NonSENS: Non-Linear SEM Estimation using Non-Stationarity

Ricardo Pio Monti[1], **Kun Zhang**[2], **Aapo Hyvärinen**[1,3]
[1]Gatsby Computational Neuroscience Unit, University College London, UK
[2]Department of Philosophy, Carnegie Mellon University, USA
[3]Department of Computer Science and HIIT, University of Helsinki, Finland

Abstract

We consider the bivariate causal discovery problem. While this problem has been extensively studied, the majority of current methods assume a linear causal relationship, and the few methods which consider non-linear dependencies usually make the assumption of additive noise. Here, we propose a framework through which we can perform causal discovery in the presence of general non-linear relationships. The proposed method exploits a correspondence between a piece-wise stationary non-linear ICA model and non-linear causal models. We show that in the case of bivariate causal discovery, non-linear ICA can be used to infer the causal direction via a series of independence tests. A series of experiments on simulated data demonstrate the capabilities of the proposed method.

Causal models play a fundamental role in modern scientific endeavor [Spirtes et al., 2000, Pearl, 2009]. While randomized control studies are the gold standard, such an approach is unfeasible or unethical in many scenarios [Spirtes and Zhang, 2016]. Furthermore, big data sets publicly available on the internet often try to be generic and thus cannot be strongly based on specific intervention. As such, it is both necessary and important to develop *causal discovery* methods through which to uncover causal structure from (potentially large-scale) passively observed data. Data collected without the explicit manipulation of certain variables is often termed *observational data*, in contrast to experimental data where certain variables are intervened upon, as in randomized controlled trials.

In this work we focus on the bivariate causal discovery problem. This corresponds to recovering the causal structure using observations from two variables, which we denote by X_1 and X_2. While bivariate causal discovery is a (simplified) special case of the more general causal discovery problem, it remains a challenging task. To date, the identification of causal relationships between two variables has primarily been studied under the assumption of linear causal dependencies [Kano and Shimizu, 2003, Hyvärinen and Smith, 2013]. Several extensions to accommodate non-linear causal dependencies have also been proposed, however, the majority of such methods typically assume additive noise [Hoyer et al., 2009, Peters et al., 2014]

Here, we propose a general method for bivariate causal discovery in the presence of general non-linearities. The proposed method is able to uncover non-linear causal relationships without requiring assumptions such as linear causal structure or additive noise. Our approach is based on the correspondence between a piece-wise stationary non-linear ICA model and a non-linear structural equation model (SEM). The proposed method therefore shares similarities with linear-ICA based causal discovery methods [Shimizu et al., 2006]. Under the assumption that we observed bivariate data generated under a specific non-linear ICA model utilizing non-stationarity, we demonstrate that if latent sources can be recovered via non-linear ICA, then a series of independence tests can be employed to uncover causal structure.

1 Background

1.1 Structural equation models

We assume we observe 2-dimensional random variables $\mathbf{X} = (X_1, X_2)$ with joint distribution $\mathbb{P}(\mathbf{X})$. The objective of causal discovery is to use the observed data, which provide an empirical estimate of $\mathbb{P}(\mathbf{X})$, to infer the associated causal graph [Pearl, 2009].

A structural equation model (SEM) is here defined (generalizing the traditional definition) as a pair $(\mathcal{S}, \mathbb{P}(\mathbf{N}))$, where $\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2)$ is a collection of structural equations:

$$\mathcal{S}_j : \quad X_j = f_j(\mathbf{PA}_j, N_j), \quad j \in \{1, 2\} \tag{1}$$

and $\mathbb{P}(\mathbf{N})$ is the joint distribution over disturbance (noise) variables, N_j, which are assumed to be mutually independent. We write $\mathbf{PA}_j$ to denote the parents of the variable X_j. The causal graph, $\mathcal{G}$, associated with a SEM in equation (1) is a graph consisting of one node corresponding to each variable X_j; throughout this work we assume $\mathcal{G}$ is a directed acyclic graph (DAG). We note that functions f_j in equation (1) can be any (possibly non-linear) functions. Special cases of SEMs include: *a)* the linear non-Gaussian acyclic model (LiNGAM; Shimizu et al., 2006), which assumes each f_j is a linear function and the N_j are non-Gaussian, *b)* the additive noise model (ANM; Hoyer et al., 2009), which assumes the noise is additive and *c)* the post-nonlinear causal model, which also captures possible nonlinear distortion in the observed variables [Zhang and Hyvärinen, 2009].

1.2 A piece-wise stationary non-linear ICA model

In this section we briefly outline a flexible class of non-linear ICA models, presented in Hyvärinen and Morioka [2016]. We assume we observe two dimensional data, $\mathbf{X}$, which is generated according to a smooth and invertible non-linear mixture of independent latent variables $\mathbf{S} = (S_1, S_2)$. Formally, we have

$$\mathbf{X} = \mathbf{f}(\mathbf{S}). \tag{2}$$

Following Hyvärinen and Morioka [2016], we further assume that both latent sources and observed data are piece-wise stationary, implying that they may be divided into non-overlapping segments such that their distributions vary across segments, indexed by $e \in \mathcal{E}$. Typically we will have $\mathcal{E} = \{1, \ldots, T\}$ and T is the number of distinct segments which, for example, may correspond to distinct experimental conditions or measurements over distinct subjects. Formally, we assume that while components S_j are mutually independent, the distribution of each component is piece-wise stationary. In particular, Hyvärinen and Morioka [2016] propose to model the log-density of the jth source in segment e as following an exponential family distribution:

$$\log p_e(S_j) = q_{j,0}(S_j) + \lambda_j(e) q_j(S_j) - \log Z(e), \tag{3}$$

where $q_{j,0}$ is a stationary baseline and q_j is a non-linear scalar function defining an exponential family for the jth source. The final term in equation (3) corresponds to a normalization constant. It is important to note that parameters $\lambda_j(e)$ are functions of the segment index, e, implying that the distribution of sources will vary across segments.

It follows from equation (2) that observations $\mathbf{X}$ may also be divided into non-overlapping segments indexed by $e \in \mathcal{E}$. We write $\mathbf{X}(i)$ to denote the ith observation and $C_i \in \mathcal{E}$ to denote its corresponding segment. Throughout this work we assume segment labels, C_i, are known.

We may formally state the relationship between the aforementioned non-linear ICA model and the non-linear SEMs, introduced in Section 1.1, as follows:

> Observations generated according to the piece-wise stationary non-linear ICA model of equations (2) and (3) will follow a (possibly non-linear) SEM where each disturbance variable, N_j, corresponds to a latent source variable, $S_{\pi(j)}$, and each structural equation, f_j, will correspond to an entry of the smooth, invertible function $\mathbf{f}$. We note that due to the permutation indeterminacies typically present in ICA, each disturbance variable N_j will only be identifiable up to some permutation π of the set $\{1, 2\}$.

1.3 Non-linear ICA by contrastive learning

We now briefly describe the Time Contrastive Learning (TCL) algorithm, through which it is possible to unmix latent sources from observed non-linear mixtures, providing the statistical assumptions described in Section 1.2 hold. For further details we refer readers to Hyvärinen and Morioka [2016].

TCL proceeds by defining a multinomial classification task, where we consider each original data point $\mathbf{X}(i) \in \mathbb{R}^d$ as a data point to be classified, and the segment indices C_i give the labels. Given the observations, $\mathbf{X}$, together with the associated segment labels, C, TCL can then be proven to estimate $\mathbf{f}^{-1}$ as well as independent components, $\mathbf{S}$, by learning to classify the observations into their corresponding segments. In particular, TCL trains a deep neural network using multinomial logistic regression to perform this classification task. The network architecture employed consists of a feature extractor corresponding to the last hidden layer, denoted $\mathbf{h}(\mathbf{X}(i); \theta)$ and parameterised by θ, together with a final linear layer. The purpose of the feature extractor is therefore to extract the relevant features which will allow for linear discrimination across segments. Intuitively, it follows that $\mathbf{h}(\mathbf{X}(i); \theta)$ must learn a representation which captures the non-stationary structure across segments. More rigorously, the optimal feature extractor learns to model the differences of the log probability density functions of the segments. The central Theorem on TCL is given in our notation as

Theorem 1 (Hyvärinen and Morioka [2016]) *Assume the following conditions hold:*

1. *We observe data generated by independent sources according to equation (3) and mixed via invertible, smooth function* $\mathbf{f}$ *as stated in equation (2).*
2. *We train a neural network consisting of a feature extractor* $\mathbf{h}(\mathbf{X}(i); \theta)$ *and a final linear layer (i.e., softmax classifier) to classify each observation to its corresponding segment label,* C_i*. We require the dimension of* $\boldsymbol{h}(\boldsymbol{X}(i); \theta)$ *be the same as* $\mathbf{X}(i)$*.*
3. *The matrix* $\mathbf{L}$ *with elements* $\mathbf{L}_{e,j} = \lambda_j(e) - \lambda_j(1)$ *for* $e \in \mathcal{E}$ *and* $j = 1, \ldots, d$ *has full column rank. Recall that* d *is the dimension of each observation* $\boldsymbol{X}(i)$*.*

Then in the limit of infinite data (i.e., infinitely many segments or experimental conditions), the outputs of the feature extractor are equal to $q(\mathbf{S})$*, up to an invertible linear transformation.*

Theorem 1 states that we may perform non-linear ICA by training a neural network to classify the segments from which each observation was generated, followed by linear ICA on the hidden representations, $\mathbf{h}(\mathbf{X}; \theta)$. More generally, the Theorem proves identifiability of this particular piecewise stationary non-linear ICA model, meaning that it is possible to recover the sources. This is not the case with many simpler attempts at non-linear ICA models [Hyvärinen and Pajunen, 1999], such as the case with a single segment in the model presented in Section 1.2 (i.e. data sampled *i.i.d*).

2 Proposed method

In this section we outline the proposed method for causal discovery over bivariate data, which we term **Non**-linear **S**EM **E**stimation using **N**on-**S**tationarity (NonSENS). We assume we observe bivariate data $\mathbf{X}(i) \in \mathbb{R}^2$ where i provides an index over all observations (e.g., i may index time but this is not necessary at all as observations are assumed to be *i.i.d* within each segment). We write $X_1(i)$ and $X_2(i)$ to denote the first and second entries of $\mathbf{X}(i)$ respectively. We will omit the i index whenever it is clear from context. We further assume data is available over a set of distinct environmental conditions or segments, $e \in \mathcal{E}$. As such, each $\mathbf{X}(i)$ is allocated to an experimental condition $C_i \in \mathcal{E}$. We write n_e to denote the number of observations within each experimental condition. The total number of distinct experimental conditions is $|\mathcal{E}| = T$.

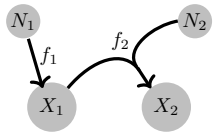
2.1 Non-linear causal discovery using contrastive learning

The objective of the proposed method is to uncover the causal direction between X_1 and X_2. Without loss of generality, we explain the basic logic assuming that $X_1 \to X_2$, such that the associated SEM is of the form:

$$X_1(i) = f_1(N_1(i)), \tag{4}$$
$$X_2(i) = f_2(X_1(i), N_2(i)), \tag{5}$$

where N_1, N_2 are latent disturbances whose distribution is also assumed to vary across experimental conditions. The DAG associated with equations (4) and (5) is shown in Figure 1. To align the proposed method with the terminology of TCL reviewed in Section 1.3, we note that we may consider each experimental condition as a distinct segment. Furthermore, as noted in Section 1.2, the latent disturbances in a bivariate SEM, $\mathbf{N}$, correspond to the independent sources in a non-linear ICA model, $\mathbf{S}$, not unlike in the original LiNGAM theory [Shimizu et al., 2006].

The proposed NonSENS algorithm consists of a two-step procedure. First, it seeks to recover latent disturbances via TCL. We note that the non-stationarity introduced by the various experimental conditions, $e \in \mathcal{E}$, implies that TCL is well suited to recover the source variables N_1 and N_2. Given estimated sources, the following property highlights how we may employ knowledge regarding the statistical independences between observed data and estimated sources in order to infer the causal structure:



$$S_1: \quad X_1 = f_1(N_1)$$
$$S_2: \quad X_2 = f_2(X_1, N_2)$$

Figure 1: Visualization of DAG, $\mathcal{G}$, associated with the SEM in equations (4) and (5). The associated structural equations are provided on the right.

Property 1 *Assume the true causal structure follows equations (4) and (5), as depicted in Figure 1. Then it follows that* $X_1 \perp\!\!\!\perp N_2$ *while* $X_1 \not\perp\!\!\!\perp N_1$ *and* $X_2 \not\perp\!\!\!\perp N_1$ *as well as* $X_2 \not\perp\!\!\!\perp N_2$.

Property 1 highlights the relationship between observations $\mathbf{X}$ and true latent sources, $\mathbf{N}$. In practice, TCL returns estimated latent sources whose ordering is identifiable only up to permutation. As a result, we must test for independence between each of the variables, X_1 and X_2, and each of the estimated latent sources, $\hat{N}_1$ and $\hat{N}_2$. This results in a total of four distinct independence tests. Each test is run at a prespecified significance level, α, and Bonferroni corrected in order to control the family-wise error rate. We employ HSIC as a test for statistical dependence [Gretton et al., 2005], but note that the proposed method can be employed in conjunction with any other suitable independence test. We note TCL is only able to recover $q(\mathbf{N})$ as opposed to $\mathbf{N}$. This is not a problem for the proposed method as we employ a general test of statistical dependence, such as HSIC, which is able to capture arbitrary (i.e., non-linear) dependencies.

The proposed method is thus recapitulated as follows: *After estimating (point-wise transformations of) latent sources by TCL, we make the four tests listed in Property 1, and conclude a causal effect in the case where, for one of the directions, there is evidence to reject the null hypothesis in three of the tests and only one of the tests fails to reject the null.* In such a case, the variable for which the null hypothesis was not rejected is identified as the cause variable.

2.2 Relationship to previous methods

The proposed NonSENS algorithm shares several similarities with prior methods. Perhaps the clearest relationship is with linear ICA-based causal discovery methods such as LiNGAM [Shimizu et al., 2006]. While original LiNGAM methods studied the ICA unmixing matrix in order to perform causal discovery, the use of independence testing in the proposed method means it is more closely aligned with the DirectLiNGAM method [Shimizu et al., 2011].

Hoyer et al. [2009] and Peters et al. [2014] propose a non-linear causal discovery method named regression and subsequent independence test (RESIT) which is able to recover the causal structure under the assumption of an additive noise model. RESIT shares the same underlying idea as the proposed method, with the difference being that it estimates latent disturbances via non-linear regression, as opposed to via non-linear ICA. One advantage of RESIT is that it only requires two independence tests, as there are no permutation related indeterminacies of the estimated latent disturbances. However, a limitation of the RESIT algorithm is that it cannot accommodate non-additive causal structure.

The proposed method also shares important similarities with the Invariant Causal Prediction (ICP) method presented by Peters et al. [2016], as both methods seek to exploit the piece-wise stationary nature of observational data in order to perform causal discovery. Formally, the ICP algorithm exploits the invariance of causal models under covariate shift in order to recover the true causal structure.

The intuition behind such an approach is that the conditional distribution of an effect given its direct causes will remain identical (i.e., invariant) given interventions on any variable other than the effect.

Finally, Zhang et al. [2017] also propose a method for causal discovery in the context of non-stationary data, termed CD-NOD. Their proposed method accounts for non-stationarity via the introduction of an surrogate variable representing the time or domain index into the causal DAG. Conditional independence testing is then employed to recover the skeleton over the augmented DAG.

3 Experimental results

In order to experimentally validate the capabilities of the proposed method, we generate synthetic data from the non-linear ICA model detailed in Section 1.2. Non-stationary disturbances, $\mathbf{N}$, were randomly generated by simulating Laplace random variables with distinct variances in each segment. This corresponds to setting $q_j(S_j) = |S_j|$ for all j. For the non-linear mixing function we employ a deep neural network ("mixing-DNN") with randomly generated weights such that:

$$\mathbf{X}^{(1)} = \mathbf{A}^{(1)}\mathbf{N}, \tag{6}$$

$$\mathbf{X}^{(l)} = \mathbf{A}^{(l)} f\left(\mathbf{X}^{(l-1)}\right), \tag{7}$$

where we write $\mathbf{X}^{(l)}$ to denote the activations at the lth layer and f corresponds to the leaky-ReLU activation function which is applied element-wise. We restrict matrices $\mathbf{A}^{(l)}$ to be lower-triangular in order to introduce acyclic causal relations. Note that equation (6) alone would be a LiNGAM. For depths $l \geq 2$, equation (7) generates bivariate data with non-linear causal dependencies.

Throughout the experiments we vary the following factors: the number of distinct experimental conditions (i.e., the number of distinct segments), the number of observations per segment, n_e, as well as the depth, l, of the mixing-DNN.

3.1 Implementation details

Our implementation of the NonSENS algorithm employed deep neural networks of varying depths as feature extractors. Following Hyvärinen and Morioka [2016], we employ maxout activation units within each layer of the network. All networks were trained on cross-entropy loss using stochastic gradient descent with minibatches of size 512. Although the proposed method may employ any independence test, we employ HSIC with a Gaussian kernel, as is the case in many alternative causal discovery algorithms [Hoyer et al., 2009, Zhang and Hyvärinen, 2009, Peters et al., 2014]. All tests are performed at the $\alpha = 5\%$ level and Bonferroni corrected to control the family-wise error rate.

We benchmark the performance of the proposed method against several state-of-the-art causal discovery algorithms. As a measure of performance against linear methods we compare against LiNGAM. In particular, we implemented the DirectLiNGAM version as introduced by Shimizu et al. [2011], with the small difference that HSIC was also employed as a measure of statistical dependence instead of mutual information. In order to highlight the need for non-linear ICA methods, we also consider the performance of the proposed method where linear ICA is employed to estimate latent disturbances. We refer to this baseline as linear-ICA NonSENS. We further compare against RESIT, where we employ Gaussian process regression and HSIC as a measure of statistical dependence, as well as the ICP[1] algorithm [Peters et al., 2016]. We also compare against the CD-NOD method presented in Zhang et al. [2017].

3.2 Results

Figure 2 shows results for bivariate causal discovery as the number of distinct experimental conditions increases and the number of observations within each condition was fixed at $n_e = 512$. Each horizontal panel shows the results as the depth of the mixing-DNN increased from $l = 1$ to $l = 5$. Each panel visualizes the proportion of times the correct cause variable was identified across 100 independent simulations. We note that as there are three potential outcomes: $X_1 \to X_2$, $X_2 \to X_1$

[1] We employ the extension of the ICP algorithm to non-linear additive noise models, presented in Section 6.1 of Peters et al. [2016]. This involves fitting non-linear regression models and testing whether residuals have the same distribution within each experimental condition.

or the causal structure is inconclusive. As such, it is unclear how to define the baseline performance for a random algorithm.

The first panel of Figure 2 corresponds to a LiNGAM where all dependencies are linear. As such, all methods are able to accurately recover the true cause variable. As the depth of the mixing-DNN increases the causal dependencies become increasingly non-linear and the performance of all methods deteriorates. While we attribute this drop in performance to the increasingly non-linear nature of causal structure, we note that the proposed method is able to out-perform all alternative methods.

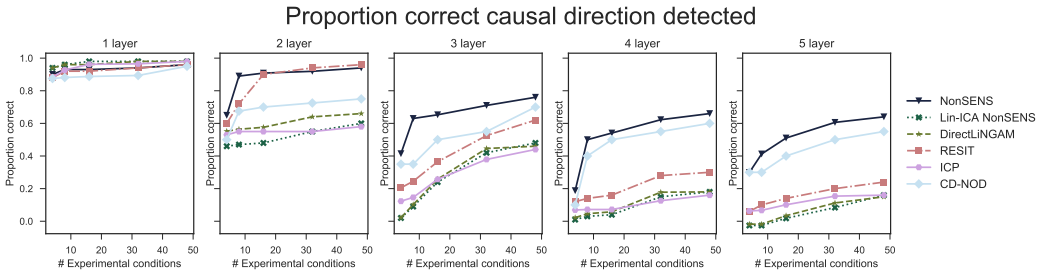


Figure 2: Experimental results indicating performance as we increase the number of experimental conditions, $|\mathcal{E}|$, whilst keeping the number of observation per condition fixed at $n_e = 512$. Each panel plots the proportion of times the correct cause variable is identified for varying depths of the mixing-DNN, ranging from $l = 1, \dots 5$.

We also consider the performance of all algorithms in the context of a fixed number of experimental conditions, , $|\mathcal{E}| = 10$, and an increasing number of observations per condition, n_e. These results are presented in Figure 3. As before, we note that all algorithms are able to accurately recover the true causal structure in the context of linear dependencies, as shown in the first panel. However, as the dependencies become increasingly non-linear there is a significant drop in the performance of all methods. It is important to note that the proposed method typically performs poorly when the number of observations per condition, n_e, is small. This is especially visible in the context of non-linear dependencies. We attribute this behavior to the fact that the NonSENS algorithm must train a neural network as a feature extractor, which will typically require a large number of observations.

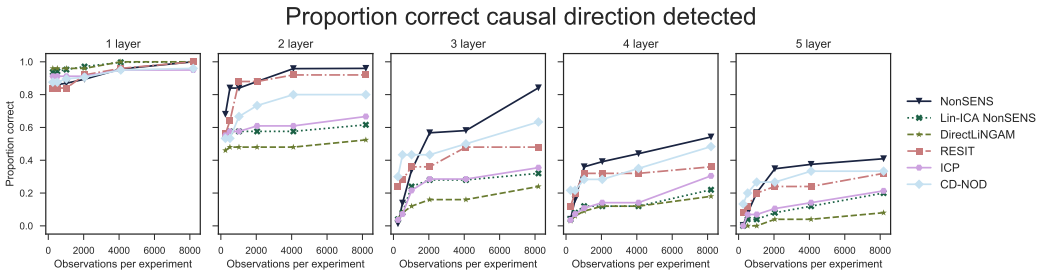


Figure 3: Experimental results indicating performance as we increase the number of observations, n_e conditions, within each experimental condition for a fixed number of experimental conditions, $|\mathcal{E}| = 10$. Each plane plots the proportion of times the correct cause variable is identified for varying depths of the mixing-DNN, ranging from $l = 1, \dots, 5$.

4 Conclusion

We proposed a method for bivariate causal discovery which exploits the correspondence between a piece-wise stationary non-linear ICA model and a non-linear SEM. The proposed method leverages the non-stationarity of data in order to uncover the underlying latent sources. While non-stationarity is crucial for the identifiability of the associated non-linear ICA model, it also constitutes a potential limitation as data is required from distinct experimental conditions which may not always be available. Future work will seek to perform causal discovery by leveraging alternative non-linear ICA methods which exploit statistical properties other than non-stationarity, for example as described in Hyvärinen and Morioka [2017].

References

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring Statistical Dependence with Hilbert-Schmidt Norms. *Int. Conf. Algorithmic Learn. Theory*, pages 63–77.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. *Neural Inf. Process. Syst.*, pages 689–696.

Hyvärinen, A. and Morioka, H. (2016). Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. *Neural Inf. Process. Syst.*

Hyvärinen, A. and Morioka, H. (2017). Nonlinear ICA of Temporally Dependent Stationary Sources. *AISTATS*, 54:460–469.

Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models. *J. Mach. Learn. Res.*, 14:111–152.

Kano, Y. and Shimizu, S. (2003). Causal inference using nonnormality. In *Int. Symp. Sci. Model. 30th Anniv. Inf. Criterion*.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. Ser. B (Statistical Methodology)*, 78:947–1012.

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, 15:2009–2053.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *J. Mach. Learn. Res.*, 7:2003–2030.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *J. Mach. Learn. Res.*, 12:1225–1248.

Spirtes, P., Glymour, C., Scheines, R., Heckerman, D., Meek, C., and Richardson, T. (2000). *Causation, Prediction and Search*. MIT Press.

Spirtes, P. and Zhang, K. (2016). Causal discovery and inference: concepts and recent methodological advances. *Appl. Informatics*.

Zhang, K., Huang, B., Zhang, J., Glymour, C., and Schölkopf, B. (2017). Causal discovery from Nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Int. Jt. Conf. Artif. Intell.*, pages 1347–1353.

Zhang, K. and Hyvärinen, A. (2009). On the Identifiability of the Post-Nonlinear Causal Model. *Proc. Twenty-Fifth Conf. Uncertain. Artif. Intell.*