Causal Inference & Deep Learning

Notes on Cause-Effect Discovery

MIT IAP 2018 Compiled by Max W. Shen

Notes are heavily derived from references and textbooks.

Previously, the basic concepts of causal learning and its relationship to statistical learning were discussed. Building off of this, the discussion turns to several approaches for inferring the causal model from observational data in the simple regime of only two observed variables. This task is known as cause-effect discovery.

Put another way, given an empirically observed joint distribution on two observed variables X and Y, cause-effect discovery is the task of deciding if the causal graph is $X \to Y$ or $X \leftarrow Y$. Unobserved confounding variables are assumed to not exist. Even in such a simple regime, the inverse problems cannot be solved without additional assumptions.

1 Additive Noise Models

Consider a causal generative process where $X \to Y$ via:

$$Y \leftarrow f_Y(X) + N_Y, X \perp\!\!\!\perp N_Y. \tag{1}$$

If this holds, $P_{Y|X}$ is said to admit an additive noise model (ANM).

This generative process assumes additive noise that is independent of the causal input, and is known as an additive noise model. While the assumption of additive independent noise is restrictive, it enables identifiability in the cause-effect discovery task.

The Tübingen dataset of causal pairs is a useful benchmark for the cause-effect discovery task and consists of tuples $((x_1, y_1), z_1), ..., ((x_n, y_n), z_n)$, where $x_i \in X$ are inputs (or co-variates), $y_i \in Y$ are outputs (or labels), and $z_i \in Z = \{\leftarrow, \to\}$ describes the causal model between X and Y. Its data are gathered from a variety of real-world and scientific settings. One example is $X =$ "altitude" and $Y =$ "temperature", where the true causal direction is "altitude" $\to$ "temperature" or $X \to Y$.

1.1 Special Case: Linear Model with Non-Gaussian Noise

In a special case of an additive noise model that is linear:

$$Y \leftarrow \alpha X + N_Y, X \perp\!\!\!\perp N_Y. \tag{2}$$

The following theorem states that the model is identifiable if and only if N_Y is non-Gaussian.

Theorem 1 (Identifiability of linear non-Gaussian models)**.** *Assume that $P_{X,Y}$ admits the linear model*

$$Y = \alpha X + N_Y, N_Y \perp\!\!\!\perp X, \tag{3}$$

with continuous random variables X, N_Y, and Y. Then there exists $\beta \in \mathbb{R}$ and a random variable N_X such that

$$X = \beta Y + N_X, N_X \perp\!\!\!\perp Y, \tag{4}$$

if and only if N_Y and X are Gaussian.

For a proof, refer to references.

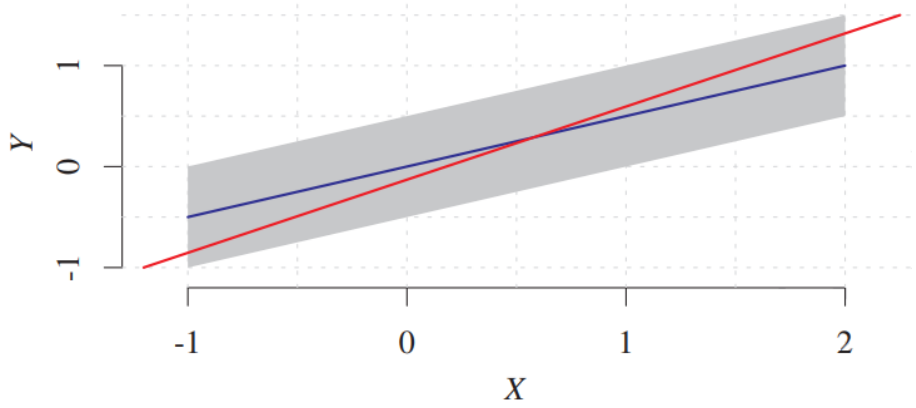

Figure 1: Linear model with Uniform Noise is identifiable

In figure 1, the blue line corresponds to the forward model $X \to Y$ with uniform independent noise represented by the gray area. The red line corresponds to the backward model $X \leftarrow Y$. In such a setting, the backward model's residuals $X - \beta Y$ are not independent of Y, therefore $X \to Y$ is identified as the causal direction.

1.2 General Conditions for Identifiability

Theorem 2. *Assume that $P_{Y|X}$ admits a "smooth" ANM $X \to Y$ such that N_Y and X have strictly positive densities p_{N_Y} and p_X, and that f_Y, p_{N_Y}, and p_X are three times differentiable. Assume also that there exists a $y \in \mathbb{R}$ such that*

$$(\log p_{N_Y})''(y - f_Y(x)) f_Y'(x) \neq 0 \tag{5}$$

for all but countably many values x.

Then, the set of log densities $\log p_X$ for which the obtained joint distribution $P_{X,Y}$ admits a smooth ANM $X \leftarrow Y$ is contained in a three-dimensional affine space.

Then, "generically", a joint distribution does not admit an ANM in both directions simultaneously. For a proof, refer to references.

1.3 Post-Nonlinear Additive Noise Models

Similar identifiability results hold for a slightly more general family of models known as post-nonlinear additive noise models where the causal generative process is assumed to be:

$$Y \leftarrow g_Y(f_Y(X) + N_Y), X \perp\!\!\!\perp N_Y. \tag{6}$$

1.4 Hilbert-Schmidt Independence Criterion

The following procedure is one method for solving cause-effect discovery problems using the structure of additive noise models.

1. Regress Y on X by learning a function $\hat{f}_Y : X \to Y$. Here, $\hat{f}_Y$ could be implemented as a deep neural network.
2. Test if the residuals $\epsilon = Y - \hat{f}_Y(X)$ is independent of X on held-out data.
3. Repeat steps 1 and 2 in the other direction, exchanging the roles of X and Y
4. If independence is accepted in one direction and rejected in the other, return the accepted direction as the causal direction. Otherwise, abstain.

A useful non-parametric independence test is the Hilbert-Schmidt Independence Criterion (HSIC).

HSIC maps distributions into an RKHS using a kernel mean embedding that is a one-to-one mapping when using a characteristic kernel. Using this approach, the HSIC score is defined as the squared RKHS distance between the embedded joint distribution $p(X, \epsilon)$ and the embedded product distribution $p(X)p(\epsilon)$, which is 0 if and only if ϵ and X are independent.

$$\text{HSIC}(P_{X,Y}) = ||\mu(P_{X,Y}) - \mu(P_X P_Y)||^2_H. \tag{7}$$

In practice, the HSIC score is computed using kernels.

Under mild assumptions, the HSIC score using empirical samples from distributions converges to the HSIC score directly using the true distributions at a rate of $\frac{1}{\sqrt{n}}$ where n is the number of empirical samples.

On the Tübingen cause-effect pairs dataset, an additive noise model approach using Gaussian process regression and HSIC achieves 59.5% weighted accuracy. A post-nonlinear additive noise model with Gaussian process regression and HSIC achieves 66.2% weighted accuracy.

Gretton, A., Bousquet, O., Smola, A., & Scholkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. Lecture Notes in Computer Science, 3734 LNAI, 63 - 77.

2 Kolmogorov Complexity

One intuitive assumption that could be used for cause-effect discovery is independence of mechanisms, meaning that the mechanism P_C producing the cause is assumed to be independent of the mechanism $P_{E|C}$ that produces the effect given the cause.

Standard statistical tools, however, are ill-equipped for mathematically representing such an assumption. Statistical independence describes independence between random variables, but independence of mechanisms requires a notion of independence on distributions of random variables.

Kolmogorov complexity enables the discussion of a type of independence between mechanisms, but further work is needed to develop it in practice.

Using a universal Turing machine, let $K(s)$ denote the length of the shortest program s^* that generates a binary string s and then stops. The program s^* can be understood as describing a "best" compression of s, and the length $K(s) = |s^*|$ is related to the information complexity of the string s. Unfortunately, it can be shown that $K(s)$ is uncomputable, that is, no algorithm is capable of computing $K(s)$ from s. This precludes this approach from being practical, but its underlying ideas may inspire related approaches.

The concept of conditional algorithmic information of s given t is denoted by $K(s|t)$ which describes the length of the shortest program that outputs s given input t and then stops. A concept of algorithmic mutual information can then be defined as:

$$I(s : t) = K(s) - K(s|t), \tag{8}$$

up to an additive constant.

For a cause C and an effect E, allow P_C and $P_{E|C}$ to admit finite descriptions via binary strings, which can be thought of as generative programs that compute $p(c)$ and $p(e|c)$ for all values c and e. Given this terminology, the assumption of independence of mechanisms can be thought of as independence of algorithms, and represented as:

> P_C and $P_{E|C}$ are algorithmically independent if
>
> $$I(P_C : P_{E|C}) = 0, \tag{9}$$
>
> up to an additive constant, and equivalently,
>
> $$K(P_{C,E}) = K(P_C) + K(P_{E|C}), \tag{10}$$
>
> up to an additive constant.

Cause-effect discovery can then be performed by returning the causal direction that better fits this property.

Information within statistical dependence relationships can be informative for causality between random variables; information within algorithmic dependence relationships can be informative for causality between objects.

3 Conditional GANs

The assumption underlying additive noise models does not apply to situations with multiplicative or heteroskedastic noise.

One approach for loosening assumptions is training a conditional GAN to generate:

$$\begin{aligned} x &\sim P(X), \\ n &\sim P(N), \\ y &\leftarrow g(x, n), \end{aligned} \tag{11}$$

with an adversary that attempts to distinguish generated data from observed (x, y):

$$d : (x, y) \to [0, 1]. \tag{12}$$

The noise n can be constructed to be independent of input x. This implies a generative structure assuming only independent noise, rather than the more restrictive assumption of additive and independent noise in additive noise models.

Conditional generative adversarial networks alternate between minimizing two objective functions with stochastic gradient descent:

$$\begin{aligned} L_d(g) &= E_{x,y}[l(d(x, y), 1)] + E_{x,z}[l(d(x, g(x, z)), 0)], \\ L_g(d) &= E_{x,y}[l(d(x, y), 0)] + E_{x,z}[l(d(x, g(x, z)), 1)]. \end{aligned} \tag{13}$$

Two conditional GANs are trained, one with generator $g_y : X \to Y$ and the other with generator $g_x : Y \to X$. The choice of the causal direction corresponds to the generator whose generated $\tilde{y} = g(x, z)$ is better able to confuse a discriminator $d : (x, y) \to [0, 1]$ when discriminating between real data (x, y) and generated data $(x, \tilde{y})$.

In practice, training conditional GANs is an unstable process. To remedy this, an ensemble approach can be used, where a set of g_y are trained with independent random seeds instead of just one g_y.

Lopez-paz benchmarked this approach on the Tübingen pairs dataset. ANM-HSIC achieves a 67% accuracy, randomized causal coefficient (a supervised approach discussed later) achieves a 76% accuracy. Lopez-paz considered conditional GANs using a discriminative neural network during training to assist the generative neural network, then used a new discriminator for the task of deciding a causal direction. When using a k-nearest-neighbors approach, the conditional GANs achieved 73% accuracy without ensembling and 82% accuracy with ensembling and deciding between the single best performer in each ensemble. His approach appears sensitive to the choice of the new discriminator, however: using a neural net with a single hidden layer as the new discriminator yielded a 70% accuracy without ensembling and 73% accuracy with ensembling and deciding between the single best performer in each ensemble.

Lopez-Paz, D., & Oquab, M. (2017). Revisiting Classifier Two-Sample Tests for GAN Evaluation and Causal Discovery. ICLR 2017.

4 Supervised Cause-Effect Discovery

In contrast to the methods discussed so far, in practice, we often have expert domain knowledge on the causal relationships between random variables. In such a setting, a reasonable approach would be to use these as supervised labels to train a model that will hopefully perform well on cause-effect discovery on held-out data.

Consider a dataset $((x_1, y_1), z_1), ..., ((x_n, y_n), z_n)$, where $x_i \in X$ are inputs (or covariates), $y_i \in Y$ are outputs (or labels), and $z_i \in Z = \{\leftarrow, \rightarrow\}$ describes the causal model between X and Y.

Note that Z can be augmented to include arbitrary causal relationship between X and Y, including bicausal (X and Y both cause each other), confounding variables (X and Y are caused by a latent parent), and no causal link.

In such a setting, a key concept is that of a "causal signature". In the realm of unsupervised causal discovery, causal signatures were assumed from general philosophical priors such as independence of mechanisms, but were further constrained by identifiability. There are likely other useful causal signatures beyond independent additive noise. In a supervised setting, rather than explicitly assuming causal signatures, the approach of implicitly learning causal signatures can be considered instead.

The following sections explore cause-effect discovery in supervised settings.

5 Convolutional Neural Networks

Motivated by the idea that statistical information in the joint distribution may be associated with the causal direction, Singh et al. applied deep convolutional neural nets on two-dimensional scatter plot images describing the empirical joint distribution between two random variables.

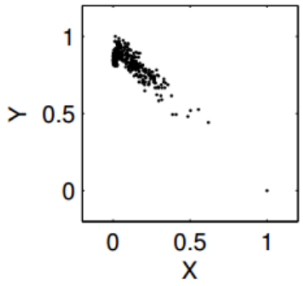

Figure 2: Example scatterplot image fed into a deep convolutional network

On a cause-effect discovery dataset from NIPS 2013, an ensemble version of this approach achieved an AUC of 0.825.

Singh, K., Gupta, G., Vig, L., Shroff, G., & Agarwal, P. (2017). Deep Convolutional Neural Networks for Pairwise Causality. Retrieved from http://arxiv.org/abs/1701.00597

6 Randomized Causation Coefficient

Cause-effect discovery can be viewed as the problem of classifying joint probability distributions. This perspective motivates a two-step approach: first, joint probability distributions are featurized using a kernel mean embedding associated with a characteristic kernel. Second, a binary classifier is trained on the embeddings with a training set of causal direction labels.

Let P be a probability distribution of a random variable Z. The kernel mean embedding of P using a kernel function $k : Z \times Z \to \mathbb{R}$ is:

$$\mu_k(P) = \int_Z k(z, \cdot) dP(z), \tag{14}$$

which is an element in H_k, the Reproducing Kernel Hilbert Space (RKHS) associated with the kernel function k.

If k is a characteristic kernel, the mapping μ_k is a one-to-one function satisfying

$$||\mu_k(P) - \mu_k(Q)||_{H_k} = 0 \iff P = Q. \tag{15}$$

Conveniently, the Gaussian kernel $k(z, z') = \exp(-\gamma||z - z'||_2^2), \gamma > 0$ is a characteristic kernel.

In practice, only finite samples $S = \{z_i\}_{i=1}^n \sim P^n$ are available, which yield the empirical distribution $P_S = \frac{1}{n} \sum_{z_i \in S} \delta(z_i)$ where $\delta(z)$ is a Dirac distribution centered at z. The empirical kernel mean embedding can be used instead:

$$\mu_k(P_S) = \frac{1}{n} \sum_{i=1}^{n} k(z_i, \cdot) \in H_k. \tag{16}$$

6.1 Low-Dimensional Embedding of Empirical Distributions

A training set $\{(\mu_{k,m}(P_{S_i}), l_i\}_{i=1}^n$ can be constructed, where $\mu_{k,m}(P_S)$ is an m-dimensional embedding of the empirical data P_S.

Assume that $Z = \mathbb{R}^d$ and that the kernel function k is real-valued and shift-invariant. Built from k, let $p_k : Z \to \mathbb{R}$ be the positive and integrable Fourier transform of the kernel function k, and let $C_k = \int_Z p_k(w) dw$ be the normalizing constant of the distribution. Then, the m-dimensional embedding is performed using:

$$\begin{aligned} w_j &\sim \frac{1}{C_k} p_k, \\ b_j &\sim U[0, 2\pi], \\ \mu_{k,m}(P_S) &= \frac{2C_k}{|S|} \sum_{z \in S} (\cos(\langle w_j, z \rangle + b_j))_{j=1}^m \in \mathbb{R}^m. \end{aligned} \tag{17}$$

When k is a Gaussian kernel, such that $k(z, z') = \exp(-\gamma||z - z'||_2^2$ with $\gamma > 0$, one can use $p_k(w) = N(w|0, 2\gamma I)$ with $C_k = 1$.

It can be shown that $\mu_{k,m}(P_S)$ is a consistent estimator of $\mu_k(P_S)$.

> Lopez-paz, D., & Sch, B. (2015). Towards a Learning Theory of Cause-Effect Inference. Proceedings of the 32nd International Conference on Machine Learning.

7 Causal Discovery with Proxy Variables

The methods for cause-effect discovery discussed thus far all attempt to estimate the causal direction between two *random* variables X and Y. This requires access to empirical samples from $p(X, Y)$.

Consider the task of cause-effect discovery between two *static* entities x and y. For example, x and y could be original artwork and a fraudulent copy; they could also be two frames in a video ("before" and "after"). In such a setting, there is only one datapoint for x and y.

The cause-effect discovery task of interest here is discovering the causal relationship between two static entities. In the example of original artwork x and a fraudulent copy y, the fraudulent copy is a causal result of the original artwork, i.e., $y \leftarrow f(x, N_x)$. Similarly, in the case of two frames of a video, the "after" frame is a causal result of the "before" frame.

Carulla et al. (2017) propose an approach using proxy variables for this task.

7.1 Proxy Variables for Images

Let x and y be images represented as vectors. Carulla proposes the following procedure: For $n >> 1$ and $j = 1, ..., n$:

1. Generate a random image mask w_j consisting of 1's inside a random rectangle, and 0's elsewhere.
2. Compute $a_j = \langle w_j, x \rangle$ and $b_j = \langle w_j, y \rangle$.

Furthermore, assume that the causal relationship between x and y holds for A and B. Then, applying standard cause-effect discovery methods on A and B suffices to discover the cause-effect relationship between x and y.

The assumption the causal relationship between x and y also holds for A and B merits a closer look. One situation where this assumption could be reasonable is when assuming that the causal mechanism mapping x to y operates "locally" – informally, this could be imagined in an image, where a given subsection of the fraudulent copy is primarily a function of the same subsection in the original image.

In this setting, the random masks w_j are called proxy variables, and they act to "expand" the single static datapoint into a set of samples from a random variable, enabling standard cause-effect discovery methods.

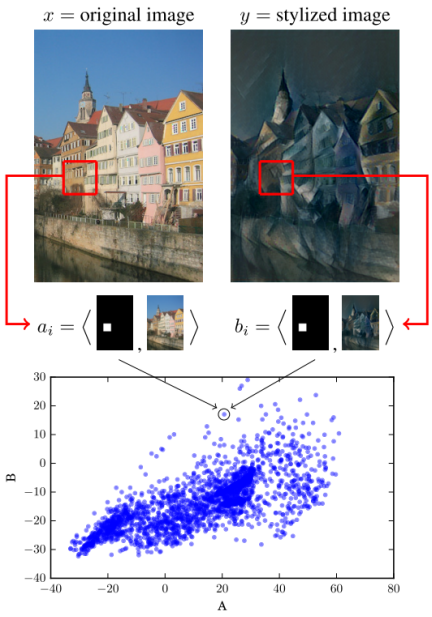

Figure 3: Sampling random patches at paired locations in two causally-related images

7.2 Proxy Variables for Natural Language

Next, Carulla et al. (2017) consider proxy variables between words. They use Amazon Mechanical Turk to build a labeled dataset of tuples (x_i, y_i, l_i) with words x_i and y_i and a human-labeled direction of causality l_i. The authors state that the canonical causal relationship between pairs of words without specific textual contexts is what they desire. Examples include "virus" $\rightarrow$ "death" and "sun" $\rightarrow$ "radiation".

In this setting, one of their proposed procedures is:

1. Generate a random word w_j drawn at frequency proportional to its occurrence in a large natural language corpus.
2. Compute $a_j = \langle \text{word2vec}(w_j), \text{word2vec}(x) \rangle$ and $b_j = \langle \text{word2vec}(w_j), \text{word2vec}(y) \rangle$, where word2vec is used to map a word w_j into a real vector.

Using this procedure followed by standard cause-effect discovery methods previously discussed, they achieve 75% accuracy on the supervised learning task.

> Rojas-carulla, M., Baroni, M., & Lopez-paz, D. (2017). Causal Discovery Using Proxy Variables. arXiv.

8 Causality and Semi-Supervised Learning

Schölkopf et al (2012) point out a connection between causality and semi-supervised learning. Specifically, they claim that whether or not semi-supervised learning is effective for a

given problem depends on the underlying causal model.

In a standard semi-supervised learning setting, consider data $(x_1, y_1), ..., (x_n, y_n) \sim P_{X,Y}$ i.i.d, augmented with additional unlabeled data points $x_{n+1}, ...x_{n+m} \sim P_X$ i.i.d., where X are data and Y are labels.

Semi-supervised learning is said to be effective if the additional m unlabeled data points improves performance on a machine learning model predicting y from x, or in other words, modeling $p(y|x)$.

The argument goes as follows: The additional m unlabeled data-points can only give information about P_X. If the true underlying causal model is $X \to Y$, the causal generative processes are $p(x)$ and $p(y|x)$. In such a setting, it is common and philosophically reasonable to assume that $p(x)$ and $p(y|x)$ are independent causal mechanisms. Therefore, additional information about $p(x)$ is not informative about $p(y|x)$, so semi-supervised learning is thought to be ineffective in such a setting.

On the other hand, if the true underlying causal model is $X \leftarrow Y$, the independent causal generative processes are $p(y)$ and $p(x|y)$. In this setting, additional information about $p(x)$ is not informative about $p(y)$ but can be informative about $p(x|y)$, and semi-supervised learning could be effective in such a setting.

In sum, the claim is that semi-supervised learning is only effective if the additional unlabeled datapoints are additional observations of *effects*, not causes.

8.1 A Big Data Approach for Improved Cause-Effect Discovery

Supervised datasets for cause-effect discovery with labels for causal relationships between random variables are useful for training cause-effect discovery algorithms that are useful in practice.

While increasing the size of these datasets will help improve performance, in practice, labeling the causal direction between random variables can be expensive, requiring expert domain knowledge. Meanwhile, in our big data era, we have access to plenty of empirical data $(x_i, y_i)^N$. Would we be able to use this large source of unlabeled data to improve performance at cause-effect discovery?

Consider two objects: $(x_i, y_i)^N$, empirical data on two random variables; and l_i, the causal relationship between the two random variables X and Y. Between these two objects, a plausible causal relationship between these two could be $l_i \to (x_i, y_i)^N$.

In cause-effect discovery, the task of interest is inferring l_i from $(x_i, y_i)^N$. The ideas from the previous section tells us approaching this task in a semi-supervised manner by obtaining additional unlabeled $(x_i, y_i)^N$, data on the *effect*, can improve performance at cause-effect discovery.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., & Mooij, J. (2012). On Causal and Anticausal Learning. Proceedings of the 29th AAAI Conference on Artificial Intelligence, 8. Retrieved from http://arxiv.org/abs/1206.6471

9 References

Peters, J. (2017) Elements of Causal Inference: Foundations and Learning Algorithms, Final Textbook Draft. https://mitpress.mit.edu/books/elements-causal-inference (Available for free, search for "This is an open access title" on left side of webpage)

Lopez-Paz, D., & Oquab, M. (2017). Revisiting Classifier Two-Sample Tests for GAN Evaluation and Causal Discovery. ICLR 2017.

Lopez-paz, D., & Sch, B. (2015). Towards a Learning Theory of Cause-Effect Inference. Proceedings of the 32nd International Conference on Machine Learning.

Mooij, Peters, Janzing, Zscheischler, & Schlkopf. (2016). Distinguishing cause from effect using observational data: methods and benchmarks, 17, 96.

https://doi.org/10.1109/TSE.2014.2322358

Schoelkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., & Mooij, J. (2012). On Causal and Anticausal Learning. Proceedings of the 29th AAAI Conference on Artificial Intelligence, 8. Retrieved from http://arxiv.org/abs/1206.6471

Singh, K., Gupta, G., Vig, L., Shroff, G., & Agarwal, P. (2017). Deep Convolutional Neural Networks for Pairwise Causality. http://arxiv.org/abs/1701.00597

Gretton, A., Bousquet, O., Smola, A., & Scholkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. Lecture Notes in Computer Science, 3734 LNAI, 6377. https://doi.org/10.1007/11564089_7

Singh, K., Gupta, G., Vig, L., Shroff, G., & Agarwal, P. (2017). Deep Convolutional Neural Networks for Pairwise Causality. http://arxiv.org/abs/1701.00597

Rojas-carulla, M., Baroni, M., & Lopez-paz, D. (2017). Causal Discovery Using Proxy Variables. arXiv.

Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., & Sebag, M. (2017). Causal Generative Neural Networks, 113. Retrieved from http://arxiv.org/abs/1711.08936