Identification of Possible Common Causes by Intrinsic Dimension Estimation

Jing Song
Hokkaido University
songjing@complex.ist.hokudai.ac.jp

Satoshi Oyama
Hokkaido University
oyama@ist.hokudai.ac.jp

Masahito Kurihara
Hokkaido University
kurihara@ist.hokudai.ac.jp

***Abstract*—The effect of confounding factors cannot be ignored in real world causal discovery tasks. A common cause is a general confounder between two variables. In this paper, we propose using intrinsic dimension estimation as a necessary condition to determine a possible common cause for two variables. Simulated application showed that the proposed method worked well for both linear and non-linear functions. Testing using different types of noise showed that it generally worked well for different types of added noise. In particular, it worked better than a kernel-based conditional independence test for Poisson noise. Testing of how the estimated intrinsic dimension is affected by different types of distributions showed that the estimated dimension is nearly not affected by the type of distribution. Simulation of mixed pattern showed that the proposed method can still tell a possible common cause when it is mixed with causal relationship. Finally, experiments using variables from the CauseEffectPairs dataset showed that the proposed method can give correct inferred results for real world data.**

***Index Terms*—common cause identification, intrinsic dimension, conditional independence test**

I. Introduction

Confounding may exist in causal analysis of real world data. Stratification or matching can be used to deal with confounding in causal analysis [1], [2]. Stratification was used to fix the value of confounders in ceitain subgroups through which the effect of confounders can be reduced or vanishes [3]. Matching can be viewed as a special case of stratification and is frequently used in case-control studies [4]. Stratification or matching variable should be carefully chose since bad choice of it would induce extra errors [4]. Some causal relationships have been analyzed using stratification. [5] analyzed the causal relationship between energy consumption and GDP growth by dividing the data into four income categories. Their experimental results showed that causal relationships differed among the four categories. Coondoo and Dinda analyzed the causal relationship between income and emissions by dividing countries into specific groups [6]. The experimental results showed that the causal relationships varied among the groups as well. The above experiments showed that the effects of confounders should be considered when doing causal analysis since the analysis result may change after considering certain confounder.

A common cause (Figure 1) or selection bias (Figure 2) can induce spurious correlation between two variables. If the data was unbiased, the possible confounding comes from unconsidered common cause. [7] discussed the responses of several causal discovery models to confounding which showed that some models can avoid the effect of confounding to a certain degree and some cannot. To correctly analyze causal relationship between two variables, finding and testing possible observable confounders is needed. Conditional independence is a necessary condition for a common cause. To test whether a third variable Z is a possible confounder for two variables X and Y, a conditional independence test can be conducted. [8] proposed a framework for crowd-based causal analysis of open data in which conditional independence test was used as an initial step to filter unrelated words. The remained related words were then used to further learn causal relationship between variables. To filter unrelated words, methods like conditional independence test are needed to identify possible common causes. However, many methods for conditional independence test have certain assumptions of data distributions, function forms or additive noise. When these assumptions are not satisfied, the conditional independence test may fail. Besides, causal relationship may be mixed with common causes in the real world problem. In this case, conditional independence test will fail to tell a possible common cause. Here we propose using intrinsic dimension estimation as an alternative to detect a possible common cause for two variables. The proposed method is a necessary condition and can be used to filter unrelated words in the candidates for confounders. It can be combined with other causal dicovery methods to realize causal relationship analysis while considering specific confounders. The proposed method generally worked well for different types of added noise. In particular, it worked better than a kernel-based conditional independence test for Poisson noise. Besides, the proposed method can infer a possible common cause when it is mixed with causal relationship.

In Section II, we introduce related machine learning methods for causal discovery. In Section III, we introduce intrinsic dimension estimation and its application to causal discovery. In Section IV, we first present the simulation results for different types of function forms. We then present the noise test results. Besides, we present the results of how the type of distribution affects the estimated dimension. Finally, we present the simulation results of mixed pattern. In Section V, we present test results for real world data. In Section VI, we summarize the key points.

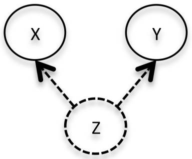

Fig. 1. Common Cause

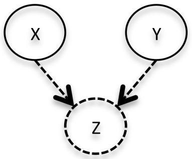

Fig. 2. Selection Bias

II. Related Work

Causal discovery [9], [10] has attracted the interest of researchers in various fields. The concept of causality was discussed and refined [11], [12]. Many causal discovery methods have been proposed until now. Some causal discovery models are based on cyclic models [13], which take into account feedback between variables in a causal relationship. Most causal discovery models use a directed acyclic graph (DAG) to express the causal relationships between variables [9], [10], [14]. The PC (Peter and Clark) algorithm [14] is a constraint-based algorithm to determine a set of DAGs in the same Markov equivalence class in which conditional independence test was used as a subroutine. Several methods have been proposed to test the conditional independence between variables and then apply them into causal discovery. [15] proposed using weakly additive noise models that use local permutation to measure conditional dependence. [16] proposed a kernel-based conditional independence (KCI) test in which there is no assumption about the specific functional form between variables and there is no constraint on the data distribution. The above methods either had additive noise assumption or used the uncorrelatedness of residual function to do the conditional independence test. When these assumptions were not satisfied, the conditional independence test may fail. The PC algorithm did not take into account the effect of confounding which is sound and complete in the assumption of causal sufficiency and faithfulness. The latterly proposed FCI (Fast Causal Inference) algorithm and RFCI algorithms extends the PC algorithm and handles the common cause and selection bias.

Cause effect pairs belong to the same Markov equivalence class and the general PC algorithm cannot learn its structure using conditional independence test. To discover the causal direction between cause and effect, additional assumptions have been added. [17] proposed expressing the relationship between cause and effect as an equation: $\text{effect} = f(\text{cause}) + \text{noise}$, where cause and additive noise are independent. The causal direction can be learned by fitting the data in both directions. If the data are generated according to the model assumption, the independence of assumed cause and corresponding residuals will exist in only one direction which can be inferred as the true one. Zhang et al. [18]–[21] proposed a post-nonlinear (PNL) model in which another non-linear function is added outside the equation: $\text{effect} = f_2(f_1(\text{cause}) + \text{noise})$. The PNL model takes into account the effect of external sensor distortion in addition to the nonlinear effect of causes and additive noise. In addition to using structural equation modeling to define the relationship between cause and effect, a probabilistic latent variable model was proposed by [22] to distinguish between cause and effect using standard Bayesian model selection. [23] proposed using information-geometric causal inference (IGCI) to distinguish cause from effect. IGCI assumes no additive noise between cause and effect and may fail in the large noise case. The above methods follows the causal sufficiency [24], [25] which assumes no unobserved common cause exists. Some methods have been proposed to deal with the unobserved common cause when telling cause from effect. Janzing et al. extended the original additive noise model [17] to identify confounding cases [26]. The proposed model is identifiable under suitable conditions. Shimizu et al. proposed a variant of the linear non-Gaussian acyclic model [27] to detect causal direction between two variables when there are latent confounding variables [28]. The above machine learning methods handle unobserved common causes and did not make it clear that which variable is confounding.

III. Proposed Method

We propose using intrinsic dimension estimation to test possible common causes. The proposed method was inspired by the IGCI causal discovery method [23] in which the asymmetries between cause and effect are identified by analyzing the complexity loss of their distributions. We studied the manifold of data in ordinary three-dimensional Euclidean space for the common cause case, as shown in Figure 1. After introducing the intrinsic dimension estimation method we used, we describe its application to causal discovery.

A. Intrinsic Dimension Estimation

The intrinsic dimension can be interpreted to mean the smallest number of variables needed to express the entire dataset; it is usually less than the total number of observed variables [29] and can be used to characterize fractals. Estimators of the intrinsic dimension are related to fractal geometry, e.g., the correlation dimension. We use the correlation dimension estimator introduced by Grassberger and Procaccia [30]. There are several methods to estimate the intrinsic dimension. We chose correlation dimension estimation because it is easy to be calculated and the estimated result can be proven theoretically. The correlation dimension is based on the assumption that the number of data points in a hypersphere with radius ε is proportional to the dimension ε^d [29]. The number of data points lying in a hypersphere within radius ε is given by

$$C(\varepsilon) = \lim_{N \to \infty} \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} H(\varepsilon - \|x_i - x_j\|), \quad (1)$$

where $H(x)$ is the Heaviside step function. Dimension d is estimated as ε approaches zero:

$$d = \lim_{\varepsilon \to 0} \frac{\log C(\varepsilon)}{\log \varepsilon}. \tag{2}$$

Since calculating the limit is impossible in practice, an approximate value is estimated in accordance with the L'Hospital rule, where ε_1 and ε_2 are set between the minimal and maximal pairwise distances measured in the data set [29]. An example of pseudocode was shown in Algorithm 1.

$$d \approx \frac{\log C(\varepsilon_2) - \log C(\varepsilon_1)}{\log \varepsilon_2 - \log \varepsilon_1} \tag{3}$$

Algorithm 1 Intrinsic dimension estimation

```
Input: Normalized input data $D$ without duplicated datapoints
Output: Estimated intrinsic dimension $d$
  $[m, n] \leftarrow size(D)$
  $distance \leftarrow list\,[\,]\,[\,]$
  while $i < m$ do
      if $i \neq m$ then
          $j = i + 1$
          $l \leftarrow zeros(m)$
          while $j \leqslant m$ do
              $\Delta x = \|x_i - x_j\|$
              $l[j] \leftarrow \Delta x$
              $j = j + 1$
          end while
      end if
      $distance.append(l)$
      $i = i + 1$
  end while
  $v \leftarrow$ no zero values of $distance$
  $\varepsilon_1 \leftarrow \mathrm{median}(v), \varepsilon_2 \leftarrow \max(v)$
  $s_1 \leftarrow \mathrm{length}(v < \varepsilon_1), s_2 \leftarrow \mathrm{length}(v < \varepsilon_2)$
  $Cr_1 \leftarrow (2/(n*(n-1)))*s_1$
  $Cr_2 \leftarrow (2/(n*(n-1)))*s_2$
  $d \leftarrow (\log(Cr_2) - \log(Cr_1))/(\log(\varepsilon_2) - \log(\varepsilon_1))$
```

B. Observable Common Cause Identification

The estimation results for the common cause case shown in Figure 1 differ from that for the selection bias case shown in Figure 2 (Table I). For the common cause case, the data are distributed in a one-dimensional manifold (a line), and the estimated intrinsic dimension is 1 ideally. For the selection bias case, the data are distributed in a two-dimensional manifold (a surface), and the estimated intrinsic dimension is 2 ideally.

For the common cause case, suppose that $X_i = \alpha m_i, Y_i = \beta m_i, Z_i = m_i$. The distance between two points x_i and x_j with coordinates (X_i, Y_i, Z_i) and (X_j, Y_j, Z_j) respectively can be calculated using

$$\begin{aligned} \|x_i - x_j\| &= \sqrt{\alpha^2 (m_i - m_j)^2 + \beta^2 (m_i - m_j)^2 + (m_i - m_j)^2} \\ &= \sqrt{\alpha^2 + \beta^2 + 1}\, |m_i - m_j| \\ &= \sqrt{\alpha^2 + \beta^2 + 1} \Delta m \propto \Delta m. \end{aligned} \tag{4}$$

Suppose that the datapoints exist in a line with unit length. The distance between two adiacency dataponits would be $\frac{1}{N}$. Thus, the convergence rate of ε is proportional to that of $\frac{1}{N}$. For the common cause case, a suitable value proportional to $\frac{1}{N}$ can be found for ε which makes that $H(\varepsilon - \|x_i - x_j\|)$ equals one only when x_j is the next nearest point $(\Delta m \to 0)$ except the last point. This means that $\sum_{j=i+1}^{N} H(\varepsilon - \|x_i - x_j\|) = 1 + \underbrace{0 + \cdots + 0}_{N-i-1}$. We thus get

$$\begin{aligned} &\log C\left(\varepsilon \to 0, \varepsilon \propto \frac{1}{N}\right) = \\ &\lim_{N \to \infty} \log \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \left(1 + \underbrace{0 + \cdots + 0}_{N-i-1}\right). \end{aligned} \tag{5}$$

The estimated dimension d can then be derived:

$$\begin{aligned} d &= \lim_{\varepsilon \to 0, \varepsilon \propto 1/N} \log C(\varepsilon) / \log \varepsilon \\ &= \lim_{\varepsilon \to 0, \varepsilon \propto 1/N} \left[\lim_{N \to \infty} \log\,(N-1)/N(N-1) \Big/ \log \varepsilon \right] \\ &= \lim_{\varepsilon \to 0, \varepsilon \propto 1/N} \left[\lim_{N \to \infty} \log\,(1/N) \Big/ \log \varepsilon \right] \end{aligned} \tag{6}$$

which is approximately equal to 1.

For the selection bias case, suppose $X_i = m_i, Y_i = n_i, Z_i = \alpha m_i + \beta n_i$. The distance between x_i and x_j is given by

$$\begin{aligned} \|x_i - x_j\| &= \sqrt{(\alpha^2 + 1)\Delta m^2 + (\beta^2 + 1)\Delta n^2 + 2\alpha\beta \Delta m \Delta n}. \\ &\propto \sqrt{\Delta m^2 + \Delta n^2 + \Delta m \Delta n}, \end{aligned} \tag{7}$$

where $\Delta m = m_i - m_j, \Delta n = n_i - n_j$. In this case, the value of $\|x_i - x_j\|$ is affected by the values of Δm and Δn. The x_j could be the next nearest point in two directions: $\Delta m \to 0, \Delta n = 0$ or $\Delta n \to 0, \Delta m = 0$. The datapoints distributed in a surface and the convergence rate of ε is proportional to that of $\frac{1}{\sqrt{N}}$. In this case, suitable values $\varepsilon_1, \varepsilon_2$ proportional to $\frac{1}{\sqrt{N}}$ can be found which makes that $H(\varepsilon - \|x_i - x_j\|)$ equals one when x_j is the next nearest point in the two directions. If $\varepsilon_1 = \varepsilon_2$, we can thus get that $\sum_{j=i+1}^{N} H(\varepsilon - \|x_i - x_j\|) = 1 + 1 + \underbrace{0 + \cdots + 0}_{N-i-2}$ when $i \neq N - 1$. When $i = N - 1$,

$\sum_{j=i+1}^{N} H(\varepsilon - \|x_i - x_j\|) = 1$. In this case,

$$\log C\left(\varepsilon \to 0, \varepsilon \propto \frac{1}{\sqrt{N}}\right) = \lim_{N\to\infty} \log \frac{1}{N(N-1)}\left[\sum_{i=1}^{N-2}\left(1+1+\underbrace{0+\cdots+0}_{N-i-2}\right)+1\right]. \tag{8}$$

The estimated dimension d is given by

$$\begin{aligned} d &= \lim_{\varepsilon\to 0, \varepsilon\propto 1/\sqrt{N}} \log C(\varepsilon) / \log\varepsilon \\ &= \lim_{\varepsilon\to 0, \varepsilon\propto 1/\sqrt{N}} \left[\lim_{N\to\infty} \log\left[2(N-2)+1\right]/N(N-1) \Big/ \log\varepsilon\right] \\ &= \lim_{\varepsilon\to 0, \varepsilon\propto 1/\sqrt{N}} \left[\lim_{N\to\infty} \log\left[2/N - 1/N(N-1)\right] \Big/ \log\varepsilon\right] \\ &\propto \lim_{N\to\infty} \log\left[2/N - 1/N(N-1)\right] / \log N^{-0.5} \\ &\propto \lim_{N\to\infty} \log N^{-1} \big/ \log N^{-0.5} . \end{aligned} \tag{9}$$

If $\varepsilon_1 \neq \varepsilon_2$, the next nearest point would lie in the direction with the smaller ε. We can get that

$$\log C\left(\varepsilon \to 0, \varepsilon \propto \frac{1}{\sqrt{N}}\right) = \lim_{N\to\infty} \log \frac{1}{N(N-1)} \sum_{i=1}^{N-1}\left(1+\underbrace{0+\cdots+0}_{N-i-1}\right). \tag{10}$$

The estimated dimension d is given by

$$\begin{aligned} d &= \lim_{\varepsilon\to 0, \varepsilon\propto 1/\sqrt{N}} \log C(\varepsilon) / \log\varepsilon \\ &= \lim_{\varepsilon\to 0, \varepsilon\propto 1/\sqrt{N}} \left[\lim_{N\to\infty} \log\,(N-1)/N(N-1) \Big/ \log\varepsilon\right] \\ &= \lim_{\varepsilon\to 0, \varepsilon\propto 1/\sqrt{N}} \left[\lim_{N\to\infty} \log\,(1/N) \Big/ \log\varepsilon\right] \\ &\propto \lim_{N\to\infty} \log N^{-1} / \log N^{-0.5} . \end{aligned} \tag{11}$$

The estimated dimension was not affected by whether ε_1 equals ε_2 or not. Eqs. 9 and 11 are both approximately equal to 2. Scatter plots of the simulation data[1] are shown in Figures 3 and 4. The manifold for the common cause case is a line with an intrinsic dimension around 1 while that for the selection bias case is a surface with an intrinsic dimension around 2. These characteristics can be used to detect possible confounders for variables x and y.

Suppose that variables x and y are dependent. To test whether a variable z is a possible common cause, z can be added into the data. If the estimated correlation dimension is around 1, there is high possibility that z is a confounding common cause variable for x and y. Here it is hard to assert that z is a common cause variable for x and y because the Markov equivalence class exists.

TABLE I
IDEAL ESTIMATED INTRINSIC DIMENSION FOR TWO CONFOUNDING CASES.

Variable 1	Variable 2	Variable 3	Estimated Dimension
x=f(z)	y=g(z)	z	around 1
x	y	z=h(x,y)	around 2

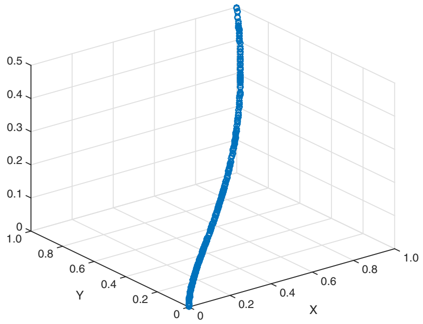

Fig. 3. Simulation: Common Cause

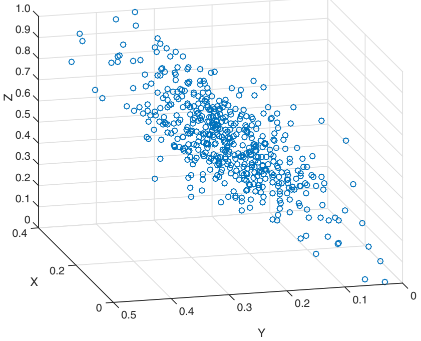

Fig. 4. Simulation: Selection Bias

Like a conditional independence test, correlation dimension estimation cannot distinguish the following three patterns belonging to the same Markov equivalence class.

1) A chain: $X \to Z \to Y$
2) Another chain: $X \leftarrow Z \leftarrow Y$
3) A fork: $X \leftarrow Z \to Y$

For the first two patterns, X and Y have an indirect causal relationship via variable Z. For the last pattern, Z works as a common cause for X and Y. For all three patterns, if variable Z is fixed, the path between X and Y can be blocked [9].

An estimated dimension of 1 works as a necessary condition for variable z to be a common cause for x and y. The necessity can be used to determine whether a third variable is a possible common cause for two observed variables. If the estimated dimension is around 1, the third variable is a possible common cause for the two observed variables. As far as we have known, no machine learning methods are sufficient to decide a common cause even the generally

[1] Simulation data for the common cause case were generated using $x = z^2, y = z^3$; those for the selection bias case were generated using $z = x + y$.

used (conditional) independence test. The necessity can help filter the unrelated variables. To distinguish the exact DAG, additional assumptions are needed. For example, an additive noise model [17] needs to be assumed. Nevertheless, the proposed method is effective and efficient as an initial step in detecting possible confounding variables and is easy to implement.

IV. Simulation

We simulated application of the proposed method for linear and non-linear functions and their combination and compared the results with those for the KCI test proposed by [16]. Our proposed method worked better in the Poisson noise case and slightly better in the multiplicative noise case. Simulation of the selection bias case (Figure 2) showed that the proposed method was not affected by Berkson's paradox. To test how the type of distribution affects the estimated correlation dimension, we generated simulation data using three types of distributions: logistic, Gaussian, and uniform. Finally, we simulated the mixed pattern with causal relationship and showed that the proposed method can tell a possible common cause even when it is mixed with causal relationship.

A. Simulation of Common Cause Case

We used linear and non-linear functions for the common cause simulation. The artificial data were generated using two equations,

$$\begin{aligned} X &= \alpha * Z + \beta * Z^2, \\ Y &= 2 * \alpha * Z + \beta * Z^3, \end{aligned} \tag{12}$$

where α and β are used to control the nonlinearity of the function. We set parameter β to 0 and varied parameter α from 0.1 to 1 for the linear function. For the nonlinear function, we set α to 0 and varied the value of β instead. The experimental results are presented below.

1) Different Types of Function Forms:

a) Linear Function: In the linear function simulation, we set β to 0 and varied the value of α. For each α, we repeated the randomized experiment 20 times and took the average of the estimated correlation dimension. As shown by the plot in Figure 5, the estimated dimension was negligibly affected by the value of alpha. In all cases, the estimated dimension was very close to 1. We did the same simulation with the method proposed by [16] in which α was varied and for each α 20 randomized experiments were run. The p-value was not affected by the value of α either. However, in 200 randomized experiments, the independence of X and Y given Z was accepted only eight times ($\alpha = 0.05$).

b) Non-linear Function: In the non-linear function simulation, we set α to 0 and varied the value of β to simulate a non-linear function in the no additive noise case. As shown in Figure 6, the estimated intrinsic dimension was very close to 1. The same simulation was run for the KCI test [16] with varied β values. The p-value was not affected by the value of β either. In 200 randomized experiments, the independence of X and Y given Z was accepted ten times.

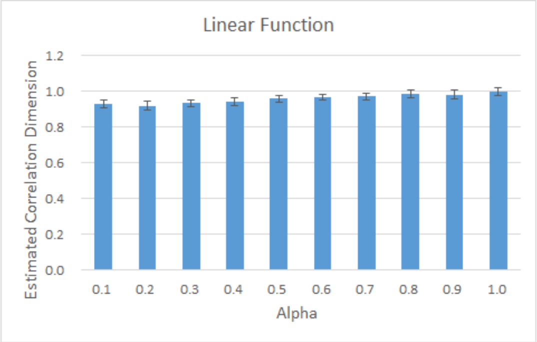

Fig. 5. Estimated Correlation Dimension for Linear Function

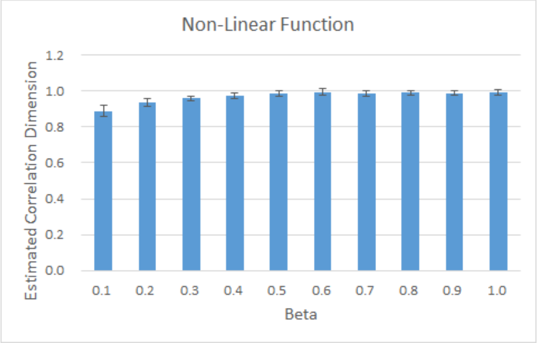

Fig. 6. Estimated Intrinsic Dimension for Non-Linear Function

c) Combination of Linear and Non-Linear Functions: In the combination simulation, both α and β were set to 1, and the randomized experiment was repeated 20 times. The average estimated intrinsic dimension was 1.0082 with a standard deviation of 0.0221, indicating that the estimated dimension is negligibly affected by the type of function between variables. The same simulation was run for the KCI test [16]. In all 20 randomized experiments, the independence of X and Y given Z was not accepted. KCI test assumed uncorrelatedness of residual function given the common cause. Less of residuals made KCI test fail. Compared with KCI test, our proposed method worked better in all the above cases.

2) Noise Test: We tested the effect of three different types of noise: Poisson noise, multiplicative noise, and Gaussian noise (Figure 7).

a) Poisson noise: Poisson noise is a general kind of image sensor noise that is not stationary and neither additive nor multiplicative. We simulated data using Eq. 12, where α and β were set to 1, and Z was generated in a standard normal distribution. We added noise into the generated data through the Poisson process. If the no-noise data was 1, the noised data was generated from a Poisson distribution with expectation 1. Randomized experiments were run 100 times, and the average estimated intrinsic dimension was 0.9053 with a standard variance of 0.0239. The proposed method worked well in the Poisson noise case. For comparison, we ran the KCI test with the same experiment settings. In the 100 randomized experiments, the conditional independence between X and Y

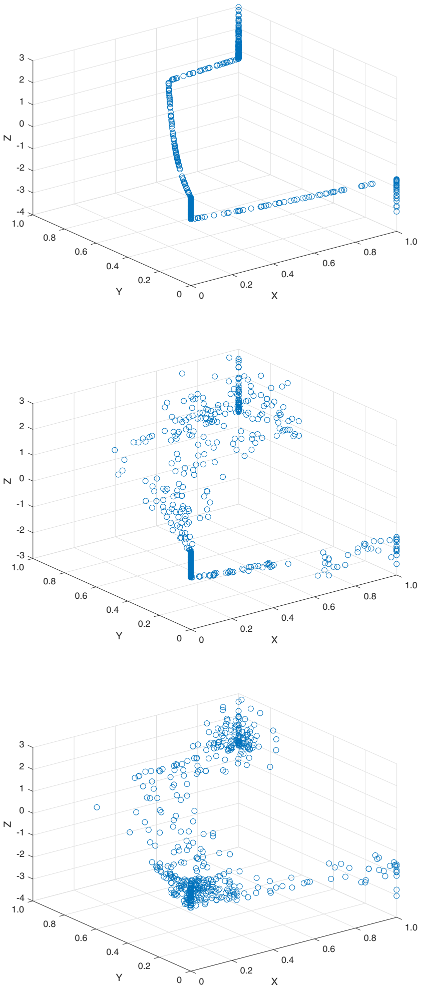

Fig. 7. Scatter Plot of Artificial Data with Poisson Noise, Multiplicative Noise, and Gaussian Noise

given Z was accepted 47 times ($\alpha = 0.05$). That is, about half of them were rejected. These experimental results showed that the proposed method worked better than the KCI test when there was Poisson noise.

b) Multiplicative noise: We added multiplicative noise to the data generation process using the equation $J = I + n * I$, where n is random noise with mean 0. We added noise with different variances from 0.01 to 0.05 into the generated artificial data. The experimental results showed that the estimated correlation dimension was negligibly affected by the variance of the noise. We conducted randomized experiments 100 times; the average estimated correlation dimension was 0.8972 with a standard variance of 0.0445. We did the same simulation for the KCI test. In the 100 randomized tests, the conditional independence of X and Y given Z was accepted 95 times ($\alpha = 0.05$).

c) Gaussian noise: We added Gaussian noise with different variances to the generated artificial data. The estimated correlation dimension with Gaussian noise for variances from 0.01 to 0.05 is shown in Figure 8. We ran randomized experiments 20 times for each variance. The estimated correlation dimension was a little larger than 1; the estimated intrinsic dimension tended to initially increase and then decrease as the degree of noise increased. We conducted additional randomized experiments with variances from 0.0001 to 10. When the variance was 0.0001, the average estimated correlation dimension was 1.1361 with a standard variance of 0.0608. When the variance was 10, the average of estimated correlation dimension was 1.0208 with a standard variance of 0.0173. The estimated intrinsic dimension was close to one when the added Gaussian noise was very little or very large. In all the randomized experiments, the estimated intrinsic dimension was less than 1.7. We did the same simulation for the KCI test. In the 100 randomized tests, the conditional independence of X and Y given Z was accepted 94 times.

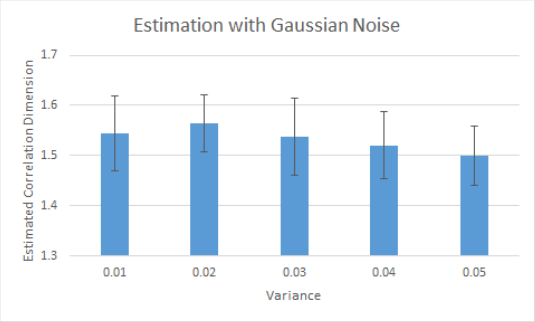

Fig. 8. Estimated Correlation Dimension for Added Gaussian Noise with Different Variances

These results show that the proposed method generally worked well for different types of added noise. The estimated value of the correlation dimension is affected by the type of noise and number of samples. The estimated correlation dimension after adding Gaussian noise was below 1.7 (Fig. 8). Thus, we think 1.7 would be a suitable upper limit value for the proposed method. The estimated correlation dimension was above 0.8 in our simulation (Figs. 5 and 6). Since the real world data has limited data points and the observed data is discrete, the estimated dimension tends to be lower. Therefore, we reduced the lower limit by 0.1 to 0.7. According to the above experimental results and considerations, we set the confidence interval to be [0.7, 1.7] when applying the correlation dimension to causal discovery.

B. *Simulation of Selection Bias Case*

We simulated the selection bias case (Figure 2) using $Z = X + Y^3$. We first randomly generated 2000 pairs of X-Y data values from a uniform distribution (0, 1) and calculated the corresponding Z values. Next, we counted the instances of Z in each sub-interval. Finally, we selected the X and Y with Z in the interval with the most instances. Figure 9 shows the count of Z in the sub-intervals of interval (0, 2.0). Figure 10 shows a scatter plot of 205 instances of the selection bias data with Z in the interval (0.9, 1.0). We used the KCI test [16] to test the independence of X and Y. The p-value was 0, which means that X and Y were not independent although they were randomly generated and thus should have been independent. However, selection bias can make two randomly generated variables dependent (Berkson's paradox). Intrinsic dimension estimation using the selected biased data shown in Figure 10 gave an estimated value of 1.8322. The average estimated dimension for 20 randomized experiments was 1.8273 with a variance of 0.0568, demonstrating that the proposed method can give a correct estimation even when the data is biased.

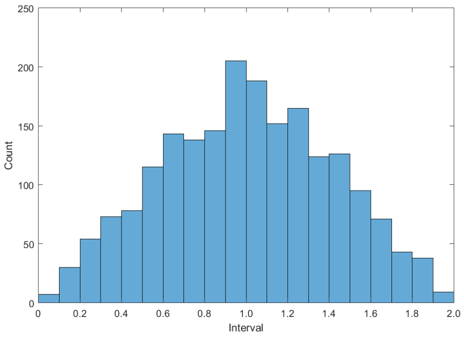

Fig. 9. Count of Z in Sub-intervals of Interval (0, 2.0)

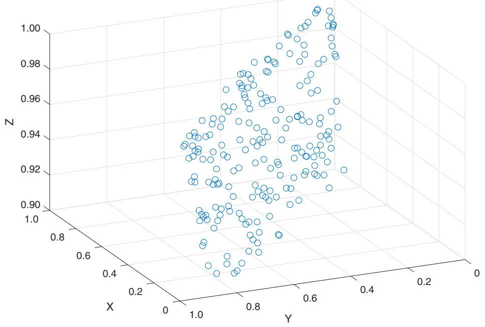

Fig. 10. Scatter Plot of Selection Bias Data for Z in (0.9, 1.0)

C. Further Study

1) Different Types of Distributions: To test how the type of distribution affects the estimated intrinsic dimension, we used three different types of distributions: logistic, Gaussian, and uniform (Figure 11) with kurtosis of 1.2, 0, and -1.2 for normalized data with mean 0 and variance 1 [2]. The estimated dimensions are shown in Table II. For all three types, the estimated dimension was close to 1, indicating that the estimated dimension is slightly affected by the data distribution.

2) Mixed Pattern of Common Cause and Causal Relationship: Common cause and causal relationship may be mixed with each other in the real world. If causal relationship exists between X and Y, the conditional independence test cannot tell a possible common cause of X and Y any more. However, even when real causal relationship exists, the effect of confounding should be considered as well. As shown by the existing researches, causal relationship may change after considering the effect of specific confounder [5], [6]. Our proposed method can still decide possible common cause when it is mixed with causal relationship. We ran simulation to show the experimental results. The artificial data was generated by $X = Z + Z^2, Y = 2 * Z + Z^3 + X$ in which Z was a common cause for X and Y, X causes Y. The average of estimated intrinsic dimension was 0.9410 with standard variance 0.0299. The simulation results showed that the proposed method can give correct inference when causal relationship and common cause are mixed with each other.

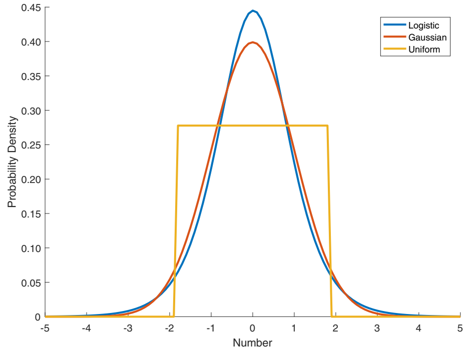

Fig. 11. Probability Density Function for Three Types of Distributions

TABLE II
ESTIMATED DIMENSION FOR DIFFERENT TYPES OF DISTRIBUTIONS.

Distribution	Kurtosis (mean: 0, variance: 1)	Estimated Dimension
Logistic	1.2	0.9204
Gaussian	0	0.9207
Uniform	-1.2	1.0009

V. REAL WORLD DATA

Finally, we used real world data obtained from the Cause-EffectPairs (CEP) dataset [31] to test our proposed method.

A. Altitude, Temperature, Sunshine

The pairs "altitude→temperature" and "altitude→sunshine" are two cause effect pairs in the CEP dataset. They were taken from the UCI Machine Learning Repository [32]. Temperature and sunshine are variables confounded by a common cause, altitude. The p-value from a KCI test for temperature and sunshine was 0.002. Given the altitude, the p-value for temperature and sunshine was 3.2779e-04. The independence of temperature and sunshine given the altitude was thus rejected ($\alpha = 0.05$). For the three variables, although altitude is a common cause for the other two variables, sunshine causes temperature as well in human cognition. Conditional independence test cannot tell a possible common cause anymore when the two are mixed with each other. Although altitude is a common cause for temperature and sunshine, the KCI test failed to detect it. With our proposed method, the estimated intrinsic dimension was 0.7132, which lies in the interval [0.7,

[2] https://en.wikipedia.org/wiki/Kurtosis

1.7], showing that the variable "altitude" can be a common cause for "temperature" and "sunshine."

B. Cement, Blast Furnace Slag, Compressive Strength

The pairs "cement→compressive strength" and "blast furnace slag→compressive strength" are two more cause effect pairs in the CEP dataset. We combined the data to get a real world example of "selection bias" (Figure 2). The p-value from a KCI test for cement and blast furnace slag was 0, which shows that the data was biased. The estimated intrinsic dimension was 2.0156. Application of dimension estimation to finding possible confounding variables did not suffer from the effects of Berkson's paradox and was useful in common cause identification.

VI. Conclusion

We proposed using intrinsic dimension estimation to detect a possible common cause for two variables. The proposed method does not need certain assumptions of function forms, data distributions. Comparison experiment was conducted with a kernel-based independence test. The experimental results showed that the proposed method worked comparatively or better than the existing method. Besides, the proposed method can still tell a possible common cause when it is mixed with causal relationship while conditional independence test cannot. The proposed method does not suffer from the effects of Berkson's paradox. The estimated intrinsic dimension was around 1 for the common cause case and around 2 for the selection bias case. In a scatter plot of data, the topology of the data generated for the common cause case would be a line while that for the selection bias case would be a surface. Simulation testing showed that the estimated correlation dimension is slightly affected by the type of distribution. Testing using variables from the CauseEffectPairs dataset showed that the proposed method can estimate the common cause for real world data.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number JP15K12148.

References

[1] C. Scholz, "Strata-a method for strategic analysis of complex systems," *European Journal of Operational Research*, vol. 24, no. 1, pp. 168–177, 1986.

[2] E. A. Stuart, "Matching methods for causal inference: a review and a look forward," *Statistical Science*, vol. 25, no. 1, p. 1, 2010.

[3] M. A. Pourhoseingholi, A. R. Baghestani, and M. Vahedi, "How to control confounding effects by statistical analysis," *Gastroenterology and Hepatology From Bed to Bench*, vol. 5, no. 2, p. 79, 2012.

[4] K. Jager, C. Zoccali, A. Macleod, and F. Dekker, "Confounding: what it is and how to deal with it," *Kidney International*, vol. 73, no. 3, pp. 256–260, 2008.

[5] B.-N. Huang, M. J. Hwang, and C. W. Yang, "Causal relationship between energy consumption and GDP growth revisited: a dynamic panel data approach," *Ecological Economics*, vol. 67, no. 1, pp. 41–54, 2008.

[6] D. Coondoo and S. Dinda, "Causality between income and emission: a country group-specific econometric analysis," *Ecological Economics*, vol. 40, no. 3, pp. 351–367, 2002.

[7] J. Song, S. Oyama, and M. Kurihara, "Tell cause from effect: models and evaluation," *International Journal of Data Science and Analytics*, vol. 4, no. 2, pp. 99–112, 2017.

[8] ——, "A framework for crowd-based causal analysis of open data," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2018, pp. 2184–2189.

[9] J. Pearl, "Causality: models, reasoning, and inference," *Econometric Theory*, vol. 19, no. 675-685, p. 46, 2003.

[10] ——, *Causality*. Cambridge University Press, 2009.

[11] C. W. Granger, "Some recent development in a concept of causality," *Journal of Econometrics*, vol. 39, no. 1-2, pp. 199–211, 1988.

[12] J. Y. Halpern, "A modification of the Halpern-Pearl definition of causality." in *International Joint Conference on Artificial Intelligence*, 2015, pp. 3022–3033.

[13] A. Hyttinen, F. Eberhardt, and P. O. Hoyer, "Learning linear cyclic causal models with latent variables," *Journal of Machine Learning Research*, vol. 13, pp. 3387–3439, 2012.

[14] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT Press, 2000.

[15] R. E. Tillman, A. Gretton, and P. Spirtes, "Nonlinear directed acyclic structure learning with weakly additive noise models," in *Advances in Neural Information Processing Systems*, 2009, pp. 1847–1855.

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

[17] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Advances in Neural Information Processing Systems*, 2009, pp. 689–696.

[18] K. Zhang and L.-W. Chan, "Extensions of ICA for causality discovery in the Hong Kong stock market," in *International Conference on Neural Information Processing*, 2006, pp. 400–409.

[19] K. Zhang and A. Hyvärinen, "Distinguishing causes from effects using nonlinear acyclic causal models," in *Proceedings of the 2008 International Conference on Causality: Objectives and Assessment*, 2008, pp. 157–164.

[20] ——, "On the identifiability of the post-nonlinear causal model," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009, pp. 647–655.

[21] K. Zhang, Z. Wang, and B. Scholkopf, "On estimation of functional causal models: post-nonlinear causal model as an example," in *IEEE 13th International Conference on Data Mining Workshops*, 2013, pp. 139–146.

[22] O. Stegle, D. Janzing, K. Zhang, J. M. Mooij, and B. Schölkopf, "Probabilistic latent variable models for distinguishing between cause and effect," in *Advances in Neural Information Processing Systems*, 2010, pp. 1687–1695.

[23] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf, "Information-geometric approach to inferring causal directions," *Artificial Intelligence*, vol. 182, pp. 1–31, 2012.

[24] J. Pearl, "An introduction to causal inference," *The International Journal of Biostatistics*, vol. 6, no. 2, pp. 1–62, 2010.

[25] P. Spirtes, "Introduction to causal inference," *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1643–1662, 2010.

[26] D. Janzing, J. Peters, J. Mooij, and B. Schölkopf, "Identifying confounders using additive noise models," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009, pp. 249–257.

[27] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-gaussian acyclic model for causal discovery," *Journal of Machine Learning Research*, vol. 7, pp. 2003–2030, 2006.

[28] S. Shimizu and K. Bollen, "Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-gaussian distributions," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2629–2652, 2014.

[29] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*. Springer Science & Business Media, 2007.

[30] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," *The Theory of Chaotic Attractors*, pp. 170–189, 2004.

[31] J. M. Mooij, D. Janzing, J. Zscheischler, and B. Schölkopf, "CauseEffectPairs repository," 2014. [Online]. Available: https://webdav.tuebingen.mpg.de/cause-effect/

[32] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml/