim p(X)}\big[\text{Vcd}[Y|x]\big]. \tag{4}$$

In following sections, we also use EVCD to denote the expectation of Vcd values. Our method assumes that the expected variance of the conditional likelihood would be smaller at the correct causal direction. With joint probability density values of two variables, we will show that it is easy to evaluate both Vcd and EVCD and thus the causal direction estimation.

A. Assumptions

To study the limit of an almost deterministic relation in a mathematically precise way, we formulate the *effect* variables E_α in the following manner:

$$E = \phi(C) + \alpha N, \tag{5}$$

where N is a noise distribution, $\phi(\cdot)$ is a function mapping and $\alpha \in \mathbb{R}^+$ is a constant to control the noise level. With the formulation for the causal function, we further adopt the following assumptions:

1) The distributions of C and N have compact support.
2) $\phi : [0,1] \to [0,1]$ is an invertible monotonic function which is two times differentiable, and $\phi(0) = 0$ and $\phi(1) = 1$ (functions with $\phi(1) = 0$ and $\phi(0) = 1$ have similar results).
3) Let $p(C)$ be the probability density function of c on interval $[0,1]$. We assume $\phi'(c)$ and $\text{Vcd}[N|c]p(c)$ are uncorrelated with each other. This condition can also be written as follows:

$$\text{Cov}\Big[\phi'(c), \text{Vcd}\big[N|c\big]p(c)\Big] = 0. \tag{6}$$

Assumption 3) takes both $\phi'(c)$ and $\text{Vcd}[N|c]p(c)$ as random variables, and a simple implication of (6) reads:

$$\begin{aligned}\int_0^1 \phi'(c)\text{Vcd}[N|c]p(c)dc &= \int_0^1 \phi'(c)dc \int_0^1 \text{Vcd}[N|c]p(c)dc \\ &= \int_0^1 \text{Vcd}[N|c]p(c)dc. \end{aligned} \tag{7}$$

Here $\int_0^1 \phi'(c)dc = 1$ under assumption 2). Compared to other methods [3, 8], assumption 3) is a relatively weak statement on the noise N and function ϕ. Furthermore, we have only three model assumptions, which implies that compared to [2], our method could apply to a greater range of noise and causal function categories. The *Vcd* values in our method can be computed based on the density values derived from the normalizing flow model or other density estimation approaches. Furthermore, we have the following definition regarding the *effect* variable: $E_\alpha = \frac{E-\alpha n_-}{1+\alpha n_+ - \alpha n_-}$, where $n_- < 0 < n_+$.

B. Theory

We start from a flow-based model $\mathbf{f}$ with a bivariate setting, i.e., the data is given by $G = [X, Y]$, with X and Y being the two variables of interest. We are interested in inferring the joint likelihood of X and Y. Let Z denote the corresponding latent space for the two variables. With the flow model $\mathbf{f}$, we could evaluate the joint probability density by the function,

$$p(g) = p(z) \cdot \left| \det\left(\frac{\partial z}{\partial g}\right) \right| = p_Z(\mathbf{f}(g)) \cdot \left| \det\left(\frac{\partial \mathbf{f}}{\partial g}\right) \right|,$$

where $g = [x, y]$ and z are concrete realizations of variable $G = [X, Y]$ and the latent variable Z, respectively. Then we can compute the marginal likelihood value of each variable by adding up the learned joint probability density as follows,

$$p_X(x) = \int_Y p_{X,Y}(x, y)\mathrm{d}y \approx \sum_{i=1}^{N} p_{X,Y}(x, y^{(i)})\delta_y. \quad (8)$$

Here $\delta_y = y^{(i+1)} - y^{(i)}$ $(\forall i, 0 \leq i \leq N)$ is a small value, and N is the number of samples evenly sampled on variable Y's range $\mathcal{Y}$. Note that a smaller δ_y gives a more accurate approximation of $P_X(x)$. Thus, the conditional likelihood estimation for both causal directions could be computed in the following forms,

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)} \approx \frac{p_{X,Y}(x, y)}{\sum_{i=1}^{N} p_{X,Y}(x, y^{(i)})\delta_y},$$
$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \approx \frac{p_{X,Y}(x, y)}{\sum_{i=1}^{N} p_{X,Y}(x^{(i)}, y)\delta_x}.$$

We have the following important lemma giving a limit regarding variable pair $\{C, E_\alpha\}$ defined by (5).

Lemma 1: Let the assumptions 1-3 hold, then the following limit holds:

$$\lim_{\alpha \to 0} \frac{\mathbb{E}\big[\mathrm{Vcd}[C|E_\alpha]\big]}{\mathbb{E}\big[\mathrm{Vcd}[E_\alpha|C]\big]} = \frac{\int_0^1 \phi'(c)^2 \mathrm{Vcd}[N|c]p(c)dc}{\int_0^1 \mathrm{Vcd}[N|c]p(c)dc}. \quad (9)$$

The proof of Lemma 1 can be found in Appendix. Based on Lemma 1, we have the following theorem (also proved in the Appendix) regarding a pair of cause-effect variables.

Theorem 1: For a pair of causal and effect variables $\{C, E\}$, the following limit holds $\lim_{\alpha \to 0} \frac{\mathbb{E}\left[\mathrm{Vcd}[C|E_\alpha]\right]}{\mathbb{E}\left[\mathrm{Vcd}[E_\alpha|C]\right]} \geq 1$, if the assumptions 1)-3) hold.

It shows that the *Vcd*-based score function could identify the correct causal direction. Meanwhile, Theorem 1 could generalize our proposed causal discovery approach to a broad family of methods where different density estimation models such as normalizing flows [5, 6, 12] and kernel based density estimation [14, 15] could be employed for density estimation. This sheds light that our proposed density estimation-based approach could be promising to improve the stability and accuracy of causal discovery.

Algorithm 1 Causal Discovery Algorithm

Input:A flow model $\mathbf{f}$; data vectors X and Y
Output: Causal direction for the bivariate case

Training $\mathbf{f}$ with data sample pairs [X, Y];
$\mathcal{T}_{Y|X} \leftarrow \mathrm{EVCD}(\mathbf{f}, X, Y)$;
$\mathcal{T}_{X|Y} \leftarrow \mathrm{EVCD}(\mathbf{f}, Y, X)$;
if $\mathcal{T}_{Y|X} < \mathcal{T}_{X|Y}$ **then**
  return X causes Y;
else if $\mathcal{T}_{Y|X} > \mathcal{T}_{X|Y}$ **then**
  return Y causes X;
else
  return No decision;
end if

Algorithm 2 Computing EVCD

Input:A flow model $\mathbf{f}$; data pairs $[A, B]$; δ_a, δ_b;
Output: EVCD$(B|A)$

Let $\mathcal{A}$ be the range of A, and $\mathcal{B}$ be the range of B;
$\widehat{A} \leftarrow$ Sample N_A points on $\mathcal{A}$ with intervals δ_a;
$\widehat{B} \leftarrow$ Sample N_B points on $\mathcal{B}$ with intervals δ_b;
$\mathcal{V} \leftarrow []$;
for $a \in \widehat{A}$ **do**
  $p(a) \leftarrow \sum_{b \in \widehat{B}} p(a, b)\delta_b$;
  $p(\frac{b}{a}) \leftarrow \frac{p(a,b)}{p(a)}$, $\forall b \in \widehat{B}$;
  $\mathcal{V}(a) \leftarrow \sum_{b \in \widehat{B}} \big(p(\frac{b}{a}) - \sum_{b \in \widehat{B}} p(\frac{b}{a})^2 \delta_b\big)^2 p(\frac{b}{a})\delta_b$;
end for
EVCD$(B|A) \leftarrow \sum_{a \in \widehat{A}} \mathcal{V}(a)p(a)\delta_a$;

C. Flow-based Causal Discovery Algorithm

Our proposed method adopts the expected variance over the conditional density (EVCD) as the score function to detect the correct causal direction. We present the detailed algorithm for our method in Algorithm 1. The steps to compute EVCD are presented in Algorithm 2, where we train a normalizing flow model with the observatory data $[X, Y]$ to infer the joint likelihood $p(a, b)$ (line 3 in Algorithm 2) following Eq. (1).

For a pair of cause-effect variables $\{C, E\}$, the flow model $\mathbf{f}$'s dimension $\dim(\mathbf{f}) = 2$. In practice, to enhance the model's performance by leveraging the expressive power of neural networks, we could set the dimension of the flow model $\dim(\mathbf{f}) > 2$ and take multiple samples drawn i.i.d. from the joint distribution of $\{C, E\}$ as $\mathbf{f}$'s input. By taking all Cs in the input as one variable and all Es as another one, one can verify that the theory we develop still applies to this case.

D. General Remarks

Though both methods develop the score functions based on variance estimation, there are significant differences between our method and RECI [2]. As discussed in section III-A, RECI relies on regression errors to determine the causal direction. Our method, on the other hand, utilizes the asymmetry conditional probability between causal and effect to discover the causal relation between two variables.

Our method is also quite different from the recent causal autoregressive flow models [11] which uses existing likelihood ratio theory for causal discovery. The bivariate causal discovery method proposed in [11] relies on the autoregressive

TABLE I
PERFORMANCE OF DIFFERENT CAUSAL DISCOVERY METHODS EVALUATED ON SIX BIVARIATE DATASETS. SPECIFICALLY, FIVE OF THEM ARE SYNTHETIC DATASETS AND THE OTHER ONE IS A REAL-WORD DATASET.

	Synthetic					Real-world	Average Scores
dataset	dream4-1	dream4-2	dream4-3	dream4-4	dream4-5	tuebingen	
ANM	0.443	0.522	0.538	0.564	0.539	0.520	0.522
Bivariate Fit	0.477	0.458	0.523	0.516	0.518	0.556	0.508
CDS	0.500	0.542	0.538	0.564	0.456	0.677	0.546
IGCI	0.591	0.546	0.502	0.545	0.523	0.616	0.554
RECI	0.346	0.466	0.497	0.336	0.383	0.636	0.444
EVCD (Ours)	**0.686**	**0.594**	**0.590**	**0.668**	**0.694**	**0.687**	**0.653**

structure of flow models in addition to the asymmetry of log-likelihood ratios regarding different orders of causal and effect variables. The dependence on neural network structures may introduce extra computation cost resulted from multiple times of model training for different possible causal directions. Obviously, our method does not rely on the flow neural net structures. As long as the data density values are available, our EVCD method is capable to perform causal discovery according to our developed theory.

Apart from this, our work also has the following three appealing properties: (i) our proposed conditional density estimation could generally be applied to any type of variables, but regression error is usually applied to continuous variables; (ii) employing conditional density values can alleviate the distribution and linearity assumption over the variables, and thereby enables our method to use effective model to capture complex causal relationship in many real-life datasets. Our method can be potentially further improved with cutting-edge deep neural network based density estimation models (iii) The proposed method can be easily plugged in other density estimation methods, such as kernel based methods. It enlarges the flexibility of causal detection and can be extended to many application scenarios. Since we proposed to compute the EVCD scores based on the joint likelihood (see Algorithm 2), we only need to train the flow model once for evaluating both causal directions, as the joint likelihood remains the same for both causal directions.

IV. EXPERIMENTS

We compare the proposed method against multiple baseline algorithms including RECI [2], ANM [8], Bivariate Fit [10], CDS [7], and IGCI [4].

A. Implementation

For the baselines methods ANM, Bivariate-Fit, CDS, IGCI, and RECI, we adopt the standard implementation from the Causal Discovery Toolbox (CDT) [10], which is an open-source package implemented in Python. There is no model hyper-parameters for the methods ANM, Bivariate-Fit and CDS. For IGCI, we adopt Gaussian as the reference measure, and entropy as the estimator. The parameters for IGCI are set as the default values specified by CDT. For RECI, the degree of polynomial feature for regression is chosen from $d = \{3, 4, 5, 6, 9, 10\}$, and the best result is reported for each dataset.

Our method (EVCD) is implemented with the PaddlePaddle deep learning platform [1]. The parameters of EVCD are coarsely selected with the ranges as learning rate $lr = \{0.001, 0.0001\}$, number of coupling blocks for the flow model $= \{2, 4, 6\}$, $\delta_x = \delta_y = 0.001$, and training epochs $= \{800, 1500, 2000\}$ for different datasets. We employ the coupling block [6] for each layer of our flow model, $\mathbf{f}$. Each coupling bock consists of three fully connect layers (hidden dimension 64) separated by two ReLU layers. The dimension of the flow model is 10, and we take 5 data samples from a cause-effect pair as the input of $\mathbf{f}$. All datasets are normalized and re-scaled to [0,1] for both EVCD and RECI methods.

B. Evaluation on Dream and Real-World Datasets

For evaluation, we consider both synthetic and real-world data sets. We evaluate each method in terms of the following evaluation metric, Accuracy $= \frac{\text{\# of correct cause-effect directions}}{\text{total \# of cause-effect pairs}}$. The synthetic datasets adopted in the experiment are the Dream datasets proposed in [18]. Specifically, Dream is a five-gene-network dataset that contains two time series corresponding to two different treatments. The data is generated by a synthetic generative network which connects 5 synthetic genes oscillating with the cell cycle. To create the bivariate setting, we consider five pair-wise connections to construct the causal discovery task.

The Tuebingen dataset is collected by the Max-Plank-Institute for Intelligent Systems at Tubingen [13]. It is an extension of a collection of datasets from a competition held in a causality workshop in 2008. The Cause Effect Pairs (CEP) collection is continuously being updated and upon writing this paper, there are in total 99 examples. The number of samples for each pair varies from 94 to 16382. The CEP collection is obtained from a variety of domains, including abalone measurements, census income, fuel consumption, and so on. We also present the statistics of this dataset in Appendix.

1) Evaluation Results: We present the evaluation results for the synthetic datasets and the real-world dataset in Table I. Note that for the baseline approaches, we adopt the standardized implementation from the Causal Discovery Toolbox [10]. Overall, our proposed method leads to promising causal discovery performance consistently across all the testified datasets. Notably, it could outperform all the compared base-

line methods in all the synthetic and real-world task domains. For most of the baseline approaches (except for IGCI), we notice that the baseline approaches are not stable and inferior performance is observed in one or more task domains, which is a result of either the deficiency over the assumption for causal asymmetry between the variables (i.e., their adopted heuristics) or the restrictive assumption over the relationship between the cause and effect variables (e.g., linearity assumption). On the contrary, our proposed density estimation approach turns out to be more effective in handling complex causal relationship among different task domains, and thereby being able to tackle a broader range of complex problems. Meanwhile, on certain task domains, our method could outperform the state-of-the-art causal discovery approaches with a large margin, e.g., on *dream4-4&5* dataset, the performance gains turn out to be 10.4% and 15.5%. Overall, our method could outperform the best baseline approach in terms of the *average scores* across all the datasets with a margin of 9.9%. Lastly, we succinctly highlight that when comparing with our most intermediately related baseline RECI, our proposed method demonstrates apparent advantage over the baseline method by consistently outperforming RECI with significant margins. This reveals that our proposed conditional density estimation approach is more effective in serving as a score function to accomplish the causal discovery task under a bivariate setting.

V. Conclusion

We present a novel method for solving causal discovery problems under a bivariate setting. A score function based on the expected variance of the conditional density is proposed, where a normalizing flow model is leveraged to facilitate the density estimation task. We compare our proposed method with various causal discovery approaches, where the results demonstrate the consistency, robustness, and effectiveness for our method on both synthetic and real-world bivariate datasets. For future work, it is promising to consider our method for real-world causal discovery problems with a multi-variate setting.

Appendix I. Dataset Statistics

TABLE II
STATISTICS FOR THE SYNTHETIC DATASETS AND THE REAL-WORLD DATASET USED FOR EVALUATING THE MODELS.

	dataset	Num. Pairs	Num. Samples
Synthetic	dream4-1	176	100
	dream4-2	249	100
	dream4-3	195	100
	dream4-4	211	100
	dream4-5	193	100
Real-world	tuebingen	99	2034 (mean)

Appendix II. Proofs of Lemma 1 and Theorem 1

A. Proof of Lemma 1

According to the definition of conditional density variance, with cause effect pair $\{C, E\}$, $\mathrm{Vcd}\big[E|c\big] = \mathbb{E}_{p(E|c)}\big[(p(E|c) - \mathbb{E}_{p(E|c)}[p(E|c)])^2\big]$. While $\mathrm{Vcd}[E|C]$ is the random variable attaining the value $\mathrm{Vcd}[E|c]$ when C attains the value c. Its expectation reads $\mathbb{E}\big[\mathrm{Vcd}[E|C]\big] = \int \mathrm{Vcd}[E|c]p(c)dc$. According to the cause effect relation, $E = \phi(C) + \alpha N$. Variables N and C may not be independent, and the noise maybe affected by C, i.e. $E(c) = \phi(c) + \alpha N(c),\ N(c) = \frac{1}{\alpha}\big(E(c) - \phi(c)\big)$. We have the conditional density value of E given a fixed c as $p(E|c) = p\big(N(c)\big)\left|\frac{dN(c)}{dE}\right| = p\big(N(c)\big)\frac{1}{\alpha} = p\big(N|c\big)\frac{1}{\alpha}$. Let $\mathbf{E}$ and $\mathbf{N}$ be the variable scope of conditional density functions $p(E|c)$ and $p(N|c)$, respectively. Then

$$\begin{aligned}
\mathrm{Vcd}\big[E|c\big] =& \mathbb{E}_{p(E|c)}\big[(p(E|c) - \mathbb{E}_{p(E|c)}[p(E|c)])^2\big] \\
=& \int_{\mathbf{E}} \left(p(e|c) - \int_{\mathbf{E}} p(e|c)^2 de\right)^2 p(e|c)de \\
=& \int_{\mathbf{N}} \left(\frac{1}{\alpha}p\big(n(c)\big) - \frac{1}{\alpha}\int_{\mathbf{N}} p(n(c))^2 dn(c)\right)^2 p(n(c))dn(c) \\
=& \frac{1}{\alpha^2}\int_{\mathbf{N}} \left(p(n|c) - \int_{\mathbf{N}} p(n|c)^2 dn(c)\right)^2 p(n|c)dn(c) \\
=& \frac{1}{\alpha^2}\mathrm{Vcd}\big[N|c\big] \qquad (10)
\end{aligned}$$

According to the definition in III-A, it is easy to verify that

$$\lim_{\alpha\to 0}\frac{\mathbb{E}\big[\mathrm{Vcd}[C|E_\alpha]\big]}{\mathbb{E}\big[\mathrm{Vcd}[E_\alpha|C]\big]} = \lim_{\alpha\to 0}\frac{\mathbb{E}\big[\mathrm{Vcd}[C|E]\big]}{\mathbb{E}\big[\mathrm{Vcd}[E|C]\big]} = \lim_{\alpha\to 0}\frac{\alpha^2\mathbb{E}\big[\mathrm{Vcd}[C|E]\big]}{\mathbb{E}\big[\mathrm{Vcd}[N|C]\big]}, \qquad (11)$$

where $E = \phi(C) + \alpha N$. Then, the denominator in RHS of (11) is not affected by α. The numerator in RHS of (11)

$$\begin{aligned}
&\lim_{\alpha\to 0}\alpha^2\mathbb{E}\big[\mathrm{Vcd}[C|E]\big] \\
&= \lim_{\alpha\to 0}\int_{\phi(0)+\alpha n^-}^{\phi(1)+\alpha n^+} \mathbb{E}_{p(C|e)}\big[(\alpha p(C|e) - \alpha\mathbb{E}_{p(C|e)}[p(C|e)])^2\big]p_E(e)de \\
&= \lim_{\alpha\to 0}\int_{\phi(0)}^{\phi(1)} \mathbb{E}_{p(C|e)}\big[(\alpha p(C|e) - \alpha\mathbb{E}_{p(C|e)}[p(C|e)])^2\big]p_E(e)de \qquad (12)
\end{aligned}$$

The later term αn^+ and αn^- vanishes due to the function $\mathbb{E}_{p(C|e)}\big[(\alpha p(C|e) - \alpha\mathbb{E}_{p(C|e)}[p(C|e)])^2\big]p_E(e)$ is uniformly bounded in α. $p_E(e)$ is uniformly bounded due to

$$p_E(e) = \int_{n_-}^{n_+} p_{\phi(C),N}(e - \alpha n, n)dn \le ||p_{C,N}||_\infty ||\phi^{-1'}||_\infty (n_+ - n_-).$$

According to the bounded convergence theorem,

$$\begin{aligned}
&\lim_{\alpha\to 0}\int_{\phi(0)}^{\phi(1)} \mathbb{E}_{p(C|e)}\big[(\alpha p(C|e) - \alpha\mathbb{E}_{p(C|e)}[p(C|e)])^2\big]p_E(e)de \qquad (13) \\
&= \int_{\phi(0)}^{\phi(1)} \lim_{\alpha\to 0}\mathbb{E}_{p(C|e)}\big[(\alpha p(C|e) - \alpha\mathbb{E}_{p(C|e)}[p(C|e)])^2\big]p_E(e)de.
\end{aligned}$$

Now we focus on the limit $\lim_{\alpha\to 0}\mathbb{E}_{p(C|e)}\big[(\alpha p(C|e) - \alpha\mathbb{E}_{p(C|e)}[p(C|e)])^2\big]$. Since ϕ is assumed to be invertible, $c = \phi^{-1}(e - \alpha n) = \phi^{-1}(e) - \alpha n\phi^{-1'}(e) - \frac{\alpha^2 n^2 \phi^{-1''}(\hat{e}(n,e))}{2}$. Here $\hat{e}(n,e)$ is a value in $[e - \alpha n, e]$. With a fixed e, $dc = \big(-\alpha\phi^{-1'}(e) - \alpha^2\phi^{-1''}(\hat{e}(n,e))n\big)dn$. Thus, given a fixed $E = e$ we get $p(C|e) = p\big(N(e)\big)\frac{1}{|\alpha\phi^{-1'}(e) + \alpha^2\phi^{-1''}(\hat{e}(N(e),e))N(e)|}$. Let $\tau(\alpha) = |\alpha\phi^{-1'}(e) + \alpha^2\phi^{-1''}(\hat{e}(N(e),e))N(e)|$, then $p(C|e) = p\big(N(e)\big)\frac{1}{\tau(\alpha)}$.

The computation of the limit can be carried as follows

$$
\begin{aligned}
&\lim_{\alpha\to 0}\mathbb{E}_{p(C|e)}\big[(\alpha p(C|e)-\alpha\mathbb{E}_{p(C|e)}[p(C|e)])^2\big]\\
=&\lim_{\alpha\to 0}\int_{\mathbf{C}}\Big(\alpha p(c|e)-\int_{\mathbf{C}}\alpha p(c|e)^2dc\Big)^2 p(c|e)dc\\
=&\lim_{\alpha\to 0}\int_{\mathbf{N}}\Big(\alpha p(c|e)-\int_{\mathbf{N}}\alpha p(c|e)p\big(n(e)\big)dn(e)\Big)^2 p\big(n(e)\big)dn(e)\\
=&\int_{\mathbf{N}}\lim_{\alpha\to 0}\Big(p(n(e))\frac{\alpha}{\tau(\alpha)}-\int_{\mathbf{N}}p(n(e))\frac{\alpha}{\tau(\alpha)}p\big(n(e)\big)\\
&\qquad dn(e)\Big)^2 p\big(n(e)\big)dn(e)
\end{aligned}
$$

$$
=\int_{\mathbf{N}}\Big(p(n(e))\frac{1}{|\phi^{-1'}(e)|}-\frac{1}{|\phi^{-1'}(e)|}\int_{\mathbf{N}}p(n(e))p\big(n(e)\big)dn(e)\Big)^2 p\big(n(e)\big)dn(e) \tag{14}
$$

$$
=\frac{1}{\phi^{-1'}(e)^2}\int_{\mathbf{N}}\Big(p(n(e))-\int_{\mathbf{N}}p(n(e))^2dn(e)\Big)^2 p\big(n(e)\big)dn(e) \tag{15}
$$

Here $\mathbf{C}$ is the scope of $p(C|e)$. With a fixed e, here $p(n|e)=p(n(e))$, then (15) becomes

$$
\begin{aligned}
&\lim_{\alpha\to 0}\mathbb{E}_{p(C|e)}\big[(\alpha p(C|e)-\alpha\mathbb{E}_{p(C|e)}[p(C|e)])^2\big] \qquad (16)\\
=&\frac{1}{\phi^{-1'}(e)^2}\int_{\mathbf{N}}\Big(p(n|e)-\int_{\mathbf{N}}p(n|e)^2dn(e)\Big)^2 p\big(n|e\big)dn(e)\\
=&\frac{1}{\phi^{-1'}(e)^2}Vcd[N|e].
\end{aligned}
$$

With the limit, here we have $e=\phi(c)$. With ϕ is invertible,

$$
\mathrm{Vcd}\big[N|e\big]=\mathrm{Vcd}\big[N|\phi(c)\big]=\mathrm{Vcd}\big[N|\phi(c),C=c\big]=\mathrm{Vcd}\big[N|c\big]. \tag{17}
$$

With (11-13), (16), (17), and $p_E(e)de=p_C(c)dc$, the limit yields

$$
\begin{aligned}
&\lim_{\alpha\to 0}\frac{\mathbb{E}\big[\mathrm{Vcd}[C|E_\alpha]\big]}{\mathbb{E}\big[\mathrm{Vcd}[E_\alpha|C]\big]}=\lim_{\alpha\to 0}\frac{\alpha^2\mathbb{E}\big[\mathrm{Vcd}[C|E]\big]}{\mathbb{E}\big[\mathrm{Vcd}[N|C]\big]}\\
=&\frac{1}{\mathbb{E}\big[\mathrm{Vcd}[N|C]\big]}\lim_{\alpha\to 0}\int_{\phi(0)}^{\phi(1)}\mathbb{E}_{p(C|e)}\big[(\alpha p(C|e)-\\
&\qquad\alpha\mathbb{E}_{p(C|e)}[p(C|e)])^2\big]p_E(e)de\\
=&\frac{1}{\mathbb{E}\big[\mathrm{Vcd}[N|C]\big]}\int_{\phi(0)}^{\phi(1)}\frac{1}{\phi^{-1'}(e)^2}\mathrm{Vcd}[N|e]p_E(e)de\\
=&\frac{\int_0^1\phi'(c)^2\mathrm{Vcd}[N|c]p(c)dc}{\int_0^1\mathrm{Vcd}[N|c]p(c)dc}.
\end{aligned}
$$

□

B. Proof of Theorem 1

According to Lemma 1,

$$
\begin{aligned}
&\lim_{\alpha\to 0}\frac{\mathbb{E}\big[\mathrm{Vcd}[C|E_\alpha]\big]}{\mathbb{E}\big[\mathrm{Vcd}[E_\alpha|C]\big]}=\frac{\int_0^1\phi'(c)^2\mathrm{Vcd}[N|c]p(c)dc}{\int_0^1\mathrm{Vcd}[N|c]p(c)dc}\\
=&\frac{1}{\big(\int_0^1\mathrm{Vcd}[N|c]p(c)dc\big)^2}\int_0^1\phi'(c)^2\mathrm{Vcd}[N|c]p(c)dc\int_0^1\mathrm{Vcd}[N|c]p(c)dc\\
\geq&\frac{1}{\big(\int_0^1\mathrm{Vcd}[N|c]p(c)dc\big)^2}\Big(\int_0^1\sqrt{\phi'(c)^2\mathrm{Vcd}[N|c]}\sqrt{\mathrm{Vcd}[N|c]}p(c)dc\Big)^2 \qquad (18)\\
=&\frac{\big(\int_0^1\phi'(c)\mathrm{Vcd}[N|c]p(c)dc\big)^2}{\big(\int_0^1\mathrm{Vcd}[N|c]p(c)dc\big)^2}=1.
\end{aligned}
$$

The step in (18) is because of Cauchy-Schwartz Inequality. The last step is due to assumption 3 and equation (7). □

REFERENCES

[1] Paddlepaddle. http://www.paddlepaddle.org.cn/.

[2] Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Analysis of cause-effect inference by comparing regression errors. *PeerJ Comput. Sci.*, 5:e169, 2019.

[3] Joshua W Comley and David L Dowe. General bayesian networks and asymmetric languages. In *Proc. Hawaii International Conference on Statistics and Related Fields*, pages 5–8, 2003.

[4] Povilas Daniusis, Dominik Janzing, Joris M. Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 143–150, Catalina Island, CA, 2010.

[5] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: nonlinear independent components estimation. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

[6] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[7] José A. R. Fonollosa. Conditional distribution variability measures for causality detection. In Isabelle Guyon, Alexander R. Statnikov, and Berna Bakir Batu, editors, *Cause Effect Pairs in Machine Learning*, pages 339–347. Springer, 2019.

[8] Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 689–696, Vancouver, Canada, 2008.

[9] Aapo Hyvärinen and Stephen M. Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *J. Mach. Learn. Res.*, 14(1):111–152, 2013.

[10] Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. Causal discovery toolbox: Uncovering causal relationships in python. *Journal of Machine Learning Research*, 21:37:1–37:5, 2020.

[11] Ilyes Khemakhem, Ricardo Pio Monti, Robert Leech, and Aapo Hyvärinen. Causal autoregressive flows. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3520–3528, Virtual Event, 2021.

[12] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10236–10245, Montréal, Canada, 2018.

[13] Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *J. Mach. Learn. Res.*, 17:32:1–32:102, 2016.

[14] Hill Peter D. Kernel estimation of a distribution function. *Communications in Statistics-Theory and Methods*, 14(3):605–620, 1985.

[15] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[16] Michael E Sobel. An introduction to causal inference. *Sociological Methods & Research*, 24(3):353–379, 1996.

[17] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[18] Changwon Yoo and Erik M Brilz. Dream project: The five-gene-network data analysis with local causal discovery algorithm using causal bayesian networks. *Annals of the New York Academy of Sciences*, 1158:93, 2009.

[19] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.