Causal Discovery with Confounding Cascade Nonlinear Additive Noise Models

JIE QIAO, School of Computer Science, Guangdong University of Technology, China
RUICHU CAI, School of Computer Science, Guangdong University of Technology, China and Guangdong Provincial Key Laboratory of Public Finance and Taxation with Big Data Application, China
KUN ZHANG, Department of Philosophy, Carnegie Mellon University, USA
ZHENJIE ZHANG, PVoice Technology, Singapore
ZHIFENG HAO, College of Science, Shantou University, China

ACM Trans. Intell. Syst. Technol., Vol. 12, No. 6, Article 80, Publication date: November 2021.
DOI: https://doi.org/10.1145/3482879 (https://doi.org/10.1145/3482879)

Identification of causal direction between a causal-effect pair from observed data has recently attracted much attention. Various methods based on functional causal models have been proposed to solve this problem, by assuming the causal process satisfies some (structural) constraints and showing that the reverse direction violates such constraints. The nonlinear additive noise model has been demonstrated to be effective for this purpose, but the model class does not allow any confounding or intermediate variables between a cause pair–even if each direct causal relation follows this model. However, omitting the latent causal variables is frequently encountered in practice. After the omission, the model does not necessarily follow the model constraints. As a consequence, the nonlinear additive noise model may fail to correctly discover causal direction. In this work, we propose a confounding cascade nonlinear additive noise model to represent such causal influences–each direct causal relation follows the nonlinear additive noise model but we observe only the initial cause and final effect. We further propose a method to estimate the model, including the unmeasured confounding and intermediate variables, from data under the variational auto-encoder framework. Our theoretical results show that with our model, the causal direction is identifiable under suitable technical conditions on the data generation process. Simulation results illustrate the power of the proposed method in identifying indirect causal relations across various settings, and experimental results on real data suggest that the proposed model and method greatly extend the applicability of causal discovery based on functional causal models in nonlinear cases.

CCS Concepts: • Theory of computation → Machine learning theory;

Additional Key Words and Phrases: Causal discovery, additive noise model, latent model

ACM Reference format:
Jie Qiao, Ruichu Cai, Kun Zhang, Zhenjie Zhang, and Zhifeng Hao. 2021. Causal Discovery with Confounding Cascade Nonlinear Additive Noise Models. *ACM Trans. Intell. Syst. Technol.* 12, 6, Article 80 (November 2021), 28 pages. https://doi.org/10.1145/3482879 (https://doi.org/10.1145/3482879)

1 INTRODUCTION

Causal discovery is an important yet challenging task in various disciplines of science, including earth system sciences, [19] and biology [6]. It is well known that using randomized experiments to identify causal influences usually encounters unethical or substantial expense issues. Fortunately, inferring causal relations from pure observations, also known as *causal discovery from observational data*, has demonstrated its power in empirical studies and has been a focus in causality research.

Various methods have been proposed to infer the causal direction by exploring properly constrained forms of **functional causal models (FCMs)**. A functional causal model represents the effect Y as a function of its direct causes X and independent noise, i.e., $Y = f(X;\epsilon), X \perp \epsilon$. Without constraints on f, then for any two variables, one can always express one of them as a function of the other, and independent noise [26]. However, it is interesting to note that with properly constrained FCMs, the causal direction between X and Y is identifiable because the independence condition between the noise and cause holds only for the true causal direction and is violated for the wrong direction. Such FCMs include the **Linear, Non-Gaussian, Acyclic Model (LiNGAM)** [20], i.e., $Y = \mathbf{a}^\intercal X + \epsilon$ with linear coefficients $\mathbf{a}$, the nonlinear **Additive Noise Model (ANM)** [9], i.e., $Y = f(X) + \epsilon$, and the **Post-Nonlinear (PNL)** causal model [24], which also considers possible nonlinear sensor or measurement distortion f_2 in the causal process: $Y = f_2(f_1(X) + \epsilon)$. It has been shown that in the generic case, for data generated by the above FCMs, the reverse direction will not admit the same FCM class with independent noise. One can then find causal direction by estimating the FCM followed by testing for independence between the hypothetical cause and estimated noise [9, 24].

In reality, we can usually record only a subset of all variables which are causally related. If some variable is a direct cause of only one measured variable and is not measured, it is considered as part of the omitted factors or noise. If a hidden variable is a direct cause of two measured variables, it is a confounder. Causal discovery in the presence of confounders is challenging, although there exist some methods with asymptotic correctness guarantees, such as the FCI algorithm [21], RFCI algorithm [5], and the ICAM method [13]. In this paper, we are concerned with unmeasured intermediate and confounding causal variables. One typical example is the unmeasured intermediate causal variables, which is recently addressed by proposing **cascade nonlinear additive noise model (CANM)** in our conference version [3]. Suppose $X_1 \rightarrow X_2 \rightarrow X_3$, with X_2 unmeasured, and that each direct causal influence can be represented by an FCM in a certain class. If the direct causal relations are linear with additive noises, then the causal influence $X_1 \rightarrow X_3$ still follows a linear model with additive noises. However, if each direct causal influence follows the ANM, the causal influence $X_1 \rightarrow X_3$ does not necessarily follow the same model class. Figure 1 gives an illustration of this phenomenon of "non-transitivity of nonlinear causal model classes", in which $X_2 = 2\tanh(5X_1) + N_2$, and $X_3 = (X_2/2)^3 + N_3$, with X_1, N_2, and N_3 mutually independent and following the uniform distribution between -0.5 and 0.5. As seen from the heterogeneity of the noise in X_3 relative to X_1, given in Figure 1(c), the causal influence from X_1 to X_3 clearly does not admit a nonlinear model with additive noise. The PNL is more general than the additive noise model – in this example, if N_3 is zero, then $X_1 \rightarrow X_3$ will follow PNL. However, the PNL model class is also non-transitive.

In this work, we extend CANM to a more general confounding **cascade nonlinear additive noise model (CCANM)** by considering there are further unobserved confounding intermediate variables that exist. As shown in Figure 2, there is a confounder C that cause the X_1 and X_3, such that $X_1 = \sin(C) + N_1$ and $X_3 = (X_2/2)^3 + cos(C) + N_3$, then the mapping of the causal direction from X_1 to X_3 is non-trivial and clearly does not admit the nonlinear additive noise model. Hence, in the correct causal direction, from X_1 to X_3, the independent noise condition is violated. As a result, existing methods for causal direction determination by checking whether the regression residual is independent of the hypothetical cause may fail.

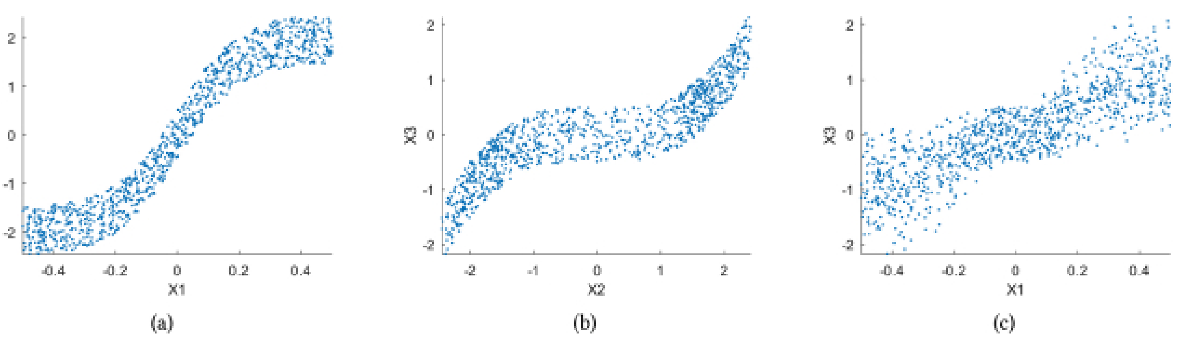


Fig. 1. Illustration of non-transitivity of nonlinear causal model classes, in which $X_1 \rightarrow X_2 \rightarrow X_3$ and each direct causal influence follows a nonlinear model with additive noise. Panels (a), (b), and (c) show the scatter plot of X_1 and $X_2 = 2\tanh(5X_1) + N_2$, that of X_2 and $X_3 = (X_2/2)^3 + N_3$, and that of X_1 and X_3, respectively.

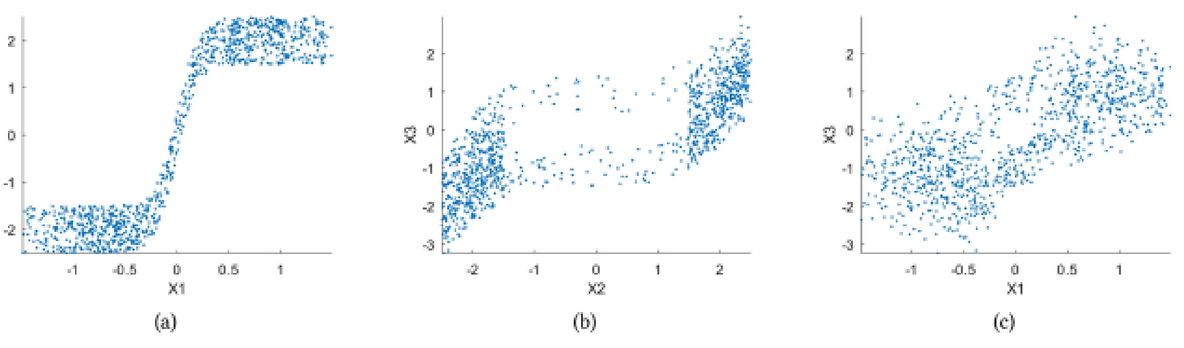


Fig. 2. Illustration of confounding of nonlinear causal model classes, in which $X_1 \rightarrow X_2 \rightarrow X_3$ with a confounding variable *C* such that $X_1 \leftarrow C \rightarrow X_3$. Panels (a), (b), and (c) show the scatter plot of $X_1 = \sin(C) + N_1$ and $X_2 = 2\tanh(5X_1) + N_2$, that of X_2 and $X_3 = (X_2/2)^3 + \cos(C) + N_3$, and that of X_1 and X_3, respectively.

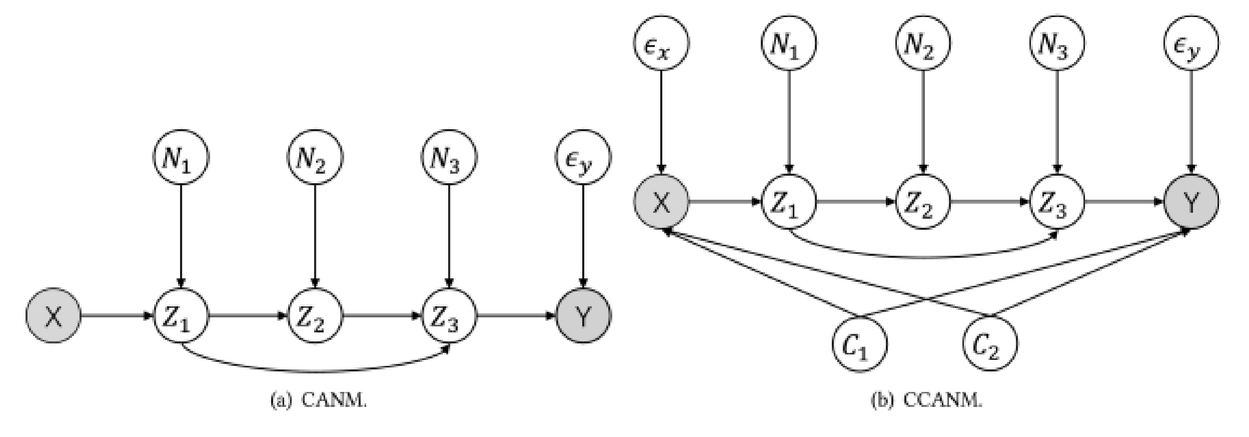


Fig. 3. Illustration of the CANM and CCANM. The causal chain from *X* to *Y* in CANM consists of three unmeasured intermediate variables Z_1, Z_2, Z_3 with their associated noises N_1, N_2, N_3, while in CCANM it further consists of the unmeasured confounding variables C_1, C_2.

This paper deals with such confounding, indirect, nonlinear causal relations, which seem to be ubiquitous in practice. Finding causal direction for such causal relations has recently been posed as an open problem [22]. In particular, we aim to find the causal direction between X and Y that are generated according to the process given in Figure 3, in which there might exist several unmeasured confounders and intermediate causal variables C_i and Z_i between X, Y. Moreover, each direct causal influence, e.g., the influences from Z_1 and Z_2 on Z_3, follows the ANM. We named the causal model from X and Y given in Figure 3(a) a **Cascade Additive Noise Model (CANM)** and the causal model from X to Y given in Figure 3(b) a **Confounding Cascade Additive Noise Model (CCANM)**. We note that the causal discovery in the presence of latent variables has been extensively investigated, including the FCI [21], RFCI [5], M3B [23] algorithms, and methods relying on stronger assumptions [13, 25]. [16] proposed an algorithm to search for the latent variables along the path X and Y, but they only considered discrete random variables.

To the best of our knowledge, this is the first study to find causal direction between confounding, indirectly, and nonlinearly related variables. The considered causal model can be seen as a cascade of processes with a confounder, where each direct cause follows the ANM, and the confounding and intermediate variables are unmeasured. Intuitively, the independence between the noise and cause is still helpful in finding causal direction–the wrong direction will not follow the independence noise condition in the generic case, allowing us to identify causal direction correctly. This will be supported by our theoretical studies and empirical results in subsequent sections.

2 CASCADE ADDITIVE NOISE MODEL

In this section, we first propose the **cascade additive noise model (CANM)** without considering the confounding variables. Then, we develop the variational solution of it. Let X be the cause of effect Y ($X \rightarrow Y$). In CANM, we aim to identify the causal direction between X and Y with unmeasured intermediate variables Z_i between them, as shown in Figure 3(a). If there is no confounder and the data generation follows the nonlinear additive noise assumption. Then, such an indirect causal mechanism can be formalized by the CANM in the following definition.

Definition 1. A CANM for cause X and effect Y is that there exists a sequence of unmeasured intermediate variables between X and Y such that no variable in the latter is the cause of the former one:

$$\begin{cases} Z_1 = f_1(X) + N_1, \\ Z_t = f_t(\mathbf{Z}_{pa(t)}) + N_t, \\ Y = f_{T+1}(\mathbf{Z}_{pa(y)}) + \epsilon_y, \end{cases} \tag{1}$$

where X, N_i, and ϵ_y are mutually independent, T denotes the depth of the chain, and $\mathbf{Z}_{pa(t)}$, $\mathbf{Z}_{pa(y)}$ denote parents of the Z_t and y, respectively. To ensure the cascade structure, the causal relations among Z_i are recursive. Let $\mathbf{f} = \{f_1, f_2, \dots, f_T\}$ and $\mathbf{N} = \{N_1, N_2, \dots, N_T\}$ denote a set of nonlinear functions and the corresponding additive noises at each depth in the chain, respectively. Naturally, here the direct cause and the noises are independent from each other.

We are given a set of data $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^m$. Let θ be the parameters of the causal mechanism. Combing all the independence relations of CANM, we can derive its marginal log-likelihood as follows:

$$\begin{aligned} &\log \prod_{i=1}^m \int p_\theta(x^{(i)}, y^{(i)}, \mathbf{z}) d\mathbf{z} \\ =\ &\log \prod_{i=1}^m \int p_\theta(x^{(i)}) p_\theta(y^{(i)}|\mathbf{z}_{pa(y)}) \prod_{t=2}^T p_\theta(z_t|\mathbf{z}_{pa(t)}) p_\theta(z_1|x^{(i)}) d\mathbf{z} \\ =\ &\log \prod_{i=1}^m \int p(x^{(i)}) p_\theta\left(\epsilon_y^{(i)} = y^{(i)} - f(x^{(i)}, \mathbf{n})\right) \prod_{t=1}^T p_\theta(n_t) d\mathbf{n} \\ =\ &\log \prod_{i=1}^m \int p_\theta\left(x^{(i)}, \epsilon_y^{(i)}, \mathbf{n}\right) d\mathbf{n}. \end{aligned} \tag{2}$$

Equation (2) first decomposes the joint likelihood based on the Markov condition [21], then applies the independence property between the cause and the noise in the second equality, i.e., $p(Z_t|\mathbf{Z}_{pa(t)}) = p(N_t = Z_t - f_t(\mathbf{Z}_{pa(t)})|\mathbf{Z}_{pa(t)}) \overset{\mathbf{Z}_{\mathrm{pa(t)}} \perp \mathrm{N_t}}{=} p(N_t = Z_t - f_t(\mathbf{Z}_{pa(t)}))$. At the same time, we replace $d\mathbf{z}$ with $d\mathbf{n}$ and rewrite function $f_{T+1}(\mathbf{Z}_{pa(y)})$ as $f(X, \mathbf{N})$, because the last unobserved direct cause $\mathbf{Z}_T \subset \mathbf{Z}_{pa(t)}$ contains all the information of the noise $\mathbf{N}$ and cause X relative to Y.

In the above derivation, we used the transformation from X and noises to Y. The property of the transformation helps study identifiability and find a practical solution. In light of the independence property of the noises, we propose a variational approach to approximate the marginal log-likelihood and identify the causal direction.

2.1 Variational Solution of CANM

Due to the intractability of the possible high dimensionality latent variables in the marginal likelihood, in this section, we develop a variational-based method for estimating and optimizing CANM. The variational solution to estimate CANM consists of two steps. First, we take advantage of the independence property between the noise variables, then replace the unobserved intermediate variables $\mathbf{Z}$ with noise $\mathbf{N}$. Second, we find an amortized inference distribution $q_\phi(\mathbf{N}|X, Y)$ with respect to the parameter ϕ to approximate the intractable posterior $p_\theta(\mathbf{N}|X, Y)$ and jointly optimize a variational lower bound of the marginal log-likelihood. Note that Y can be seen as a function of X, N. Thus, N is a function of X and Y, and we need to recover N from both X and Y. Furthermore, the process in estimating N is similar to some works of nonlinear independent component analysis [10, 11] but the process in generating is different because we need to consider the specific causal process. According to Equation (2), the (log) marginal likelihood, as the sum over of the marginal likelihoods over individual data points:

$$\begin{aligned} &\log \prod_{i=1}^m \int p_\theta\left(x^{(i)}, \epsilon_y^{(i)}, \mathbf{n}\right) d\mathbf{n} \\ =\ &\sum_{i=1}^m \underbrace{E_{\mathbf{n} \sim q_\phi(\mathbf{n}|x^{(i)}, y^{(i)})} \left[\log \frac{p_\theta(x^{(i)}, \epsilon_y^{(i)}, \mathbf{n})}{q_\phi(\mathbf{n}|x^{(i)}, y^{(i)})}\right]}_{:=\mathcal{L}(\theta, \phi; x^{(i)}, y^{(i)})} + KL(q_\phi(\mathbf{n}|x^{(i)}, y^{(i)}) \| p_\theta(\mathbf{n}|x^{(i)}, y^{(i)})) \\ \geqslant &\sum_{i=1}^m \mathcal{L}(\theta, \phi; x^{(i)}, y^{(i)}), \end{aligned} \tag{3}$$

where $\mathcal{L}(\theta, \phi; x^{(i)}, y^{(i)})$ be the lower bound at data point $(x^{(i)}, y^{(i)})$, resulting from approximating an intractable posterior $p_\theta(\mathbf{n}|x^{(i)}, y^{(i)})$ by $q_\phi(\mathbf{n}|x^{(i)}, y^{(i)})$. Under the framework of CANM, the lower bound of the total marginal likelihood can be further estimated as follows:

$$\begin{aligned} &\sum_{i=1}^m \mathcal{L}(\theta, \phi; x^{(i)}, y^{(i)}) \\ =&\sum_{i=1}^m E_{\mathbf{n} \sim q_\phi(\mathbf{n}|x^{(i)}, y^{(i)})} \left[-\log q_\phi\left(\mathbf{n}|x^{(i)}, y^{(i)}\right) + \log p_\theta\left(x^{(i)}, \epsilon_y^{(i)}, \mathbf{n}\right)\right] \\ =\ &\sum_{i=1}^m \log p(x^{(i)}) - KL(q_\phi(\mathbf{n}|x^{(i)}, y^{(i)}) \| p_\theta(\mathbf{n})) \\ &+ E_{\mathbf{n} \sim q_\phi(\mathbf{n}|x^{(i)}, y^{(i)})} \left[\log p\left(\epsilon_y^{(i)} = y^{(i)} - f\left(x^{(i)}, \mathbf{n}; \theta\right)\right)\right]. \end{aligned} \tag{4}$$

The details of derivation can be found in Appendix A. As shown in Equation (3), the lower bound $\mathcal{L}$ is tight at $KL(q_\phi(\mathbf{n}|x^{(i)}, y^{(i)}) \| p_\theta(\mathbf{n}|x^{(i)}, y^{(i)})) = 0$. That is, when $q_\phi(\mathbf{n}|x^{(i)}, y^{(i)}) = p_\theta(\mathbf{n}|x^{(i)}, y^{(i)})$, the marginal log-likelihood is equal to the lower bound. Below we will maximize the variational lower bound.

Here, we assume the distributions of noise $\mathbf{N}$ can be factorized as $p_\theta(\mathbf{N}) = \prod_{t=1}^T p_\theta(N_t)$. Note that if $\mathbf{N}$ is an empty set, the above lower bound is equivalent to the log-likelihood of the standard additive noise model.

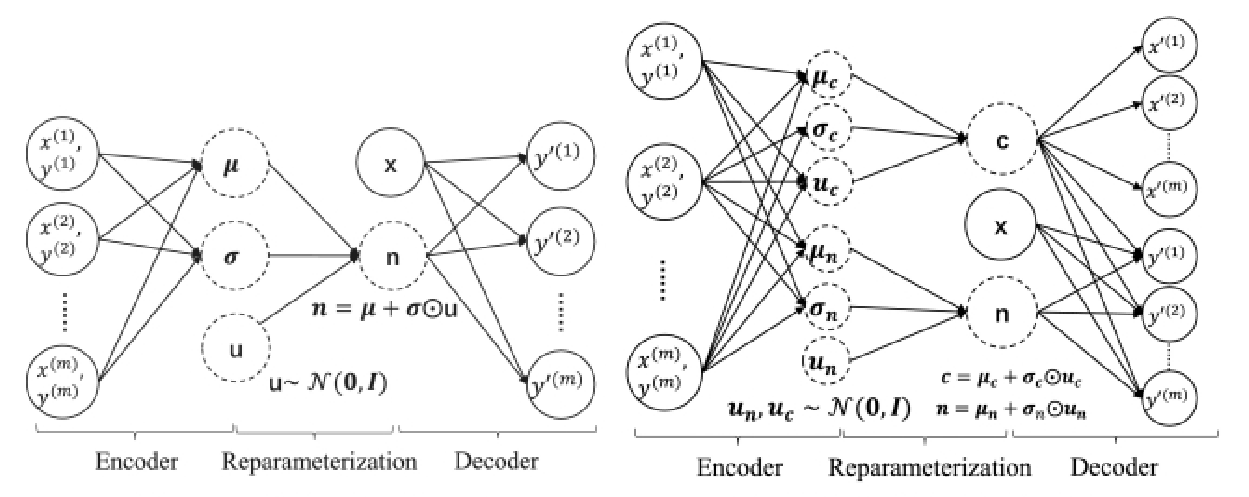
2.2 Variational Auto-Encoder for CANM

Given the lower bound of CANM in Equation (4), however, due to the complex structure in the generative distribution and the possible high dimensionality of the inference distribution, one may need a flexible and efficacious method for optimizing. Fortunately, by utilizing the expressiveness of neural network, the **variational autoencoder (VAE)** based method has been proposed [15], which is flexible enough for our task. In the following, we will show how to design the architecture in VAE for CANM. The design of VAE generally follows the typical configuration in [15]. We denote q_ϕ as *encoder* and p_θ as *decoder*, using a **multilayer perceptron (MLP)** as a universal approximator for these two functions.

In CANM, in the encoder phase, the noises are inferred by an encoder network with a reparameterization trick. That is, reparameterize the random variable $\mathbf{n} \sim q_\phi(\mathbf{n}|x, y)$ using a differentiable transformation $h_\phi(x, y, u)$ such that $\mathbf{n} \sim h_\phi(x, y, u)$ with $u \sim p(u)$. Then the expectation of lower bound $E_{\mathbf{n} \sim q_\phi(\mathbf{n}|x,y)}[p(\epsilon_y^{(i)} = y^{(i)} - f(x^{(i)}, \mathbf{n}; \theta))]$ can be estimated by using Monte Carlo.

In the decoder phase, we estimate the $\epsilon_y^{(i)}$ by subtracting sample $y^{(i)}$ from the reconstruction value of decoder $f(x^{(i)}, h_\phi(x^{(i)}, y^{(i)}, u^{(l)}); \theta)$, where $u^{(l)} \sim p(u)$. Then, alternatively processes the encoder and decoder phases, we can optimize the lower bound until it converges.

Figure 4(a) shows a toy example of the structure of the CANM variational auto-encoder with $q_\phi(\mathbf{n}|x^{(i)}, y^{(i)}) = \mathcal{N}(\mathbf{n}; \mu_\phi(x^{(i)}, y^{(i)}), \sigma_\phi(x^{(i)}, y^{(i)})\mathbf{I})$, where μ_ϕ and σ_ϕ are deterministic functions with parameter ϕ. In the encoder phase, we encode the samples into the noises using a reparameterization trick $\mathbf{n}^{(l)} = \mu_\phi(x^{(i)}, y^{(i)}) + \sigma_\phi(x^{(i)}, y^{(i)}) \odot u^{(l)}$ where $u^{(l)} \sim \mathcal{N}(0, \mathbf{I})$. In the decoder phase, the sample $y^{(i)}$ is reconstructed by the decoder $y\prime^{(i)} = f(x^{(i)}, \mathbf{n}^{(l)}; \theta)$.



(a) Variational Auto-encoder for CANM.

(b) Variational Auto-encoder for CCANM.

Fig. 4. The variational auto-encoder framework on both CANM and CCANM. It contains an encoder and decoder part in both models. The main difference is that, in CCANM, the encoder further infer the confounding noise c while CANM does not. In particular, the encoder uses samples $x^{(i)}, y^{(i)}$ to infer the mean μ and standard variance σ of the latent noise distribution. Then,

in reparameterization phase, we sample the latent noise variables using $\mathbf{n} = \mu + \sigma \odot u$, where $\boldsymbol{u}$ is the sample from standard Gaussian distribution. In CCANM, we will further sample the latent confounders using $\mathbf{c} = \mu_{\mathbf{c}} + \sigma_{\mathbf{c}} \odot u_c$. In the decoder phase, CANM reconstruct the $y'^{(i)}$ from $x, \mathbf{n}$ while CCANM reconstruct $x'^{(i)}$, $y'^{(i)}$ from $\mathbf{c}$ and $x^{(i)}, \mathbf{c}, \mathbf{n}$, respectively.

3 CONFOUNDING CASCADE ADDITIVE NOISE MODEL

In this section, as shown in the toy example in Figure 3(b), we consider a causal pair $X \to Y$ that exists unobserved intermediate variables Z_i as well as the confounding variables C_i. Such a causal mechanism can be formalized as follows.

Definition 2. A CCANM for cause X and effect Y is that there exists a set of unmeasured intermediate and the confounding variables between X and Y such that:

$$\begin{cases} X = f_x(\mathbf{C}) + \epsilon_x, \\ Z_1 = f_1(X) + N_1, \\ Z_t = f_t(Z_{pa(t)}) + N_t, \\ Y = f_{T+1}(Z_{pa(y)}, \mathbf{C}) + \epsilon_y, \end{cases} \tag{5}$$

where $X, N_i, \epsilon_x, \epsilon_y$ are mutually independent, $\mathbf{C}$ denote the set of confounding variables that are independent to each other, T is the depth of the chain, and $Z_{pa(t)}, Z_{pa(y)}$ denote the parents of the Z_t and y, respectively. Each Z_i follows the cascade structure with a recursive additive noise structure.

Similar to CANM, given the dataset $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^m$ generated from CCANM with the generated parameters θ, the marginal log-likelihood can be derived as follows:

$$\begin{aligned} &\log \prod_{i=1}^m \int p_\theta(x^{(i)}, y^{(i)}, \mathbf{z}, \mathbf{c}) d\mathbf{z} d\mathbf{c} \\ =\ &\log \prod_{i=1}^m \int p_\theta(\mathbf{c}) p_\theta(x^{(i)}|\mathbf{c}) p_\theta(y^{(i)}|\mathbf{z}_{pa(y)}, \mathbf{c}) \prod_{t=2}^T p_\theta(z_t|\mathbf{z}_{pa(t)}) p_\theta(z_1|x^{(i)}) d\mathbf{z} d\mathbf{c} \\ =\ &\log \prod_{i=1}^m \int p(\mathbf{c}) p\left(\epsilon_x^{(i)} = x^{(i)} - f_x(\mathbf{c})\right) p_\theta\left(\epsilon_y^{(i)} = y^{(i)} - f(x^{(i)}, \mathbf{n}, \mathbf{c})\right) \prod_{t=1}^T p_\theta(n_t) d\mathbf{n} d\mathbf{c} \\ =\ &\log \prod_{i=1}^m \int p_\theta\left(\epsilon_x^{(i)}, \epsilon_y^{(i)}, \mathbf{n}, \mathbf{c}\right) d\mathbf{n} d\mathbf{c}, \end{aligned} \tag{6}$$

where the first equality is based on the Markov condition and the second equality is based on the fact that the noise of the intermediate variables is independent of its parents. At the same time, we further replace $d\mathbf{z}$ with $d\mathbf{n}$. Compared with the marginal likelihood of CANM in Equation (2), the Equation (6) further introduces the confounding variables such that the likelihood of X becomes the likelihood of the residual ϵ_x.

3.1 Variational Solution of CCANM

In this section, we proposed the variational solution of CCANM. To infer the latent variables in Equation (6), we introduce the variational distribution $q_\phi(\mathbf{C}, \mathbf{N}|X, Y)$ with parameters ϕ aiming to approximate the posterior distribution $p_\theta(\mathbf{C}, \mathbf{N}|X, Y)$. Given the variational distribution, the *evident lower bound* **(ELBO)** can be given as follows:

$$\begin{aligned} &\log \prod_{i=1}^m \int p_\theta\left(\epsilon_x^{(i)}, \epsilon_y^{(i)}, \mathbf{n}, \mathbf{c}\right) d\mathbf{n} d\mathbf{c} \\ =\ &\sum_{i=1}^m \underbrace{E_{\mathbf{n} \sim q_\phi(\mathbf{c}, \mathbf{n}|x^{(i)}, y^{(i)})}\left[\log \frac{p_\theta(\epsilon_x^{(i)}, \epsilon_y^{(i)}, \mathbf{n}, \mathbf{c})}{q_\phi(\mathbf{c}, \mathbf{n}|x^{(i)}, y^{(i)})}\right]}_{:=\mathcal{L}_C(\theta, \phi; x^{(i)}, y^{(i)})} + KL(q_\phi(\mathbf{c}, \mathbf{n}|x^{(i)}, y^{(i)}) \| p_\theta(\mathbf{c}, \mathbf{n}|x^{(i)}, y^{(i)})) \\ \geqslant &\sum_{i=1}^m \mathcal{L}_C(\theta, \phi; x^{(i)}, y^{(i)}), \end{aligned} \tag{7}$$

where $\mathcal{L}_C(\theta, \phi; x^{(i)}, y^{(i)})$ is the variational lower bound of CCANM in data point $(x^{(i)}, y^{(i)})$ which can be further rewritten as follows:

$$\begin{aligned} &\sum_{i=1}^m \mathcal{L}_C(\theta, \phi; x^{(i)}, y^{(i)}) \\ =&\sum_{i=1}^m E_{\mathbf{c}, \mathbf{n} \sim q_\phi(\mathbf{c}, \mathbf{n}|x^{(i)}, y^{(i)})}\left[-\log q_\phi\left(\mathbf{c}, \mathbf{n}|x^{(i)}, y^{(i)}\right) + \log p_\theta\left(\epsilon_x^{(i)}, \epsilon_y^{(i)}, \mathbf{c}, \mathbf{n}\right)\right] \\ =&\sum_{i=1}^m E_{\mathbf{c}, \mathbf{n} \sim q_\phi(\mathbf{c}, \mathbf{n}|x^{(i)}, y^{(i)})}\left[\log p\left(\epsilon_x^{(i)} = x^{(i)} - f_x(\mathbf{c}; \theta)\right) + \log p\left(\epsilon_y^{(i)} = y^{(i)} - f\left(x^{(i)}, \mathbf{c}, \mathbf{n}; \theta\right)\right)\right] \\ &- KL(q_\phi(\mathbf{c}, \mathbf{n}|x^{(i)}, y^{(i)}) \| p_\theta(\mathbf{c}, \mathbf{n})). \end{aligned} \tag{8}$$

The details of derivation can be found in Appendix B. Here we assume that the $\mathbf{c}, \mathbf{n}$ in the prior distribution are independent of each other, i.e., $p(\mathbf{c}, \mathbf{n}) = \prod_i p(c_i) \prod_j p(n_j)$. And the lower bound in Equation (8) is tight if and only if $q_\phi(\mathbf{c}, \mathbf{n}|x^{(i)}, y^{(i)}) = p_\theta(\mathbf{c}, \mathbf{n}|x^{(i)}, y^{(i)})$. It is easy to see that if $\mathbf{c}$ is an empty set or a constant, CCANM can be reduced to CANM.

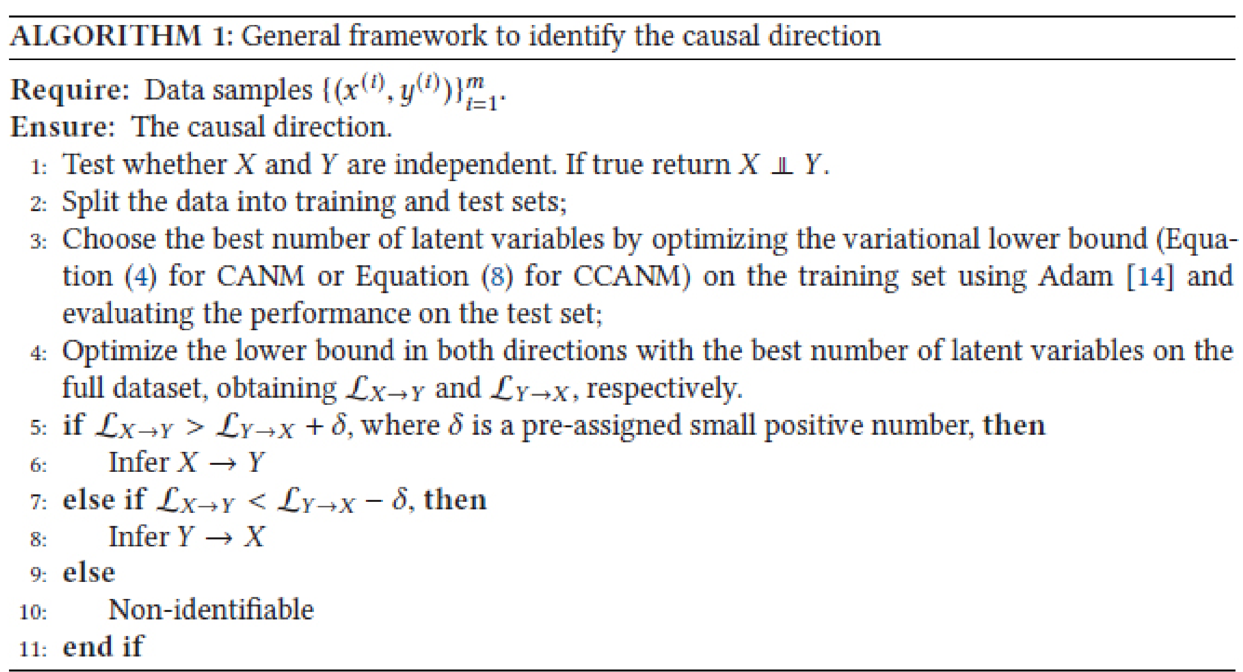
3.2 Variational Auto-Encoder for CCANM

Given the lower bound of CCANM in Equation (8), we can also optimize it using the VAE based method.

The procedure is given in Figure 4(b). In the encode phase, we first infer the $\mathbf{c}, \mathbf{n}$ such that $q_\phi(\mathbf{c}, \mathbf{n}|x^{(i)}, y^{(i)}) = \mathcal{N}(\mathbf{c}, \mathbf{n}; \mu_n(x^{(i)}, y^{(i)}; \phi), \sigma_n(x^{(i)}, y^{(i)}; \phi)\mathbf{I})$. Then, to estimate the expectation in Equation (8), we apply the reparameterization trick, such that the expectation $E_{\mathbf{c}, \mathbf{n} \sim q_\phi(\mathbf{c}, \mathbf{n}|x^{(i)}, y^{(i)})}$ becomes $E_{u_c, u_n \sim \mathcal{N}(0, \mathbf{I})}$ and the Monte Carlo method is used. At the l-th sample in Monte Carlo, we have $\mathbf{c}^{(l)} = \mu_c(x^{(i)}, y^{(i)}; \phi) + \sigma_c(x^{(i)}, y^{(i)}; \phi) \odot u_c^{(l)}$ and $\mathbf{n}^{(l)} = \mu_n(x^{(i)}, y^{(i)}; \phi) + \sigma_n(x^{(i)}, y^{(i)}; \phi) \odot u_n^{(l)}$. Using those samples, in the decoder phase, we use the MLP as the universal approximator to reconstruct X using $x'^{(i)} = f_x(\mathbf{c}^{(l)})$, and Y using $y'^{(i)} = f(x^{(i)}, \mathbf{c}^{(l)}, \mathbf{n}^{(l)})$.

4 PRACTICAL ALGORITHM

Finally, we propose a framework that makes use of the VAE to estimate the marginal log-likelihood as well as identify the causal direction. The framework for causal discovery is given in Algorithm 1.



```
ALGORITHM 1: General framework to identify the causal direction
Require: Data samples $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$.
Ensure: The causal direction.
 1: Test whether $X$ and $Y$ are independent. If true return $X \perp\!\!\!\perp Y$.
 2: Split the data into training and test sets;
 3: Choose the best number of latent variables by optimizing the variational lower bound (Equation (4) for CANM or Equation (8) for CCANM) on the training set using Adam [14] and evaluating the performance on the test set;
 4: Optimize the lower bound in both directions with the best number of latent variables on the full dataset, obtaining $\mathcal{L}_{X \to Y}$ and $\mathcal{L}_{Y \to X}$, respectively.
 5: if $\mathcal{L}_{X \to Y} > \mathcal{L}_{Y \to X} + \delta$, where $\delta$ is a pre-assigned small positive number, then
 6:     Infer $X \to Y$
 7: else if $\mathcal{L}_{X \to Y} < \mathcal{L}_{Y \to X} - \delta$, then
 8:     Infer $Y \to X$
 9: else
10:     Non-identifiable
11: end if
```

Algorithm 1 consists of two phases; the first is model selection, selecting the best number of latent variables, and the second is to identify the causal direction. In phase 1, by splitting the data into training and testing sets, the best number of latent variables is selected based on the performance on the test set (Lines 2–3). Note that it is reasonable to select the sparser number of latent variables in practice when there is lack of sample in the test set. In phase 2, we use the number of the latent variables obtained in phase 1 to optimize the variational lower bound on the full dataset and then identify causal direction according to the likelihood for both directions (Lines 4–11).

5 IDENTIFIABILITY

In this section, we investigate whether there exist any CANMs or CCANMs where the generated data also admit itself in the reverse (anti-causal) direction. In the following theorem, we first analyze the identifiability of CANM and propose a way to derive the noise distribution for the reverse direction $p(\hat{\epsilon})$ by making use of the theory of Fourier transform [1]. The causal direction is unidentifiable according to the CANM if $\hat{\epsilon}$ is independent of Y and $\hat{\mathbf{N}}$ (i.e., the marginal likelihoods for both directions are equal).

Theorem 1. Let $X \to Y$ follow the cascade additive noise model, while there exists a backward model following the same form, i.e.,

$$\begin{aligned} Y = f(X, \mathbf{N}) + \epsilon, &\qquad X, \mathbf{N}, \text{ and } \epsilon\text{~are independent}, \\ X = g(Y, \hat{\mathbf{N}}) + \hat{\epsilon}, &\qquad Y, \hat{\mathbf{N}}, \text{ and } \hat{\epsilon}\text{~are independent}, \end{aligned} \tag{9}$$

then the noise distribution of the reverse direction $p_{\hat{\epsilon}}$ must be

$$p_{\hat{\epsilon}}\left(\hat{\epsilon}\right) = \int e^{2\pi i \hat{\epsilon} \cdot \nu} \frac{\iint p(x) p(\mathbf{n}) p_\epsilon(y - f(x, \mathbf{n})) e^{-2\pi i x \cdot \nu} d\mathbf{n} dx}{p(y) \int p\left(\hat{\mathbf{n}}\right) e^{-2\pi i g(y, \hat{\mathbf{n}}) \cdot \nu} d\hat{\mathbf{n}}} d\nu, \tag{10}$$

where f, g denote the functions implied by the cascade process.

Proof. See Appendix C for the proof.□

Roughly speaking, regardless of the linear case, Theorem 1 implies that the noise distribution in the reverse direction is generally coherent with y. To ensure such noise is independent from Y, one strict condition must hold, i.e., $\hat{\epsilon}$ should be independent from Y in the sense that $\forall y_1, y_2, \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\iint p(x)p(\mathbf{n})p_\epsilon(y_1-f(x,\mathbf{n}))e^{-2\pi ix\cdot\nu}d\mathbf{n}dx}{p(y_1)\int p(\hat{\mathbf{n}})e^{-2\pi ig(y_1,\hat{\mathbf{n}})\cdot\nu}d\hat{\mathbf{n}}}d\nu = \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\iint p(x)p(\mathbf{n})p_\epsilon(y_2-f(x,\mathbf{n}))e^{-2\pi ix\cdot\nu}d\mathbf{n}dx}{p(y_2)\int p(\hat{\mathbf{n}})e^{-2\pi ig(y_2,\hat{\mathbf{n}})\cdot\nu}d\hat{\mathbf{n}}}d\nu$. That is, for all $y_1 \neq y_2$, we have $p(\hat{\epsilon}|y_1) = p(\hat{\epsilon}|y_2) = p(\hat{\epsilon})$. However, in general, it seems that such a condition holds only in restrictive cases. Therefore, in most cases, after the latent noise is recovered, we can identify the causal direction by using the independence property for $(X, \mathbf{N}, \epsilon)$.

To further illustrate the implication of Theorem 1, we provide two special cases in the following corollaries. In Corollary 1, we show that CANM is unidentifiable if the generation process is linear Gaussian. In Corollary 2, based on Theorem 1, we show the connection with ANM when there is no unmeasured intermediate variables and shows a generic choice of f, $p_X(x)$, and $p_\epsilon(\epsilon)$ for the identification of the model. Those two special cases are consistent with the previous results.

Corollary 1. Assume that CANM is linear Gaussian, i.e.,

$$Y = aX + bN + \epsilon, \tag{11}$$

where $X, N, \epsilon \sim \mathcal{N}(0,1)$, then there exists a backward CANM

$$X = \frac{a}{a^2+b^2+1}Y + \frac{a}{\sqrt{a^2+b^2+1}}\hat{N} + \hat{\epsilon}, \tag{12}$$

where $\hat{N}, \hat{\epsilon} \sim \mathcal{N}(0,1)$ and $\hat{\epsilon}$ is independent of Y and $\hat{N}$.

Proof. See Appendix D for the proof.□

Corollary 1 states that for any linear Gaussian CANM in Equation (11), then there exists a backward model as Equation (12) shown. That is, in linear Gaussian, CANM is generally unidentifiable, which is also consistent with the previous results [9].

Corollary 2. Suppose that there is no unmeasured intermediate noise in CANM, if the solution of Equation (10) exists, then for any 3-times differentiable triple (f, p_X, p_ϵ), it must satisfy the differential equation from ANM [9Theorem 1] for all x, y with $\nu''(y - f(x))f'(x) \neq 0$:

$$\xi''' = \xi''\left(-\frac{\nu'''f'}{\nu''} + \frac{f''}{f'}\right) - 2\nu''f''f' + \nu'f''' + \frac{\nu'\nu'''f''f'}{\nu''} - \frac{\nu'(f'')^2}{f'}, \tag{13}$$

where $\nu := \log p_\epsilon$, $\xi := \log p_X$.

Proof. See Appendix E for the proof.□

Corollary 2 is directly derived from Theorem 1 which verifies the generality of Theorem 1 and shows that it will be degenerated to the identifiability of ANM when there is no intermediate noise. That is, in a 3-dimensional affine space, if the condition 13 holds and for any $x, y \in \mathbb{R}$, satisfying $\nu''(y - f(x))f'(x) \neq 0$, then there exists a backward ANM.

For the CCANM, in the following theorem, we analyze the identifiability of CCANM and derive the noise distribution on both directions $p(\hat{\epsilon}_x), p(\hat{\epsilon}_y)$. The causal direction is unidentifiable according to the CCANM if both $\hat{\epsilon}_x$ is independent of Y, $\hat{\mathbf{N}}$, $\mathbf{Z}_c$ as well as $\hat{\epsilon}_y$ is independent of $\mathbf{Z}_c$.

Theorem 2. Let $X \to Y$ follow the confounding cascade additive noise model, while there exists a backward model following the same form:

$$\begin{aligned} X &= g(\mathbf{Z_c}) + \epsilon_x \quad Y = f(X, \mathbf{N}, \mathbf{Z_c}) + \epsilon_y, \\ Y &= \hat{g}(\mathbf{Z_c}) + \hat{\epsilon}_y \quad X = \hat{f}(Y, \hat{\mathbf{N}}, \mathbf{Z_c}) + \hat{\epsilon}_x, \end{aligned} \tag{14}$$

where X is independent of $\mathbf{N}, \epsilon_y$, Y is independent of $\hat{\mathbf{N}}, \hat{\epsilon}_x$, f and $\hat{f}$ denote the functional classes implied by confounding cascade process. Then the noises distribution on the anti-causal direction $p_{\hat{\epsilon}_x}, p_{\hat{\epsilon}_y}$ must be

$$p(\hat{\epsilon}_x) = \int e^{2\pi i\hat{\epsilon}_x\cdot\nu}\frac{\int p(x,y)e^{-2\pi ix\cdot\nu}dx}{\int p(\hat{\mathbf{n}})p(\hat{\mathbf{c}})p(\hat{\epsilon}_y = y - \hat{g}(\hat{\mathbf{c}}))e^{-2\pi i\hat{f}(y,\hat{\mathbf{n}},\hat{\mathbf{c}})\cdot\nu}d\hat{\mathbf{n}}d\hat{\mathbf{c}}}d\nu, \tag{15}$$

$$p(\hat{\epsilon}_y) = \int e^{2\pi i\hat{\epsilon}_y\cdot\mu}\frac{\int p(y)e^{-2\pi iy\cdot\mu}dy}{\int p(\hat{\mathbf{c}})e^{-2\pi i\hat{g}(\hat{\mathbf{c}})\cdot\mu}d\hat{\mathbf{c}}}d\mu \tag{16}$$

Proof. See Appendix F for the proof.□

Intuitively, Theorem 2 states that CCANM is not identifiable if the Equation (15) is independent of Y. That is, for any $y_1 \neq y_2$, we must have $\int e^{2\pi i\hat{\epsilon}_x\cdot\nu}\frac{\int p(x,y_1)e^{-2\pi ix\cdot\nu}dx}{\int p(\hat{\mathbf{n}})p(\hat{\mathbf{c}})p(\hat{\epsilon}_y=y_1-\hat{g}(\hat{\mathbf{c}}))e^{-2\pi i\hat{f}(y_1,\hat{\mathbf{n}},\hat{\mathbf{c}})\cdot\nu}d\hat{\mathbf{n}}d\hat{\mathbf{c}}}d\nu \neq \int e^{2\pi i\hat{\epsilon}_x\cdot\nu}\frac{\int p(x,y_2)e^{-2\pi ix\cdot\nu}dx}{\int p(\hat{\mathbf{n}})p(\hat{\mathbf{c}})p(\hat{\epsilon}_y=y_2-\hat{g}(\hat{\mathbf{c}}))e^{-2\pi i\hat{f}(y_2,\hat{\mathbf{n}},\hat{\mathbf{c}})\cdot\nu}d\hat{\mathbf{n}}d\hat{\mathbf{c}}}d\nu$. Such condition does not necessarily hold in practice. To further illustrate the Theorem 2, we provide the following corollary to show the unidentifiable case in linear Gaussian system.

Corollary 3. Assume that CCANM is linear Gaussian, i.e.,

$$X = aC + \epsilon_x \quad Y = bX + cN + dC + \epsilon_y,$$

where $C, N, \epsilon_x, \epsilon_y \sim \mathcal{N}(0,1)$, then there exists a backward CCANM

$$Y = \hat{a}\hat{C} + \hat{\epsilon}_y \quad X = eY + f\hat{N} + g\hat{C} + \hat{\epsilon}_x,$$

where $\hat{N}, \hat{C} \sim \mathcal{N}(0,1)$, $\hat{\epsilon}_x \sim \mathcal{N}(0, \sigma^2_{\hat{\epsilon}_x})$, $\hat{\epsilon}_y \sim \mathcal{N}(0, \sigma^2_{\hat{\epsilon}_y})$ and $\hat{\epsilon}_x, \hat{\epsilon}_y$ are independent of Y and $\hat{C}$ if the following four conditions hold:

$$\begin{aligned} &\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y} + \hat{a}^2\right)} + e - \frac{b(a^2+1)}{(c^2+d^2+a^2+2)} = 0 \\ &\left(\frac{(a^2+1)}{(c^2+d^2+a^2+2)}\right)^2 - \frac{(a^2+1)}{(c^2+d^2+a^2+2)} - \frac{\hat{a}^2}{2\sigma^2_{\hat{\epsilon}_y}\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)} + \frac{1}{2\sigma^2_{\hat{\epsilon}_y}} = 0, \\ &\frac{z^2\sigma^2_{\hat{\epsilon}_y}(c^2+d^2+a^2+2)}{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2} = \frac{(a^2+1)(c^2+d^2+1)}{c^2+d^2+a^2+2} + f^2 + \frac{g^2\sigma^2_{\hat{\epsilon}_y}}{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2}, \\ &(b^2(a^2+1) + c^2 + d^2 + 1) - \hat{a}^2 = \sigma^2_{\hat{\epsilon}_y}. \end{aligned} \tag{17}$$

Proof. See Appendix G for the proof.□

In Corollary 3, we investigate the conditions that CCANM has a backward model. We find that if the four conditions in Equation (17) holds, then in a linear Gaussian cased, a backward CCANM exists. We can see that such a condition is also stricter than CANM, which is benefited from the existence of confounders. In the following experiments, we will further investigate the effectiveness of the proposed methods.

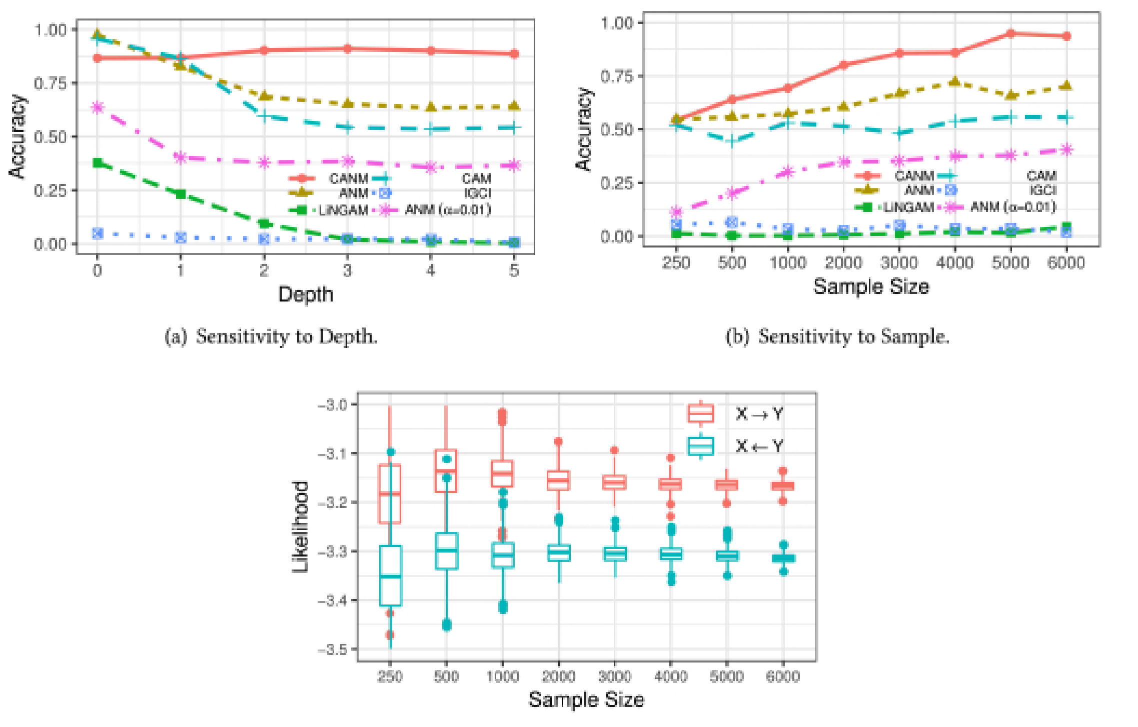
6 SYNTHETIC DATA EXPERIMENTS

In this section, we test the performance of CANM and CCANM on synthetic data, respectively. For comparison, the following four algorithms are taken as baseline methods: ANM [9], CAM [2], IGCI [12], and LiNGAM [20]. We also improve the implementation for ANM by using the XGBoost [4] for regression and the **Hilbert-Schmidt independence criterion (HSIC)** [7] as the independence test. Therefore, ANM can be evaluated in two ways. First, we compare the HSIC statistic to determine the direction, and second, we select the best significance level ($p = 0.01$) ranging from 0.01 to 1 to determine the causal direction. At the same time, the best parameter setting of IGCI is chosen. For the other baseline methods, we use the parameter settings in their original papers. The implementation and the parameter settings of LiNGAM and CAM are based on the CompareCausalNetworks packages in R [8].

6.1 Unobserved Intermediate Variables

To test the performance of CANM, we design three experiments with known ground truth, with the depth $= \{0, 1, 2, \mathbf{3}, 4, 5\}$, sample size $= \{250, 500, 1000, 2000, \mathbf{3000}, 4000, 5000, 6000\}$, and with different sample sizes for some fixed structure. The default setting is marked in bold. All experimental results are averaged over 1,000 randomly generated causal pairs by the cascade additive noise model. Code for CANM is available online.[1]

To make the synthetic data general enough, in each depth, we randomly generate an additive noise model and then stack it together to obtain the cascade additive noise model. In detail, the cause (X) is sampled from a random Gaussian Mixture model of three components. For each layer $x_t = f_t(x_{t-1}) + n_t$ where $n_t \sim \mathcal{N}(0,1)$ and f_t is generated from a cubic spline interpolation using a 6-dimensional grid from $\min(x_{t-1})$ to $\max(x_{t-1})$ as input with respect to six random generated points as knots for the interpolation; the generated points are sampled from $\mathcal{N}(0,1)$ and the number of knots is used to control non-linearity of the function. Such generative process follows the instrument given in [18].



(a) Sensitivity to Depth.

(b) Sensitivity to Sample.

(c) Sensitivity to Sample in a Fixed Structure.

Fig. 5. Sensitivity experiments of CANM.

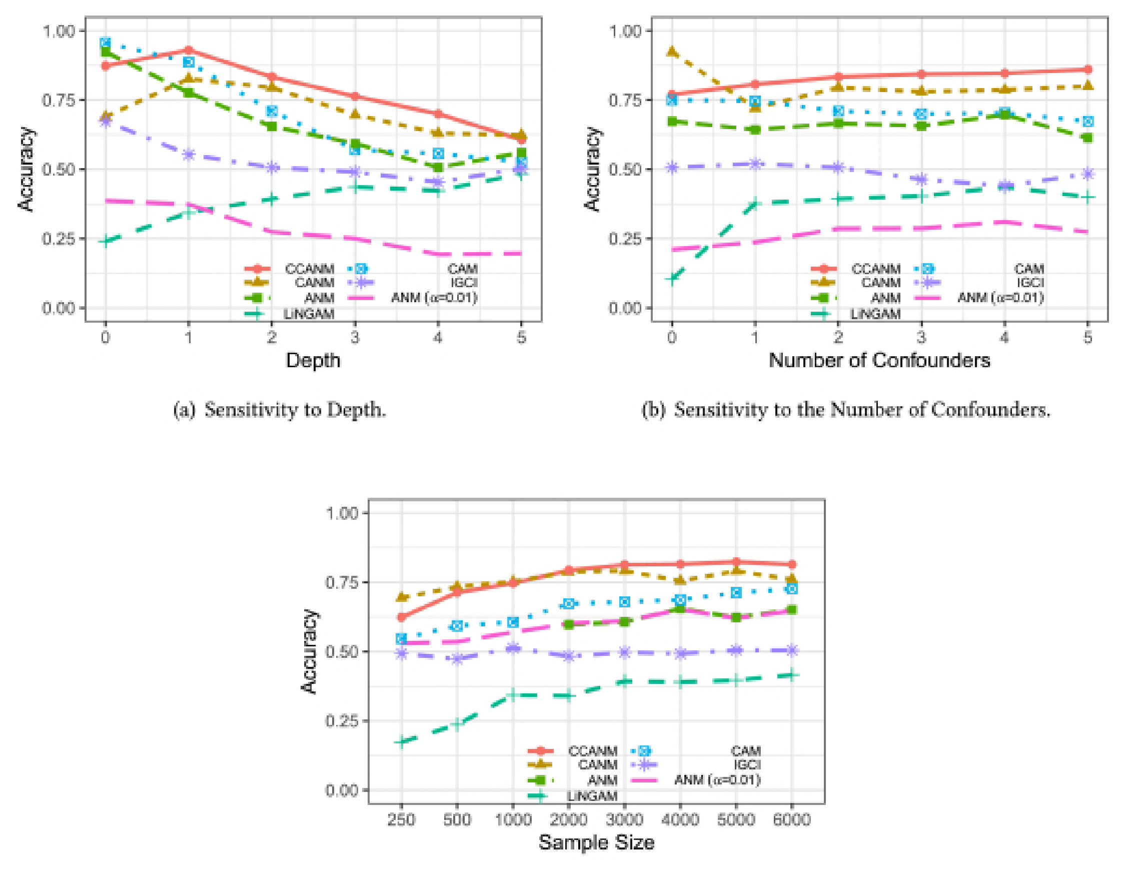
Sensitivity to Depth. Figure 5(a) shows the accuracy with different depths in 3,000 samples. First, when the depth equals 0 (the original ANM), all CANM, ANM, and CAM achieve high accuracy. Note that CANM still has similar performance compared with ANM even though CANM assumes that unmeasured intermediate variables might exist, which demonstrates the robustness of our method. Second, as the depth increases, the accuracy of CANM is stable and around 90% accuracy with a slight decrease. In contrast, the performance of the other methods decreases rapidly as the depth grows. In particular, the ANM with the significance level of 0.01 gives almost random decisions when the cascade structure exists.

Sensitivity to Sample Size. Figure 5(b) shows the accuracy with different sample sizes while the depth is fixed at 3. The result shows that even in the small sample size, CANM still outperforms the other methods. As the sample size increases, the accuracy of CANM grows faster than the other methods. Thus, large samples are beneficial to CANM because of the variational auto-encoder framework employed in CANM. A similar result also can be observed in ANM and CAM, while the other methods are less sensitive to the sample size due to the model restriction.

Sensitivity to Sample Size in a Fixed Structure. Figure 5(c) shows the accuracy with different numbers of samples. At the same time, we use a fixed causal mechanism, which was randomly generated with depth = 3. When the sample size is small, the variance of the likelihood is significant; however, the asymmetry in the causal direction is still explicit. As the sample size increases, the variance of the likelihood decreases, and the accuracy increases, which implies the effectiveness and robustness of CANM as the sample size grows.

6.2 Unobserved Confounding Intermediate Variables

To test the performance of CCANM, we design three sensitivity experiments with different depths, sample sizes, and numbers of confounders, in which the default setting is set as depth $= 2$, sample size $= 5000$, and number of confounders $= 2$.



(a) Sensitivity to Depth.

(b) Sensitivity to the Number of Confounders.

(c) Sensitivity to Sample.

Fig. 6. Sensitivity experiments of CCANM.

To synthetic the data with unobserved confounding and intermediate variables, based on the generating process of CANM, we further randomly generate additive noise models for the confounders such that $X = \sum_{i=1}^{|\mathbf{C}|} f_x^{(i)}(C_i) + \epsilon_x, Y = f_{T+1}(Z_{pa(y)}) + \sum_{i=1}^{|\mathbf{C}|} f_y^{(i)}(C_i) + \epsilon_y$ where $|\mathbf{C}|$ is the number of confounders, $f_x^{(i)}$ and $f_y^{(i)}$ denote the i-th random generated function for C_i, respectively. The standard error of the Gaussian noises in between the intermediate variables are set to 0.1 to adjust the signal to noise rate to provide more convincing experiment results.

Sensitivity to Depth. Figure 6(a) shows the accuracy with different depths in the presence of confounders. In general, the performance of data in CCANM is not significant as the data in CANM. The reason is that CCANM has a more complex generating causal process. In particular, the number of noises is much larger than the data in CANM, which will hinge the performance of all methods. At first, CCANM, CAM, ANM achieve relatively high accuracy at depth = 0. Though the data will degenerate to the standard additive noise model when depth = 0, CCANM also has competitive performance compared with ANM and CAM, which shows the robustness of our method. As the depth increases, CCANM outperforms other baseline methods. Furthermore, given the confounders, all methods are sensitive to depth. Thus a relatively simple structure is still required to achieve a better result.

Sensitivity to the number of confounders. Figure 6(b) shows the accuracy with different numbers of confounders. First, when the number of confounders equals 0 (the original CANM), the CANM achieves the best result due to consistency between the model assumption and the data itself. As the number of confounders increase, CCANM has the best performance compared with all other baseline methods. In conclusion, under the data with unobserved confounding cascade variables, CCANM has a better performance.

Sensitivity to Sample Size. Figure 6(c) shows the accuracy with different numbers of sample sizes. The result shows that all methods are sensitive to the sample size, and relatively large sample size is required in such a complex cause mechanism. According to the experimental results, 1,000 or more samples should be enough to achieve good performance.

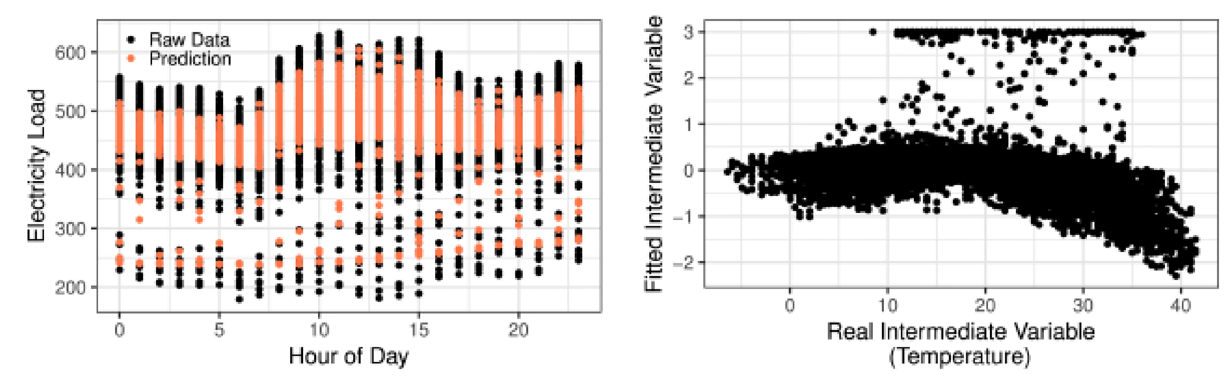
RMSE of the estimated number of latent variables. We further test the **Root-Mean-Square Error (RMSE)** between the true number of latent variables and the estimated number of latent variables in the learning procedure. All experiments run in the default setting where sample size = 5000, depth = 2, and the number of confounders = 2. The RMSE of depth is 0.535 while the RMSE of the number of confounders is 0.713. We can see that the error is relatively small in both depth and the number of confounders, which shows the effectiveness of the model selection procedure.

7 REAL-WORLD DATA EXPERIMENTS

In this section, to test the effectiveness of our proposed methods, we use the real-world causal pairs from Tübingen causal effect benchmark[2] [17]. Note that it is unlikely that all causal pairs in this benchmark meet the assumption of our method according to [17], in which most causal discovery methods that do not allow latent variables among the causal pair are well-performed in this benchmark. Thus, in the following, we select three datasets that meet the assumptions of CANM and CCANM, respectively, to verify the effectiveness of our methods.

7.1 Unobserved Intermediate Variables

Electricity Consumption Dataset. The electricity consumption dataset [18] has 9504-hour measurements from the energy industry, containing the *hour of data*, outside *temperature* and the *electricity load* on the power station. The causal mechanism among the three variables is $hour\ of\ day \rightarrow temperature$ and $temperature \rightarrow electricity\ load$. The first pair is generally caused by the heating of sunlight and the second pair is based on the fact that the usage of heating or air condition depends upon temperature. We are interested in knowing whether we can identify whether the $hour\ of\ day\ (X)$ is the cause of the $electricity\ load\ (Y)$ and what intermediate variable will be inferred via CANM.



(a) Hour of Day Against Electricity Load.

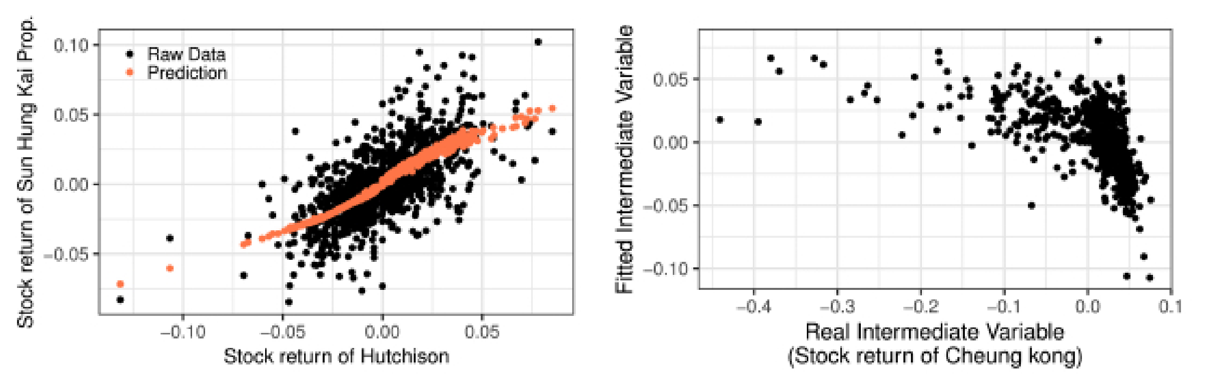
(b) Temperature Against Fitted Intermediate Variable.

Fig. 7. Electricity Consumption.

In general, we successfully identify the correct causal direction with average score $\mathcal{L}_{X \rightarrow Y} = -2.62 > \mathcal{L}_{Y \rightarrow X} = -2.67$ while ANM fails on this pair (the p-value $= 0$ on both directions). The prediction of *electricity* is given in Figure 7(a). Interestingly, more than one unmeasured variable might exist, e.g., season, causing a different electricity load at the same hour of day. Such unmeasured variables are successfully captured by CANM as the prediction separating into both upper and lower parts. Furthermore, the intermediate variable inferred by our method has rather high correlation ($\rho = -0.35$) with the temperature as shown in Figure 7(b), which means that CANM recovers not only the information of the season but also the information of the temperature.

Stock Market Dataset. The stock market dataset is collected by Tübingen causal effect benchmark as pairs 66–67. It contains the stock return of *Hutchison*, *Cheung Kong*, and *Sun Hung Kai* with the causal relationship: $Hutchison \rightarrow Cheung\ Kong$ and $Cheung\ Kong \rightarrow Sun\ Hung\ Kai$. The reason for the first pair is that Cheung Kong owns about 50% of Hutchison. For the second pair, Sun Hung Kai Prop., a typical stock in the Hang Seng

Property subindex, is believed to depend on the major stock Cheung Kong. Similar to the previous experiment, we are interested in whether we can identify that $Hutchison$ (X) is the cause of the $Sun\ Hung\ Kai$ (Y).



(a) Stock return of Hutchison Against Stock return of Sun Hung Kai Prop. (b) Stock return of Cheung kong Against Fitted Intermediate Variable.

Fig. 8. Stock Market.

Since these three stocks form a causal chain such that $Hutchison \rightarrow Cheung\ Kong \rightarrow Sun\ Hung\ Kai$, using CANM, we successfully identify the indirect causal direction with average score $\mathcal{L}_{X\rightarrow Y} = -2.49 > \mathcal{L}_{Y\rightarrow X} = -2.51$ while ANM fails on this pair (the p-value $= 0.006 < 0.05$ on the causal direction and p-value $= 0.29 > 0.05$ on the reverse direction). Figure 8(a) shows the prediction of the stock return of the $Sun\ Hung\ Kai$. We also find that the fitted intermediate variable has a high correction ($\rho = -0.54$) with the stock return of $Cheung\ Kong$ as shown in Figure 8(b).

7.2 Unobserved Confounding Intermediate Variables

Whistler Daily Snowfall Dataset. The Whistler daily snowfall dataset is collected by Tübingen causal effect benchmark as pairs 87. It concerns the historical daily *Temperature* and *Total snow* in Whistler, BC, Canada, over the period July 1, 1972, to December 31, 2009. In common sense, the temperature is one of the causes of total snow, but confounders are expected to be present (e.g., clouds or air pollution). Therefore, in this case, we are interested in whether we can identify that $Temperature$ (X) is the cause of the $Total\ snow$ (Y) with possible existence of confounders via CCANM.

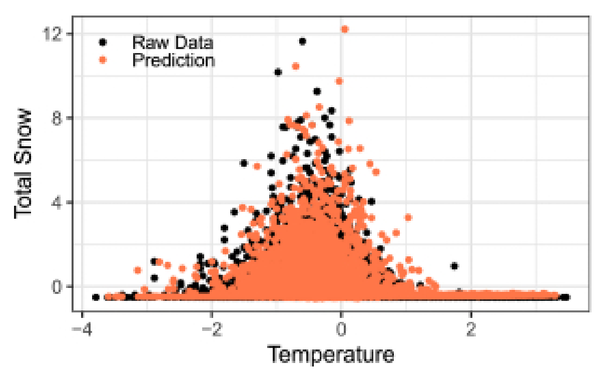


Fig. 9. Whistler Daily Snowfall.

Using CCANM, we successfully identify the causal direction with average score $\mathcal{L}_{X\rightarrow Y} = -3.01 > \mathcal{L}_{Y\rightarrow X} = -3.34$ while the ANM and CANM fail on this pair. In ANM, the p-values in both directions are zeros. Figure 9 shows the reconstruction of the causal pair, in which the reconstructed data is very close to the original data, which shows the effectiveness of CCANM for approximating the latent confounders.

8 CONCLUSION

This paper proposes the confounding cascade nonlinear additive noise model as an extension of the nonlinear additive noise model. We first develop the cascade nonlinear additive noise model to represent indirect causal influences, which result from unmeasured intermediate causal variables. Then, we further develop a general confounding cascade nonlinear additive noise model to model the unobserved confounder and intermediate variables. We demonstrate that the independence between the noise and cause is still helpful in determining the causal direction between two variables. We propose to estimate the confounding and intermediate causal model using the variational auto-encoder framework, and the produced likelihood indicates the asymmetry between cause and effect. As supported by our theoretical and empirical results, the proposed approach effectively determines the causal direction from data generated by nonlinear, confounding indirect causal relations. Future research can extend CCANM to the more general measurement error cases allowing the causal relationship between the latent confounders.

APPENDICES

A THE LOWER BOUND FOR CASCADE NONLINEAR ADDITIVE NOISE MODEL

The following derivation is based on Equation (2). To obtain the lower bound, we will introduce a variational distribution q and we obtain the lower bound as follows:

$$
\begin{aligned}
&\log p(x^{(i)}, y^{(i)}) \\
=\ & \log p\left(x^{(i)}, \epsilon_y^{(i)}, \mathbf{n}\right) - \log p(\mathbf{n}|x^{(i)}, y^{(i)}) \\
=\ & \log\left(\frac{p(x^{(i)}, \epsilon_y^{(i)}, \mathbf{n})}{q(\mathbf{n}|x^{(i)}, y^{(i)})}\right) - \log\left(\frac{p(\mathbf{n}|x^{(i)}, y^{(i)})}{q(\mathbf{n}|x^{(i)}, y^{(i)})}\right) \\
=\ & \log p\left(x^{(i)}, \epsilon_y^{(i)}, \mathbf{n}\right) - \log q(\mathbf{n}|x^{(i)}, y^{(i)}) - \log\left(\frac{p(\mathbf{n}|x^{(i)}, y^{(i)})}{q(\mathbf{n}|x^{(i)}, y^{(i)})}\right) \\
=\ & \log p\left(x^{(i)}, \epsilon_y^{(i)}\right) + \log p(\mathbf{n}) - \log q(\mathbf{n}|x^{(i)}, y^{(i)}) - \log\left(\frac{p(\mathbf{n}|x^{(i)}, y^{(i)})}{q(\mathbf{n}|x^{(i)}, y^{(i)})}\right) \\
=\ & \log p\left(\epsilon_y^{(i)} = y - f(\mathbf{n}, x^{(i)})\right) + \log p(x^{(i)}) + \log p(\mathbf{n}) - \log q(\mathbf{n}|x^{(i)}, y^{(i)}) - \log\left(\frac{p(\mathbf{n}|x^{(i)}, y^{(i)})}{q(\mathbf{n}|x^{(i)}, y^{(i)})}\right) \\
\overset{(i)}{=} & \log p(x^{(i)}) + \int q(\mathbf{n}|x^{(i)}, y^{(i)}) \log p\left(\epsilon_y^{(i)} = y^{(i)} - f\left(\mathbf{n}, x^{(i)}\right)\right) d\mathbf{n} \\
& + \int q(\mathbf{n}|x^{(i)}, y^{(i)}) \log \frac{p(\mathbf{n})}{q(\mathbf{n}|x^{(i)}, y^{(i)})} d\mathbf{n} - \int q(\mathbf{n}|x^{(i)}, y^{(i)}) \log\left(\frac{p(\mathbf{n}|x^{(i)}, y^{(i)})}{q(\mathbf{n}|x^{(i)}, y^{(i)})}\right) d\mathbf{n} \\
=\ & \log p(x^{(i)}) + E_{\mathbf{n}\sim q(\mathbf{n}|x^{(i)}, y^{(i)})}\left[\log p\left(\epsilon_y^{(i)} = y^{(i)} - f\left(\mathbf{n}, x^{(i)}\right)\right)\right] - KL(q(\mathbf{n}|x^{(i)}, y^{(i)}) \| p(\mathbf{n})) \\
& + KL(q(\mathbf{n}|x^{(i)}, y^{(i)}) \| p(\mathbf{n}|x^{(i)}, y^{(i)})) \\
\overset{(ii)}{\geqslant} & \log p(x^{(i)}) + E_{\mathbf{n}\sim q(\mathbf{n}|x^{(i)}, y^{(i)})}\left[\log p\left(\epsilon_y^{(i)} = y^{(i)} - f\left(\mathbf{n}, x^{(i)}\right)\right)\right] - KL(q(\mathbf{n}|x^{(i)}, y^{(i)}) \| p(\mathbf{n}))
\end{aligned}
\tag{A.1}
$$

In equality (i), we take the integral of both sides of the equation. Because the left hand side $\int q(\mathbf{n}|x^{(i)}, y^{(i)}) \log p(x^{(i)}, y^{(i)}) dn = \log p(x^{(i)}, y^{(i)})$ does not change, we obtain the equality. In (ii), because $KL(q(\mathbf{n}|x^{(i)}, y^{(i)}) \| p(\mathbf{n})) \geq 0$, we obtain the last inequality.

B THE LOWER BOUND FOR CONFOUNDING CASCADE NONLINEAR ADDITIVE NOISE MODEL

The derivation for CCANM is similar to CANM, and the only difference is that we take both $\mathbf{n}$ and $\mathbf{c}$ into account.

$$
\begin{aligned}
&\log p(x^{(i)}, y^{(i)}) \\
=\ & \log p\left(\epsilon_x^{(i)}, \epsilon_y^{(i)}, \mathbf{n}, \mathbf{c}\right) - \log p(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)}) \\
=\ & \log\left(\frac{p(\epsilon_x^{(i)}, \epsilon_y^{(i)}, \mathbf{n}, \mathbf{c})}{q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}\right) - \log\left(\frac{p(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}{q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}\right) \\
=\ & \log p\left(\epsilon_x^{(i)}, \epsilon_y^{(i)}, \mathbf{n}, \mathbf{c}\right) - \log q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)}) - \log\left(\frac{p(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}{q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}\right) \\
=\ & \log p\left(\epsilon_x^{(i)}, \epsilon_y^{(i)}\right) + \log p(\mathbf{n}, \mathbf{c}) - \log q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)}) - \log\left(\frac{p(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}{q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}\right) \\
=\ & \log p\left(\epsilon_x^{(i)} = x^{(i)} - g(\mathbf{c})\right) + \log p\left(\epsilon_y^{(i)} = y - f\left(x^{(i)}, \mathbf{n}, \mathbf{c}\right)\right) \\
& + \log p(\mathbf{n}, \mathbf{c}) - \log q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)}) - \log\left(\frac{p(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}{q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}\right) \\
\overset{(i)}{=} & \int q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)}) \left[\log p\left(\epsilon_x^{(i)} = x^{(i)} - g(\mathbf{c})\right) + \log p\left(\epsilon_y^{(i)} = y^{(i)} - f\left(x^{(i)}, \mathbf{n}, \mathbf{c}\right)\right)\right] d\mathbf{n} d\mathbf{c} \\
& + \int q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)}) \log \frac{p(\mathbf{n}, \mathbf{c})}{q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})} d\mathbf{n} d\mathbf{c} - \int q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)}) \log\left(\frac{p(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}{q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}\right) d\mathbf{n} d\mathbf{c} \\
=\ & E_{\mathbf{n}, \mathbf{c}\sim q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}\left[\log p\left(\epsilon_x^{(i)} = x^{(i)} - g(\mathbf{c})\right) + \log p\left(\epsilon_y^{(i)} = y^{(i)} - f\left(x^{(i)}, \mathbf{n}, \mathbf{c}\right)\right)\right] \\
& - KL(q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)}) \| p(\mathbf{n}, \mathbf{c})) + KL(q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)}) \| p(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})) \\
\geqslant & E_{\mathbf{n}, \mathbf{c}\sim q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)})}\left[\log p\left(\epsilon_x^{(i)} = x^{(i)} - g(\mathbf{c})\right) + \log p\left(\epsilon_y^{(i)} = y^{(i)} - f\left(x^{(i)}, \mathbf{n}, \mathbf{c}\right)\right)\right] \\
& - KL(q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)}) \| p(\mathbf{n}, \mathbf{c}))
\end{aligned}
\tag{B.1}
$$

Similarly, in (i), we take integral of $\int q(\mathbf{n}, \mathbf{c}|x^{(i)}, y^{(i)}) d\mathbf{n} d\mathbf{c}$ on both sides, and due to the left hand side does not have $\mathbf{n}$ and $\mathbf{c}$, we obtain the equality.

C PROOF OF THEOREM 1

Theorem 1. Let $X \rightarrow Y$ follow the cascade additive noise model, while there exists a backward model following the same form, i.e.

$$
\begin{aligned}
Y &= f(X, \mathbf{N}) + \epsilon, && X, \mathbf{N}, \text{ and } \epsilon\text{-are independent,} \\
X &= g(Y, \hat{\mathbf{N}}) + \hat{\epsilon}, && Y, \hat{\mathbf{N}}, \text{ and } \hat{\epsilon}\text{-are independent,}
\end{aligned}
\tag{C.1}
$$

then the noise distribution of the reverse direction $p_{\hat{\epsilon}}$ must be

$$p_{\hat{\epsilon}}\left(\hat{\epsilon}\right)=\int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\iint p(x)p(\mathbf{n})p_{\epsilon}(y-f(x,\mathbf{n}))e^{-2\pi ix\cdot\nu}d\mathbf{n}dx}{p(y)\int p\left(\hat{\mathbf{n}}\right)e^{-2\pi ig(y,\hat{\mathbf{n}})\cdot\nu}d\hat{\mathbf{n}}}d\nu, \tag{C.2}$$

where f, g denote the function implied by the cascade process.

Sketch of Proof: Based on the derivation of the marginal log-likelihood at Equation (2) in Section 2, if Equation (C.1) holds, we have

$$\begin{aligned} p(y|x) &= \int p(\mathbf{n})p_{\epsilon}(\epsilon = y - f(x,\mathbf{n}))d\mathbf{n}, \\ p(x|y) &= \int p(\hat{\mathbf{n}})p_{\hat{\epsilon}}(\hat{\epsilon} = x - g\left(y,\hat{\mathbf{n}}\right))d\hat{\mathbf{n}}. \end{aligned} \tag{C.3}$$

Applying Fourier transform to $p(x|y)$, we obtain

$$\begin{aligned} \mathcal{F}(\nu) &= \int p(x|y)e^{-2\pi ix\cdot\nu}dx \\ &= \int p\left(\hat{\mathbf{n}}\right)\int p_{\hat{\epsilon}}\left(x - g\left(y,\hat{\mathbf{n}}\right)\right)e^{-2\pi ix\cdot\nu}dxd\hat{\mathbf{n}}. \end{aligned} \tag{C.4}$$

Since $\hat{\epsilon} = x - g\left(y,\hat{\mathbf{n}}\right)$, we have $d\hat{\epsilon} = dx$. By making use of the convolution theorem, the above equation can be rewritten as follows,

$$\begin{aligned} \mathcal{F}(\nu) &= \int p\left(\hat{\mathbf{n}}\right)\int p_{\hat{\epsilon}}\left(\hat{\epsilon}\right)e^{-2\pi i(\hat{\epsilon}+g(y,\hat{\mathbf{n}}))\cdot\nu}d\hat{\epsilon}\,d\hat{\mathbf{n}} \\ &= \int p\left(\hat{\mathbf{n}}\right)e^{-2\pi ig(y,\hat{n})\cdot\nu}d\hat{\mathbf{n}}\int p_{\hat{\epsilon}}\left(\hat{\epsilon}\right)e^{-2\pi i\hat{\epsilon}\cdot\nu}d\hat{\epsilon}. \end{aligned} \tag{C.5}$$

Combining Equations (C.4) and (C.5), we have

$$\int p_{\hat{\epsilon}}\left(\hat{\epsilon}\right)e^{-2\pi i\hat{\epsilon}\cdot\nu}d\hat{\epsilon} = \frac{\int p(x|y)e^{-2\pi ix\cdot\nu}dx}{\int p\left(\hat{\mathbf{n}}\right)e^{-2\pi ig(y,\hat{\mathbf{n}})\cdot\nu}d\hat{\mathbf{n}}}. \tag{C.6}$$

Then, applying the inverse Fourier transform, we conclude

$$p_{\hat{\epsilon}}\left(\hat{\epsilon}\right) = \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\int p(x|y)e^{-2\pi ix\cdot\nu}dx}{\int p\left(\hat{\mathbf{n}}\right)e^{-2\pi ig(y,\hat{\mathbf{n}})\cdot\nu}d\hat{\mathbf{n}}}d\nu. \tag{C.7}$$

Based on Bayes' theorem, $p(x|y) = \frac{p(x)p(y|x)}{p(y)} = \frac{p(x)\int p(\mathbf{n})p_{\epsilon}(y-f(x,\mathbf{n}))d\mathbf{n}}{p(y)}$, and we further have

$$p_{\hat{\epsilon}}\left(\hat{\epsilon}\right)=\int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\iint p(x)p(\mathbf{n})p_{\epsilon}(y-f(x,\mathbf{n}))e^{-2\pi ix\cdot\nu}d\mathbf{n}dx}{p(y)\int p\left(\hat{\mathbf{n}}\right)e^{-2\pi ig(y,\hat{\mathbf{n}})\cdot\nu}d\hat{\mathbf{n}}}d\nu.$$

□

D PROOF OF COROLLARY 1

Corollary 1. Assume that CANM is linear Gaussian, i.e.,

$$Y = aX + bN + \epsilon,$$

where $X, N, \epsilon \sim \mathcal{N}(0,1)$, then there exists a backward CANM

$$X = \frac{a}{a^2+b^2+1}Y + \frac{a}{\sqrt{a^2+b^2+1}}\hat{N} + \hat{\epsilon},$$

where $\hat{N}, \hat{\epsilon} \sim \mathcal{N}(0,1)$ and $\hat{\epsilon}$ is independent of Y and $\hat{N}$.

Proof. Based on Theorem 1, the noise distribution on the reverse direction can be expressed as

$$p_{\hat{\epsilon}}\left(\hat{\epsilon}\right) = \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\int p(x|y)e^{-2\pi ix\cdot\nu}dx}{\int p\left(\hat{n}\right)e^{-2\pi ig(y,\hat{n})\cdot\nu}d\hat{n}}d\nu. \tag{D.1}$$

Based on the Bayes' theorem, $p(x|y) = \frac{p(x)p(y|x)}{p(y)} = \frac{p(x)p_{\tilde{\epsilon}}(y-ax)}{p(y)}$, where $\tilde{\epsilon} \sim \mathcal{N}(0, b^2+1)$ is the distribution of the $\tilde{\epsilon} = bn + \epsilon$. Without loss of generality, let $g(y,\hat{n}) = cy + d\hat{n}$, we have

$$p_{\hat{\epsilon}}\left(\hat{\epsilon}\right) = \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\int p_{\tilde{\epsilon}}(y-ax)p(x)e^{-2\pi ix\cdot\nu}dx}{p(y)\int p\left(\hat{n}\right)e^{-2\pi i(cy+d\hat{n})\cdot\nu}d\hat{n}}d\nu.$$

The following derivation using the fact that the Fourier transform of the Gaussian distribution is

$$\mathcal{F}_x\left[\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\right](\nu) = e^{-2\pi i\mu\cdot\nu}e^{-2\pi^2\sigma^2\cdot\nu^2},$$

then we have

$$\begin{aligned} p_{\hat{\epsilon}}\left(\hat{\epsilon}\right) &= \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\int p_{\tilde{\epsilon}}(y-ax)p(x)e^{-2\pi ix\cdot\nu}dx}{p(y)\int p\left(\hat{n}\right)e^{-2\pi i(cy+d\hat{n})\cdot\nu}d\hat{n}}d\nu \\ &= \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\int \frac{1}{\sqrt{2\pi\left(b^2+1\right)}}e^{-\frac{(y-ax)^2}{2\left(b^2+1\right)}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}e^{-2\pi ix\cdot\nu}dx}{\frac{1}{\sqrt{2\pi\left(a^2+b^2+1\right)}}e^{-\frac{y^2}{2\left(a^2+b^2+1\right)}}e^{-2\pi icy\cdot\nu}\int p\left(\hat{n}\right)e^{-2\pi id\hat{n}\cdot\nu}d\hat{n}}d\nu \\ &= \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\int \frac{\sqrt{2\pi\left(a^2+b^2+1\right)}}{\sqrt{2\pi}\sqrt{2\pi\left(b^2+1\right)}}e^{-\frac{(y-ax)^2}{2\left(b^2+1\right)}-\frac{x^2}{2}+\frac{y^2}{2\left(a^2+b^2+1\right)}}e^{-2\pi ix\cdot\nu}dx}{e^{-2\pi icy\cdot\nu}\int p\left(\hat{n}\right)e^{-2\pi id\hat{n}\cdot\nu}d\hat{n}}d\nu \\ &= \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\int \frac{\sqrt{\left(a^2+b^2+1\right)}}{\sqrt{2\pi\left(b^2+1\right)}}e^{-\left(a^2+b^2+1\right)x^2-a^2y^2+2axy\left(a^2+b^2+1\right)}e^{-2\pi ix\cdot\nu}dx}{e^{-2\pi icy\cdot\nu}\int p\left(\hat{n}\right)e^{-2\pi id\hat{n}\cdot\nu}d\hat{n}}d\nu \\ &= \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\int \frac{\sqrt{\left(a^2+b^2+1\right)}}{\sqrt{2\pi\left(b^2+1\right)}}e^{-\frac{\left(\left(a^2+b^2+1\right)x-ay\right)^2}{2\left(b^2+1\right)\left(a^2+b^2+1\right)}}e^{-2\pi ix\cdot\nu}dx}{e^{-2\pi icy\cdot\nu}\int p\left(\hat{n}\right)e^{-2\pi id\hat{n}\cdot\nu}d\hat{n}}d\nu \\ &= \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{\int \frac{1}{\sqrt{2\pi\left(b^2+1\right)/\left(a^2+b^2+1\right)}}e^{-\frac{\left(x-\frac{a}{a^2+b^2+1}y\right)^2}{2\left(b^2+1\right)/\left(a^2+b^2+1\right)}}e^{-2\pi i\left(x-\frac{a}{a^2+b^2+1}y\right)\cdot\nu}e^{-2\pi i\frac{a}{a^2+b^2+1}y\cdot\nu}dx}{e^{-2\pi icy\cdot\nu}\int p\left(\hat{n}\right)e^{-2\pi id\hat{n}\cdot\nu}d\hat{n}}d\nu \\ &= \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{e^{-2\pi i\frac{a}{a^2+b^2+1}y\cdot\nu}e^{-2\pi^2\left(b^2+1\right)/\left(a^2+b^2+1\right)\cdot\nu^2}}{e^{-2\pi icy\cdot\nu}\int p\left(\hat{n}\right)e^{-2\pi id\hat{n}\cdot\nu}d\hat{n}}d\nu \\ &= \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{e^{-2\pi i\frac{a}{a^2+b^2+1}y\cdot\nu}e^{-2\pi^2\left(b^2+1\right)/\left(a^2+b^2+1\right)\cdot\nu^2}}{e^{-2\pi icy\cdot\nu}\int \frac{1}{\sqrt{2\pi d^2}}e^{-\frac{(d\hat{n})^2}{2d^2}}e^{-2\pi id\hat{n}\cdot\nu}d\left(d\hat{n}\right)}d\nu \\ &= \int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{e^{-2\pi i\frac{a}{a^2+b^2+1}y\cdot\nu}e^{-2\pi^2\left(b^2+1\right)/\left(a^2+b^2+1\right)\cdot\nu^2}}{e^{-2\pi icy\cdot\nu}e^{-2\pi^2d^2\cdot\nu^2}}d\nu. \end{aligned}$$

Let $c = \frac{a}{a^2+b^2+1}$, $d^2 = \frac{a^2}{a^2+b^2+1}$, we obtain

$$\begin{aligned} &\int e^{2\pi i\hat{\epsilon}\cdot\nu}\frac{e^{-2\pi i\frac{a}{a^2+b^2+1}y\cdot\nu}e^{-2\pi^2\left(b^2+1\right)/\left(a^2+b^2+1\right)\cdot\nu^2}}{e^{-2\pi icy\cdot\nu}e^{-2\pi^2d^2\cdot\nu^2}}d\nu \\ =\ & \int e^{2\pi i\hat{\epsilon}\cdot\nu-2\pi^2\cdot\nu^2}d\nu \\ =\ & \int e^{-\left(\sqrt{2}\pi\nu-\frac{i\hat{\epsilon}}{\sqrt{2}}\right)^2-\frac{\hat{\epsilon}^2}{2}}d\nu \\ =\ & \frac{1}{\sqrt{2\pi}}e^{-\frac{\hat{\epsilon}^2}{2}}. \end{aligned}$$

Thus, we have $p_{\hat{\epsilon}}\left(\hat{\epsilon}\right) = \frac{1}{\sqrt{2\pi}}e^{-\frac{\hat{\epsilon}^2}{2}}$, which is a Gaussian distribution and independent of $y, \hat{\mathbf{n}}$.□

E PROOF OF COROLLARY 2

Corollary 2. Suppose that there is no unmeasured intermediate noise in CANM, if the solution of Equation (10) exists, then for any 3-times differentiable triple (f, p_X, p_ϵ), it must satisfy the differential equation from ANM [9Theorem 1] for all x, y with $\nu''(y-f(x))f'(x) \neq 0$:

$$\xi''' = \xi''\left(-\frac{\nu'''f'}{\nu''} + \frac{f''}{f'}\right) - 2\nu''f''f' + \nu'f''' + \frac{\nu'\nu'''f''f'}{\nu''} - \frac{\nu'(f'')^2}{f'}, \tag{E.1}$$

where $\nu := \log p_\epsilon$, $\xi := \log p_X$

Proof. Since there are no unobserved intermediate noises, based on Theorem 1, we have

$$p_{\hat{\epsilon}}\left(\hat{\epsilon}\right) = e^{2\pi i(\hat{\epsilon}-g(y))\cdot\nu}\int \frac{p(x)p_\epsilon(y-f(x))}{p(y)}e^{-2\pi ix\cdot\nu}dx. \tag{E.2}$$

Let $\hat{\epsilon} = x - g(y)$, then based on the Fourier inverse theorem, the existence of the solution of Equation (E.2) is equivalent to the existence of following equation,

$$p_{\hat{\epsilon}}(x - g(y)) = \frac{p(x)p_{\epsilon}(y - f(x))}{p(y)}, \tag{E.3}$$

which is the standard identifiability for additive noise model, then applying the [9Theorem 1], the triple (f, p_x, p_ϵ) must satisfy the following differential equation for all x, y with $\nu''(y - f(x))f'(x) \neq 0$:

$$\xi''' = \xi''\left(-\frac{\nu'''f'}{\nu''} + \frac{f''}{f'}\right) - 2\nu''f''f' + \nu'f''' + \frac{\nu'\nu'''f''f'}{\nu''} - \frac{\nu'(f'')^2}{f'}.$$

□

F PROOF OF THEOREM 2

Theorem 2. Let $X \to Y$ follow the confounding cascade additive noise model, while there exists a backward model following the same form:

$$\begin{aligned} X &= g(\mathbf{Z_c}) + \epsilon_x \quad Y = f(X, \mathbf{N}, \mathbf{Z_c}) + \epsilon_y, \\ Y &= \hat{g}(\mathbf{Z_c}) + \hat{\epsilon}_y \quad X = \hat{f}(Y, \hat{\mathbf{N}}, \mathbf{Z_c}) + \hat{\epsilon}_x, \end{aligned} \tag{F.1}$$

where X is independent of $\mathbf{N}, \epsilon_y$ and Y is independent of $\hat{\mathbf{N}}, \hat{\epsilon}_x$ and $f, \hat{f}$ denote the functional class implied by confounding cascade process. Then the noises distribution on the anti-causal direction $p_{\hat{\epsilon}_x}, p_{\hat{\epsilon}_y}$ must be

$$p(\hat{\epsilon}_x) = \int e^{2\pi i\hat{\epsilon}_x\cdot\nu}\frac{\int p(x,y)e^{-2\pi ix\cdot\nu}dx}{\int p(\hat{\mathbf{n}})p(\hat{\mathbf{c}})p(\hat{\epsilon}_y = y - \hat{g}(\hat{\mathbf{c}}))e^{-2\pi i\hat{f}(y,\hat{\mathbf{n}},\hat{\mathbf{c}})\cdot\nu}d\hat{\mathbf{n}}d\hat{\mathbf{c}}}d\nu, \tag{F.2}$$

$$p(\hat{\epsilon}_y) = \int e^{2\pi i\hat{\epsilon}_y\cdot\mu}\frac{\int p(y)e^{-2\pi iy\cdot\mu}dy}{\int p(\hat{\mathbf{c}})e^{-2\pi i\hat{g}(\hat{\mathbf{c}})\cdot\mu}d\hat{\mathbf{c}}}d\mu \tag{F.3}$$

Proof. First recall that the marginal likelihood of CCANM can be given as follows:

$$p(x,y) = \int p(\mathbf{n})p(\mathbf{c})p(x - g(\mathbf{c}))p(y - f(x,\mathbf{n},\mathbf{c}))d\mathbf{n}d\mathbf{c} \tag{F.4}$$

$$p(x,y) = \int p(\hat{\mathbf{n}})p(\hat{\mathbf{c}})p(y - \hat{g}(\hat{\mathbf{c}}))p(x - \hat{f}(y,\hat{\mathbf{n}},\hat{\mathbf{c}}))d\hat{\mathbf{n}}d\hat{\mathbf{c}} \tag{F.5}$$

Applying the Fourier transform on $p(x, y)$ respect to x, we have

$$\begin{aligned} \mathcal{F}(\nu) &= \int p(x,y)e^{-2\pi ix\cdot\nu}dx \\ &= \int p(\hat{\mathbf{n}})p(\hat{\mathbf{c}})p(\hat{\epsilon}_y = y - \hat{g}(\mathbf{c}))p(\hat{\epsilon}_x = x - \hat{f}(y,\hat{\mathbf{n}},\hat{\mathbf{c}}))e^{-2\pi ix\cdot\nu}d\hat{\mathbf{n}}d\hat{\mathbf{c}}dx \end{aligned} \tag{F.6}$$

Since $\hat{\epsilon}_x = x - \hat{f}(y,\hat{\mathbf{n}},\hat{\mathbf{c}})$, we have $d\hat{\epsilon}_x = dx$. Then we have

$$\begin{aligned} \mathcal{F}(\nu) &= \int p(\hat{\mathbf{n}})p(\hat{\mathbf{c}})p(\hat{\epsilon}_y = y - \hat{g}(\hat{\mathbf{c}}))p(\hat{\epsilon}_x)e^{-2\pi i(\hat{\epsilon}_x + \hat{f}(y,\hat{\mathbf{n}},\hat{\mathbf{c}}))\cdot\nu}d\hat{\mathbf{n}}d\hat{\mathbf{c}}d\hat{\epsilon}_x \\ &= \int p(\hat{\mathbf{n}})p(\hat{\mathbf{c}})p(\hat{\epsilon}_y = y - \hat{g}(\hat{\mathbf{c}}))e^{-2\pi i\hat{f}(y,\hat{\mathbf{n}},\hat{\mathbf{c}})\cdot\nu}d\hat{\mathbf{n}}d\hat{\mathbf{c}}\int p(\hat{\epsilon}_x)e^{-2\pi i\hat{\epsilon}_x\cdot\nu}d\hat{\epsilon}_x \end{aligned} \tag{F.7}$$

Combining Equations (F.6) and (F.7), we have

$$\int p(\hat{\epsilon}_x)e^{-2\pi i\hat{\epsilon}_x\cdot\nu}d\hat{\epsilon}_x = \frac{\int p(x,y)e^{-2\pi ix\cdot\nu}dx}{\int p(\hat{\mathbf{n}})p(\hat{\mathbf{c}})p(\hat{\epsilon}_y = y - \hat{g}(\hat{\mathbf{c}}))e^{-2\pi i\hat{f}(y,\hat{\mathbf{n}},\hat{\mathbf{c}})\cdot\nu}d\hat{\mathbf{n}}d\hat{\mathbf{c}}} \tag{F.8}$$

Then, applying the inverse Fourier transform, the noise distribution $p(\hat{\epsilon}_x)$ on the reverse direction is given as follows:

$$p(\hat{\epsilon}_x) = \int e^{2\pi i\hat{\epsilon}_x\cdot\nu}\frac{\int p(x,y)e^{-2\pi ix\cdot\nu}dx}{\int p(\hat{\mathbf{n}})p(\hat{\mathbf{c}})p(\hat{\epsilon}_y = y - \hat{g}(\hat{\mathbf{c}}))e^{-2\pi i\hat{f}(y,\hat{\mathbf{n}},\hat{\mathbf{c}})\cdot\nu}d\hat{\mathbf{n}}d\hat{\mathbf{c}}}d\nu \tag{F.9}$$

Similarly, we can apply the Fourier transform respect to y as follows:

$$\begin{aligned} \mathcal{F}(\mu) &= \int p(y,\hat{\mathbf{c}})e^{-2\pi iy\cdot\mu}dyd\hat{\mathbf{c}} \\ &= \int p(\hat{\mathbf{c}})p(\hat{\epsilon}_y = y - \hat{g}(\hat{\mathbf{c}}))e^{-2\pi iy\cdot\mu}d\hat{\mathbf{c}}dy \end{aligned} \tag{F.10}$$

Since $\hat{\epsilon}_y = y - \hat{g}(\hat{\mathbf{c}})$, we have $d\hat{\epsilon}_y = dy$. Then we have

$$\begin{aligned} \mathcal{F}(\mu) &= \int p(\hat{\mathbf{c}})p(\hat{\epsilon}_y)e^{-2\pi i(\hat{\epsilon}_y + \hat{g}(\hat{\mathbf{c}}))\cdot\mu}d\hat{\mathbf{c}}d\hat{\epsilon}_y, \\ &= \int p(\hat{\mathbf{c}})e^{-2\pi i\hat{g}(\hat{\mathbf{c}})\cdot\mu}d\hat{\mathbf{c}}\int p(\hat{\epsilon}_y)e^{-2\pi i\hat{\epsilon}_y\cdot\mu}d\hat{\epsilon}_y. \end{aligned} \tag{F.11}$$

Then combining Equations (F.10) and (F.11), we have

$$\int e^{-2\pi i\hat{\epsilon}_y\cdot\mu}p(\hat{\epsilon}_y)d\hat{\epsilon}_y = \frac{\int p(y,\hat{\mathbf{c}})e^{-2\pi iy\cdot\mu}dyd\hat{\mathbf{c}}}{\int p(\hat{\mathbf{c}})e^{-2\pi i\hat{g}(\hat{\mathbf{c}})\cdot\mu}d\hat{\mathbf{c}}} = \frac{\int p(y)e^{-2\pi iy\cdot\mu}dy}{\int p(\hat{\mathbf{c}})e^{-2\pi i\hat{g}(\hat{\mathbf{c}})\cdot\mu}d\hat{\mathbf{c}}} \tag{F.12}$$

Finally, applying the inverse Fourier transfer, we conclude

$$p(\hat{\epsilon}_y) = \int e^{2\pi i\hat{\epsilon}_y\cdot\mu}\frac{\int p(y)e^{-2\pi iy\cdot\mu}dy}{\int p(\hat{\mathbf{c}})e^{-2\pi i\hat{g}(\hat{\mathbf{c}})\cdot\mu}d\hat{\mathbf{c}}}d\mu \tag{F.13}$$

□

G PROOF OF COROLLARY 3

Corollary 3. Assume that CCANM is linear Gaussian, i.e.,

$$X = aC + \epsilon_x \quad Y = bX + cN + dC + \epsilon_y,$$

where $C, N, \epsilon_x, \epsilon_y \sim \mathcal{N}(0,1)$, then there exists a backward CCANM

$$Y = \hat{a}\hat{C} + \hat{\epsilon}_y \quad X = eY + f\hat{N} + g\hat{C} + \hat{\epsilon}_x,$$

where $\hat{N}, \hat{C} \sim \mathcal{N}(0,1)$, $\hat{\epsilon}_x \sim \mathcal{N}(0, \sigma^2_{\hat{\epsilon}_x})$, $\hat{\epsilon}_y \sim \mathcal{N}(0, \sigma^2_{\hat{\epsilon}_y})$ and $\hat{\epsilon}_x, \hat{\epsilon}_y$ are independent of Y and $\hat{C}$ if the following four conditions hold:

$$\begin{aligned} &\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y} + \hat{a}^2\right)} + e - \frac{b(a^2+1)}{(c^2+d^2+a^2+2)} = 0 \\ &\left(\frac{(a^2+1)}{(c^2+d^2+a^2+2)}\right)^2 - \frac{(a^2+1)}{(c^2+d^2+a^2+2)} - \frac{\hat{a}^2}{2\sigma^2_{\hat{\epsilon}_y}\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)} + \frac{1}{2\sigma^2_{\hat{\epsilon}_y}} = 0, \\ &\frac{z^2\sigma^2_{\hat{\epsilon}_y}\left(c^2+d^2+a^2+2\right)}{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2} = \frac{\left(a^2+1\right)\left(c^2+d^2+1\right)}{c^2+d^2+a^2+2} + f^2 + \frac{g^2\sigma^2_{\hat{\epsilon}_y}}{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2}, \\ &(b^2(a^2+1) + c^2 + d^2 + 1) - \hat{a}^2 = \sigma^2_{\hat{\epsilon}_y}. \end{aligned} \tag{G.1}$$

Proof. Based on Theorem 2, the noise distribution on the reverse direction can be expressed as:

$$p(\hat{\epsilon}_x) = \int e^{2\pi i\hat{\epsilon}_x\cdot\nu}\frac{\int p(x,y)e^{-2\pi ix\cdot\nu}dx}{\int p(\hat{\mathbf{n}})p(\hat{\mathbf{c}})p(\hat{\epsilon}_y = y - \hat{g}(\hat{\mathbf{c}}))e^{-2\pi i\hat{f}(y,\hat{\mathbf{n}},\hat{\mathbf{c}})\cdot\nu}d\hat{\mathbf{n}}d\hat{\mathbf{c}}}d\nu, \tag{G.2}$$

$$p(\hat{\epsilon}_y) = \int e^{2\pi i\hat{\epsilon}_y\cdot\mu}\frac{\int p(y)e^{-2\pi iy\cdot\mu}dy}{\int p(\hat{\mathbf{c}})e^{-2\pi i\hat{g}(\hat{\mathbf{c}})\cdot\mu}d\hat{\mathbf{c}}}d\mu \tag{G.3}$$

In the following, we will frequently use the formula on Fourier transformation on Gaussian distribution:

$$\mathcal{F}_x\left[\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\right](\nu) = e^{-2\pi i\mu\cdot\nu}e^{-2\pi^2\sigma^2\cdot\nu^2}. \tag{G.4}$$

Due to the linear relation, we have $\hat{f}(y,\hat{\mathbf{n}},\hat{\mathbf{c}}) = eY + f\hat{n} + gc$, $\hat{g}(\hat{\mathbf{c}}) = \hat{a}\hat{\mathbf{c}}$, then we have

$$p(\hat{\epsilon}_x) = \int e^{2\pi i\hat{\epsilon}_x\cdot\nu}\frac{\int p(x,y)e^{-2\pi ix\cdot\nu}dx}{\int p(\hat{\mathbf{n}})e^{-2\pi if\hat{n}\cdot\nu}d\hat{\mathbf{n}}\int p(\hat{\mathbf{c}})e^{-2\pi ig\hat{\mathbf{c}}\cdot\nu}p(\hat{\epsilon}_y = y - \hat{a}\hat{\mathbf{c}})e^{-2\pi iey\cdot\nu}d\hat{\mathbf{c}}}d\nu, \tag{G.5}$$

(G.6)

$$p(\hat{\epsilon}_y) = p(\hat{\epsilon}_y) = \int e^{2\pi i\hat{\epsilon}_y\cdot\mu}\frac{\int p(y)e^{-2\pi iy\cdot\mu}dy}{\int p(\hat{\mathbf{c}})e^{-2\pi i\hat{a}\hat{\mathbf{c}}\cdot\mu}d\hat{\mathbf{c}}}d\mu.$$

For the term $\int p(x,y)e^{-2\pi ix\cdot\nu}dx$, due to that $p(x)\sim N(0,\sqrt{a^2+1}^2)$, $p(y|x)\sim N(0,\sqrt{c^2+d^2+1}^2)$, we have

$$\begin{aligned}
&\int p(x,y)e^{-2\pi ix\cdot\nu}dx\\
=\ &\frac{1}{\sqrt{2\pi}}\int\frac{1}{\sqrt{2\pi\left(a^2+1\right)\left(c^2+d^2+1\right)}}e^{-\frac{1}{2\left(a^2+1\right)}x^2-\frac{1}{2\left(c^2+d^2+1\right)}(y-bx)^2}e^{-2\pi ix\cdot\nu}dx\\
=\ &\frac{1}{\sqrt{2\pi}}\int\frac{1}{\sqrt{2\pi\left(a^2+1\right)\left(c^2+d^2+1\right)}}e^{-\frac{1}{2\left(a^2+1\right)}x^2-\frac{y^2-2bxy+x^2}{2\left(c^2+d^2+1\right)}}e^{-2\pi ix\cdot\nu}dx\\
=\ &\frac{1}{\sqrt{2\pi}}\int\frac{1}{\sqrt{2\pi\left(a^2+1\right)\left(c^2+d^2+1\right)}}e^{-\frac{\left(c^2+d^2+1\right)x^2+\left(a^2+1\right)\left(y^2-2bxy+x^2\right)}{2\left(a^2+1\right)\left(c^2+d^2+1\right)}}e^{-2\pi ix\cdot\nu}dx\\
=\ &\frac{1}{\sqrt{2\pi}}\int\frac{1}{\sqrt{2\pi\left(a^2+1\right)\left(c^2+d^2+1\right)}}e^{-\frac{\left(c^2+d^2+a^2+2\right)x^2-2b\left(a^2+1\right)xy+\left(a^2+1\right)y^2}{2\left(a^2+1\right)\left(c^2+d^2+1\right)}}e^{-2\pi ix\cdot\nu}dx\\
=\ &\frac{1}{\sqrt{2\pi}}\int\frac{1}{\sqrt{2\pi\left(a^2+1\right)\left(c^2+d^2+1\right)}}ev^{-\frac{x^2-\frac{2b\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}xy+\frac{\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}y^2}{2\left(a^2+1\right)\left(c^2+d^2+1\right)/\left(c^2+d^2+a^2+2\right)}}e^{-2\pi ix\cdot\nu}dx\\
=\ &\frac{1}{\sqrt{2\pi}}\int\frac{1}{\sqrt{2\pi\left(a^2+1\right)\left(c^2+d^2+1\right)}}e^{-\frac{x^2-\frac{2b\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}xy+\left(\frac{\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}\right)^2y^2-\left(\frac{\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}\right)^2y^2+\frac{\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}y^2}{2\left(a^2+1\right)\left(c^2+d^2+1\right)/\left(c^2+d^2+a^2+2\right)}}e^{-2\pi ix\cdot\nu}dx\\
=\ &\frac{e^{\left(\frac{\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}\right)^2y^2-\frac{\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}y^2}}{\sqrt{2\pi\left(c^2+d^2+a^2+2\right)}}\int\frac{e^{-\frac{\left(x-\frac{b\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}y\right)^2}{2\left(a^2+1\right)\left(c^2+d^2+1\right)/\left(c^2+d^2+a^2+2\right)}}}{\sqrt{2\pi\left(a^2+1\right)\left(c^2+d^2+1\right)/\left(c^2+d^2+a^2+2\right)}}e^{-2\pi ix\cdot\nu}dx\\
=\ &\frac{e^{\left(\frac{\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}\right)^2y^2-\frac{\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}y^2}}{\sqrt{2\pi\left(c^2+d^2+a^2+2\right)}}e^{-2\pi i\frac{b\left(a^2+1\right)y}{c^2+d^2+a^2+2}\cdot\nu}e^{-2\pi^2\frac{\left(a^2+1\right)\left(c^2+d^2+1\right)}{c^2+d^2+a^2+2}\cdot\nu^2}
\end{aligned}$$

Because, $p(\hat{\mathbf{c}}),p(\hat{\mathbf{n}})$ follow the standard Gaussian distribution. Then, for the terms $\int p(\hat{\mathbf{n}})e^{-2\pi if\hat{n}\cdot\nu}d\hat{\mathbf{n}}$ and $\int p(\hat{\mathbf{c}})e^{-2\pi i\hat{g}(\hat{\mathbf{c}})\cdot\mu}d\hat{\mathbf{c}}$, by using Equation (G.4), we have

(G.8)

$$\begin{aligned}
\int p(\hat{\mathbf{n}})e^{-2\pi if\hat{n}\cdot\nu}d\hat{\mathbf{n}} &= \int\frac{1}{\sqrt{2\pi f^2}}e^{-\frac{(f\hat{\mathbf{n}})^2}{2f^2}}e^{-2\pi if\hat{\mathbf{n}}\cdot\nu}d(f\hat{\mathbf{n}}) &&= e^{-2\pi^2f^2\cdot\nu^2}\\
\int p(\hat{\mathbf{c}})e^{-2\pi i\hat{a}\hat{\mathbf{c}}\cdot\mu}d\hat{\mathbf{c}} &= \int\frac{1}{\sqrt{2\pi\hat{a}^2}}e^{-\frac{(\hat{a}\hat{\mathbf{c}})^2}{2\hat{a}^2}}e^{-2\pi i\hat{a}\hat{\mathbf{c}}\cdot\nu}d(\hat{a}\hat{\mathbf{c}}) &&= e^{-2\pi^2\hat{a}^2\cdot\nu^2}.
\end{aligned}$$

Here, we assume $\hat{\epsilon}_y\sim N(0,\sigma^2_{\hat{\epsilon}_y})$. Then, for the term $\int p(\hat{\mathbf{c}})e^{-2\pi ig\hat{\mathbf{c}}\cdot\nu}p(\hat{\epsilon}_y=y-\hat{a}\hat{\mathbf{c}})d\hat{\mathbf{c}}$, we have

$$\begin{aligned}
&\int p(\hat{\mathbf{c}})e^{-2\pi ig\hat{\mathbf{c}}\cdot\nu}p(\hat{\epsilon}_y=y-\hat{a}\hat{\mathbf{c}})d\hat{\mathbf{c}}\\
=\ &\frac{1}{\sqrt{2\pi\sigma^2_{\hat{\epsilon}_y}}}\int\frac{1}{\sqrt{2\pi}}e^{-\frac{\hat{\mathbf{c}}^2}{2}}e^{-\frac{(y-\hat{a}\hat{\mathbf{c}})^2}{2\sigma^2_{\hat{\epsilon}_y}}}e^{-2\pi ig\hat{\mathbf{c}}\cdot\nu}d\hat{\mathbf{c}}\\
=\ &\frac{1}{\sqrt{2\pi\sigma^2_{\hat{\epsilon}_y}}}\int\frac{1}{\sqrt{2\pi}}e^{-\frac{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)\hat{\mathbf{c}}^2-2\hat{a}\hat{\mathbf{c}}y+y^2}{2\sigma^2_{\hat{\epsilon}_y}}}e^{-2\pi ig\hat{\mathbf{c}}\cdot\nu}d\hat{\mathbf{c}}
\end{aligned}$$

(G.9)

$$\begin{aligned}
=\ &\frac{1}{\sqrt{2\pi\sigma^2_{\hat{\epsilon}_y}}}\int\frac{1}{\sqrt{2\pi}}e^{-\frac{\hat{\mathbf{c}}^2-\frac{2\hat{a}\hat{\mathbf{c}}y}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}+\frac{y^2}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}{2\sigma^2_{\hat{\epsilon}_y}/\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}e^{-2\pi ig\hat{\mathbf{c}}\cdot\nu}d\hat{\mathbf{c}}\\
=\ &\frac{1}{\sqrt{2\pi\sigma^2_{\hat{\epsilon}_y}}}\int\frac{1}{\sqrt{2\pi}}e^{-\frac{g^2\hat{\mathbf{c}}^2-\frac{2\hat{a}\hat{\mathbf{c}}g^2y}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}+\frac{g^2y^2}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}{2g^2\sigma^2_{\hat{\epsilon}_y}/\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}e^{-2\pi ig\hat{\mathbf{c}}\cdot\nu}d\hat{\mathbf{c}}\\
=\ &\frac{1}{\sqrt{2\pi\sigma^2_{\hat{\epsilon}_y}}}\int\frac{1}{\sqrt{2\pi}}e^{-\frac{g^2\hat{\mathbf{c}}^2-\frac{2\hat{a}\hat{\mathbf{c}}g^2y}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}+\left(\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}\right)^2y^2-\left(\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}\right)^2y^2+\frac{g^2y^2}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}{2g^2\sigma^2_{\hat{\epsilon}_y}/\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}e^{-2\pi ig\hat{\mathbf{c}}\cdot\nu}d\hat{\mathbf{c}}\\
=\ &\frac{e^{\frac{\left(\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}\right)^2y^2-\frac{g^2y^2}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}{2g^2\sigma^2_{\hat{\epsilon}_y}/\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}}{g\sqrt{2\pi}}\int\frac{1}{\sqrt{2\pi}}e^{-\frac{\left(g\hat{\mathbf{c}}-\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}y\right)^2}{2g^2\sigma^2_{\hat{\epsilon}_y}/\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}e^{-2\pi ig\hat{\mathbf{c}}\cdot\nu}d(g\hat{\mathbf{c}})\\
=\ &\frac{e^{\frac{\left(\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}\right)^2y^2-\frac{g^2y^2}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}{2g^2\sigma^2_{\hat{\epsilon}_y}/\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}}{g\sqrt{2\pi}}\sqrt{\frac{g^2\sigma^2_{\hat{\epsilon}_y}}{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2}}\int\frac{1}{\sqrt{2\pi\frac{g^2\sigma^2_{\hat{\epsilon}_y}}{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2}}}e^{-\frac{\left(g\hat{\mathbf{c}}-\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}y\right)^2}{2g^2\sigma^2_{\hat{\epsilon}_y}/\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}e^{-2\pi ig\hat{\mathbf{c}}\cdot\nu}dg\hat{\mathbf{c}}\\
=\ &\frac{\sigma_{\hat{\epsilon}_y}}{\sqrt{2\pi\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}e^{\frac{\left(\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}\right)^2y^2-\frac{g^2y^2}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}{2g^2\sigma^2_{\hat{\epsilon}_y}/\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}\int\frac{1}{\sqrt{2\pi\frac{g^2\sigma^2_{\hat{\epsilon}_y}}{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2}}}e^{-\frac{\left(g\hat{\mathbf{c}}-\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}y\right)^2}{2g^2\sigma^2_{\hat{\epsilon}_y}/\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}e^{-2\pi ig\hat{\mathbf{c}}\cdot\nu}dg\hat{\mathbf{c}}
\end{aligned}$$

$$=\frac{\sigma_{\hat{\epsilon}_y}}{\sqrt{2\pi\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}e^{\frac{\left(\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}\right)^2y^2-\frac{g^2y^2}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}{2g^2\sigma^2_{\hat{\epsilon}_y}/\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}e^{-2\pi i\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}y\cdot\nu}e^{-2\pi^2\frac{g^2\sigma^2_{\hat{\epsilon}_y}}{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2}\cdot\nu^2}$$

Combining Equations (G.7), (G.8), (G.9), we have

$$\begin{aligned}
p(\hat{\epsilon}_x) &= \int e^{2\pi i\hat{\epsilon}_x\cdot\nu}\frac{\int p(x,y)e^{-2\pi ix\cdot\nu}dx}{\int p(\hat{\mathbf{n}})p(\hat{\mathbf{c}})p(\hat{\epsilon}_y=y-\hat{a}\hat{\mathbf{c}})e^{-2\pi i(ey+f\hat{n}+g\hat{\mathbf{c}})\cdot\nu}d\hat{\mathbf{n}}d\hat{\mathbf{c}}}d\nu\\
&= \int e^{2\pi i\hat{\epsilon}_x\cdot\nu}\frac{\int p(x,y)e^{-2\pi ix\cdot\nu}dx}{\int p(\hat{\mathbf{n}})e^{-2\pi if\hat{n}\cdot\nu}d\hat{\mathbf{n}}\int p(\hat{\mathbf{c}})e^{-2\pi ig\hat{\mathbf{c}}\cdot\nu}p(\hat{\epsilon}_y=y-\hat{a}\hat{\mathbf{c}})e^{-2\pi iey\cdot\nu}d\hat{\mathbf{c}}}d\nu
\end{aligned}$$

((

$$\begin{aligned}
=\ &\int e^{2\pi i\hat{\epsilon}_x\cdot\nu}\frac{\frac{e^{\left(\frac{\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}\right)^2y^2-\frac{\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}y^2}}{\sqrt{2\pi\left(c^2+d^2+a^2+2\right)}}e^{-2\pi i\frac{b\left(a^2+1\right)y}{\left(c^2+d^2+a^2+2\right)}\cdot\nu}e^{-2\pi^2\frac{\left(a^2+1\right)\left(c^2+d^2+1\right)}{c^2+d^2+a^2+2}\cdot\nu^2}}{e^{-2\pi^2f^2\cdot\nu^2}\left(\frac{\sigma_{\hat{\epsilon}_y}}{\sqrt{2\pi\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}e^{\frac{\left(\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}\right)^2y^2-\frac{g^2y^2}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}{2g^2\sigma^2_{\hat{\epsilon}_y}/\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}e^{-2\pi i\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}y\cdot\nu}e^{-2\pi^2\frac{g^2\sigma^2_{\hat{\epsilon}_y}}{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2}\cdot\nu^2}\right)e^{-2\pi iey\cdot\nu}}dv\\
=\ &\frac{\sqrt{2\pi\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}{\sigma_{\hat{\epsilon}_y}\sqrt{2\pi\left(c^2+d^2+a^2+2\right)}}\int e^{2\pi i\hat{\epsilon}_x\cdot\nu}\frac{e^{\left(\frac{\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}\right)^2y^2-\frac{\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}y^2-2\pi i\frac{b\left(a^2+1\right)}{\left(c^2+d^2+a^2+2\right)}y\cdot\nu-\frac{2\pi^2\left(a^2+1\right)\left(c^2+d^2+1\right)}{c^2+d^2+a^2+2}\cdot\nu^2}}{e^{-2\pi^2f^2\cdot\nu^2+\frac{\left(\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}\right)^2y^2-\frac{g^2y^2}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}}{2g^2\sigma^2_{\hat{\epsilon}_y}/\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}-2\pi i\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}y\cdot\nu-2\pi^2\frac{g^2\sigma^2_{\hat{\epsilon}_y}}{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2}\cdot\nu^2-2\pi iey\cdot\nu}}dv
\end{aligned}$$

To ensure $\hat{\epsilon}_x$ is independent to Y, we choose the coefficient to make sure y^2 and y vanish in Equation (G.10). Thus, setting

(G.11)

$$\frac{\hat{a}g}{\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}+e-\frac{b(a^2+1)}{(c^2+d^2+a^2+2)}=0$$

(G.12)

$$\left(\frac{(a^2+1)}{(c^2+d^2+a^2+2)}\right)^2-\frac{(a^2+1)}{(c^2+d^2+a^2+2)}-\frac{\hat{a}^2}{2\sigma^2_{\hat{\epsilon}_y}\left(\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2\right)}+\frac{1}{2\sigma^2_{\hat{\epsilon}_y}}=0,$$

we have

(G.13)

$$p(\hat{\epsilon}_x)=\frac{\sqrt{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2}}{\sigma_{\hat{\epsilon}_y}\sqrt{c^2+d^2+a^2+2}}\int e^{2\pi i\hat{\epsilon}_x\cdot\nu}\frac{e^{-2\pi^2\frac{\left(a^2+1\right)\left(c^2+d^2+1\right)}{c^2+d^2+a^2+2}\cdot\nu^2}}{e^{-2\pi^2f^2\cdot\nu^2-2\pi^2\frac{g^2\sigma^2_{\hat{\epsilon}_y}}{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2}\cdot\nu^2}}dv$$

To make sure $p(\hat{\epsilon}_x)$ is a valid distribution, we first let $\frac{(a^2+1)(c^2+d^2+1)}{c^2+d^2+a^2+2}+f^2+\frac{g^2\sigma^2_{\hat{\epsilon}_y}}{\sigma^2_{\hat{\epsilon}_y}+\hat{a}^2}=z^2$, and we have

$$\begin{aligned}
p(\hat{\epsilon}_x) &= \frac{\sqrt{\sigma_{\hat{\epsilon}_y}^2 + \hat{a}^2}}{\sigma_{\hat{\epsilon}_y}\sqrt{c^2+d^2+a^2+2}} \int e^{2\pi i\hat{\epsilon}_x\cdot\nu} e^{-2\pi^2 z^2\cdot\nu^2} dv \\
&= \frac{\sqrt{\sigma_{\hat{\epsilon}_y}^2 + \hat{a}^2}}{\sigma_{\hat{\epsilon}_y}\sqrt{c^2+d^2+a^2+2}} \int e^{-\left(\sqrt{2}\pi z\nu - \frac{i\hat{\epsilon}_x}{z\sqrt{2}}\right)^2 - \frac{\hat{\epsilon}_x^2}{2z^2}} d\nu \\
&= \frac{\sqrt{\sigma_{\hat{\epsilon}_y}^2 + \hat{a}^2}}{\sigma_{\hat{\epsilon}_y}\sqrt{c^2+d^2+a^2+2}} \frac{1}{z\sqrt{2}\pi} e^{-\frac{\hat{\epsilon}_x^2}{2z^2}} \int e^{-\left(\sqrt{2}\pi z\nu - \frac{i\hat{\epsilon}_x}{z\sqrt{2}}\right)^2} d\left(\sqrt{2}\pi z\nu - \frac{i\hat{\epsilon}_x}{z\sqrt{2}}\right) \\
&= \frac{\sqrt{\sigma_{\hat{\epsilon}_y}^2 + \hat{a}^2}}{\sigma_{\hat{\epsilon}_y}\sqrt{c^2+d^2+a^2+2}} \frac{1}{z\sqrt{2}\pi} e^{-\frac{\hat{\epsilon}_x^2}{2z^2}} \sqrt{\pi} \\
&= \frac{1}{\sqrt{2\pi\frac{z^2\sigma_{\hat{\epsilon}_y}^2(c^2+d^2+a^2+2)}{\sigma_{\hat{\epsilon}_y}^2+\hat{a}^2}}} e^{-\frac{\hat{\epsilon}_x^2}{2z^2}}.
\end{aligned} \tag{G.14}$$

Then $p(\hat{\epsilon}_x)$ is a valid distribution if the following equation holds:

$$\frac{z^2\sigma_{\hat{\epsilon}_y}^2(c^2+d^2+a^2+2)}{\sigma_{\hat{\epsilon}_y}^2+\hat{a}^2} = z^2 = \frac{(a^2+1)(c^2+d^2+1)}{c^2+d^2+a^2+2} + f^2 + \frac{g^2\sigma_{\hat{\epsilon}_y}^2}{\sigma_{\hat{\epsilon}_y}^2+\hat{a}^2}. \tag{G.15}$$

Similarly, based on Equation (G.8), for $p(\hat{\epsilon}_y)$, we have

$$\begin{aligned}
p(\hat{\epsilon}_y) &= \int e^{2\pi i\hat{\epsilon}_y\cdot\mu} \frac{\int p(y)e^{-2\pi iy\cdot\mu}dy}{e^{-2\pi^2\hat{a}^2\cdot\nu^2}} d\mu. \\
&= \int e^{2\pi i\hat{\epsilon}_y\cdot\mu} \frac{\int \frac{1}{\sqrt{2\pi\left(b^2\left(a^2+1\right)+c^2+d^2+1\right)}} e^{-\frac{y^2}{2\left(b^2\left(a^2+1\right)+c^2+d^2+1\right)}} e^{-2\pi iy\cdot\mu}dy}{e^{-2\pi^2\hat{a}^2\cdot\nu^2}} d\mu \\
&= \int e^{2\pi i\hat{\epsilon}_y\cdot\mu} \frac{e^{-2\pi^2\left(b^2\left(a^2+1\right)+c^2+d^2+1\right)\cdot\nu^2}}{e^{-2\pi^2\hat{a}^2\cdot\nu^2}} d\mu.
\end{aligned} \tag{G.16}$$

To make sure $p(\hat{\epsilon}_y)$ is a valid distribution, we set $(b^2(a^2+1)+c^2+d^2+1)-\hat{a}^2 = \sigma_{\hat{\epsilon}_y}^2$, and we have

$$\begin{aligned}
p(\hat{\epsilon}_y) &= \int e^{2\pi i\hat{\epsilon}_y\cdot\mu} \frac{e^{-2\pi^2\left(b^2\left(a^2+1\right)+c^2+d^2+1\right)\cdot\nu^2}}{e^{-2\pi^2\hat{a}^2\cdot\nu^2}} d\mu \\
&= \int e^{2\pi i\hat{\epsilon}_y\cdot\nu - 2\pi^2\sigma_{\hat{\epsilon}_y}^2\cdot\nu^2} d\nu \\
&= \frac{1}{\sqrt{2\pi\sigma_{\hat{\epsilon}_y}^2}} e^{-\frac{\hat{\epsilon}_x^2}{2\sigma_{\hat{\epsilon}_y}^2}}.
\end{aligned} \tag{G.17}$$

Therefore, we conclude that the reverse direction exists if the four conditions in Equation (G.1) hold.□

ACKNOWLEDGMENTS

We appreciate the comments from anonymous reviewers, which greatly helped to improve the paper.

REFERENCES

[1] Ronald Newbold Bracewell. 1986. *The Fourier Transform and its Applications*. Vol. 31999. McGraw-Hill, New York. Navigate to

[2] Peter Bühlmann, Jonas Peters, Jan Ernest, et al. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics* 42, 6 (2014), 2526–2556. Navigate to

[3] Ruichu Cai, Jie Qiao, Kun Zhang, Zhenjie Zhang, and Zhifeng Hao. 2019. Causal discovery with cascade nonlinear additive noise model. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. 1609–1615. Navigate to

[4] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794. Navigate to

[5] Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. 2012. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics* (2012), 294–321. Navigate to

[6] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in Genetics* 10 (2019), 524. Navigate to

[7] Arthur Gretton, Kenji Fukumizu, Choon H. Teo, Le Song, Bernhard Schölkopf, and Alex J. Smola. 2008. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*. 585–592. Navigate to

[8] Christina Heinze-Deml, Marloes H. Maathius, and Nicolai Meinshausen. 2018. Causal structure learning. *Annual Review of Statistics and Its Application* 8 (2018). Navigate to

[9] Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. 2009. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*. 689–696. Navigate to

[10] Aapo Hyvarinen and Hiroshi Morioka. 2016. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. Navigate to

[11] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. 2019. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 859–868. Navigate to

[12] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. 2012. Information-geometric approach to inferring causal directions. *Artificial Intelligence* 182 (2012), 1–31. Navigate to

[13] Dominik Janzing, Jonas Peters, Joris Mooij, and Bernhard Schölkopf. 2009. Identifying confounders using additive noise models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 249–257. Navigate to

[14] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

[15] Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014)*. Navigate to

[16] Murat Kocaoglu, Sanjay Shakkottai, Alexandros G. Dimakis, Constantine Caramanis, and Sriram Vishwanath. 2018. Entropic latent variable discovery. *arXiv preprint arXiv:1807.10399* (2018). Navigate to

[17] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. 2016. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research* 17, 32 (2016), 1–102. http://jmlr.org/papers/volume17/14-518/14-518.pdf (http://jmlr.org/papers/volume17/14-518/14-518.pdf). Navigate to

[18] S. D. Prestwich, S. A. Tarim, and I. Ozkan. 2016. Causal discovery by randomness test. In *Proceedings of the 14th International Symposium on Artificial Intelligence and Mathematics*. Navigate to

[19] Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, et al. 2019. Inferring causation from time series in Earth system sciences. *Nature Communications* 10, 1 (2019), 1–13. Navigate to

[20] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7, Oct (2006), 2003–2030. Navigate to

[21] Peter Spirtes, Clark N. Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. MIT Press. Navigate to

[22] Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics* 3, 1 (18 Feb 2016), 3. Navigate to

[23] Kui Yu, Lin Liu, Jiuyong Li, and Huanhuan Chen. 2018. Mining Markov blankets without causal sufficiency. *IEEE Transactions on Neural Networks and Learning Systems*99 (2018), 1–15. Navigate to

[24] K. Zhang and A. Hyvärinen. 2009. On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*. AUAI Press, 647–655. Navigate to

[25] K. Zhang, B. Schölkopf, and D. Janzing. 2010. Invariant Gaussian process latent variable models and application in causal discovery. In *26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*. AUAI Press, 717–724. Navigate to

[26] Kun Zhang, Zhikun Wang, Jiji Zhang, and Bernhard Schölkopf. 2015. On estimation of functional causal models: General results and application to the post-nonlinear causal model. *ACM Trans. Intell. Syst. Technol.* 7, 2, Article 13 (Dec. 2015), 22 pages. Navigate to

Footnotes

[1] https://github.com/DMIRLAB-Group/CANM **(https://github.com/DMIRLAB-Group/CANM)**.

[2] https://webdav.tuebingen.mpg.de/cause-effect/ **(https://webdav.tuebingen.mpg.de/cause-effect/)**.

This research was supported in part by NSFC-Guangdong Joint Found (U1501254), Natural Science Foundation of China (61876043), Natural Science Foundation of Guangdong (2014A030306004, 2014A030308008), Guangdong High-level Personnel of Special Support Program (2015TQ01X140), Pearl River S&T Nova Program of Guangzhou (201610010101), and Guangdong Provincial Science and Technology Innovation Strategy Fund (019B121203012). KZ would like to acknowledge the support by National Institutes of Health (NIH) under Contract No. NIH-1R01EB022858-01, FAINR01EB022858, NIH-1R01LM012087, NIH-5U54HG008540-02, and FAIN- U54HG008540, by the United States Air Force under Contract No. FA8650-17-C-7715, and by National Science Foundation (NSF) EAGER Grant No. IIS-1829681. The NIH, the U.S. Air Force, and the NSF are not responsible for the views reported here.

Authors' addresses: J. Qiao, School of Computer Science, Guangdong University of Technology, Guangzhou, Guangdong, China, 510006; email: qiaojie.chn@gmail.com **(mailto:qiaojie.chn@gmail.com)**; R. Cai (corresponding author), School of Computer Science, Guangdong University of Technology, Guangzhou, Guangdong, China, 510006 and Guangdong Provincial Key Laboratory of Public Finance and Taxation with Big Data Application, Guangzhou, Guangdong, China, 510320; email: cairuichu@gdut.edu.cn **(mailto:cairuichu@gdut.edu.cn)**; K. Zhang, Department of philosophy, Carnegie Mellon University, Pittsburgh, PA, USA, 15213; Z. Hao, College of Science, Shantou University, Shantou, Guangdong, China, 515063; email: kunz1@cmu.edu **(mailto:kunz1@cmu.edu)**; Z. Zhang, PVoice Technology, Singapore; email: zhenjie.zhang@pvoice.io **(mailto:zhenjie.zhang@pvoice.io)**; Z. Hao, College of Science, Shantou University, Shantou, Guangdong, China, 515063; email: zfhao@gdut.edu.cn **(mailto:zfhao@gdut.edu.cn)**.

2157-6904/2021/11-ART80 $15.00
DOI: https://doi.org/10.1145/3482879 **(https://doi.org/10.1145/3482879)**

Publication History: Received December 2020; revised July 2021; accepted August 2021