Causal Discovery with General Non-Linear Relationships Using Non-Linear ICA

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 problem of inferring causal relationships between two or more passively observed variables. While the problem of such causal discovery has been extensively studied, especially in the bivariate setting, the majority of current methods assume a linear causal relationship, and the few methods which consider non-linear relations 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 is based on recent progress in non-linear independent component analysis (ICA) and exploits the non-stationarity of observations in order to recover the underlying sources. We show rigorously that in the case of bivariate causal discovery, such non-linear ICA can be used to infer causal direction via a series of independence tests. We further propose an alternative measure for inferring causal direction based on asymptotic approximations to the likelihood ratio, as well as an extension to multivariate causal discovery. We demonstrate the capabilities of the proposed method via a series of simulation studies and conclude with an application to neuroimaging data.

1 INTRODUCTION

Causal models play a fundamental role in modern scientific endeavor (Pearl, 2009). While randomized control studies are the gold standard, such an approach is unfeasible or unethical in many scenarios (Spirtes and Zhang, 2016). Even when it is possible to run randomized control trials, the number of experiments required may raise practical challenges (Eberhardt et al., 2005). Furthermore, big data sets publicly available on the internet often try to be generic and thus cannot be strongly based on specific interventions; a prominent example is the Human Connectome Project which collects resting state fMRI data from over 500 subjects (Van Essen et al., 2012). As such, it is important to develop *causal discovery* methods through which to uncover causal structure from (potentially large-scale) passively observed data. Such data collected without explicit manipulation of certain variables is often termed *observational data*.

The intrinsic appeal of causal discovery methods is that they allow us to uncover the underlying causal structure of complex systems, providing an explicit description of the underlying generative mechanisms. Within the context of machine learning, causal knowledge has also been shown to play an important role in many domains such as semi-supervised and transfer learning (Schölkopf et al., 2012; Zhang et al., 2013), covariate shift and algorithmic fairness (Kusner et al., 2017). A wide range of methods have been proposed to discover casual knowledge (Shimizu et al., 2006; Hoyer et al., 2009; Zhang and Hyvärinen, 2009; Peters et al., 2016; Zhang et al., 2017). However, many of the current methods rely on restrictive assumptions regarding the nature of the causal relationships. For example, Shimizu et al. (2006) assume linear causal models with non-Gaussian disturbances and demonstrate that independent component analysis (ICA) may be employed to uncover causal structure. Hoyer et al. (2009) provide an extension to non-linear causal models but under the assumption of additive noise.

In this paper 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 ap-

KZ acknowledges the support by the U.S. Air Force under Contract No. FA8650-17-C-7715 and by NSF EAGER Grant No. IIS-1829681. The U.S. Air Force and the NSF are not responsible for the views reported in this article.

proach exploits a correspondence between a non-linear ICA model and non-linear causal models, and is specifically tailored for observational data which are collected across a series of distinct experimental conditions or regimes. Given such data, we seek to exploit the non-stationarity introduced via distinct experimental conditions in order to perform causal discovery. We demonstrate that if latent sources can be recovered via non-linear ICA, then a series of independence tests may be employed to uncover causal structure. As an alternative to independence testing, we further propose a novel measure of non-linear causal direction based on an asymptotic approximation to the likelihood ratio.

2 PRELIMINARIES

In this section we introduce the class of causal models to be studied. We also present an overview of non-linear ICA methods based on contrastive learning, upon which we base the proposed method.

2.1 MODEL DEFINITION

Suppose we observe d-dimensional random variables $\mathbf{X} = (X_1, \ldots, X_d)$ with joint distribution $\mathbb{P}(\mathbf{X})$. The objective of causal discovery is to use the observed data, which give the empirical version of $\mathbb{P}(\mathbf{X})$, to infer the associated causal graph which describes the data generating procedure (Spirtes et al., 2000; Pearl, 2009).

A structural equation model (SEM) is here defined (generalizing the traditional definition) as a collection of d structural equations:

$$X_j = f_j(\mathbf{PA}_j, N_j), \quad j = 1, \ldots, d \tag{1}$$

together with a joint distribution, $\mathbb{P}(\mathbf{N})$, 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).

While functions f_j in equation (1) can be any (possibly non-linear) functions, to date the causal discovery community has focused on specific special cases in order to obtain identifiability results as well as provide practical algorithms. Pertinent examples 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 non-linear distortion in the observed variables (Zhang and Hyvärinen, 2009).

The aforementioned approaches enforce strict constraints on the functional class of the SEM. Otherwise, without suitable constraints on the functional class, for any two variables one can always express one of them as a function of the other and independent noise (Hyvärinen and Pajunen, 1999). We are motivated to develop novel causal discovery methods which benefit from new identifiability results established from a different angle, in the context of general non-linear (and non-additive) relationships. A key component of our method exploits some recent advances in non-linear ICA, which we review next.

2.2 NON-LINEAR ICA VIA TCL

We briefly outline the recently proposed Time Contrastive Learning (TCL) algorithm, through which it is possible to demix (or disentangle) latent sources from observed non-linear mixtures; this algorithm provides hints as to the identifiability of causal direction between two variables in general non-linear cases under certain assumptions and is exploited in our causal discovery method. For further details we refer readers to Hyvärinen and Morioka (2016) but we also provide a brief review in Supplementary Material A. We assume we observe d-dimensional data, $\mathbf{X}$, which are generated according to a smooth and invertible non-linear mixture of independent latent variables $\mathbf{S} = (S_1, \ldots, S_d)$. In particular, we have

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

The goal of non-linear ICA is then to recover $\mathbf{S}$ from $\mathbf{X}$.

TCL, as introduced by Hyvärinen and Morioka (2016), is a method for non-linear ICA which is premised on the assumption that both latent sources and observed data are non-stationary time series. Formally, they assume that while components S_j are mutually independent, the distribution of each component is piece-wise stationary, implying they can be divided into non-overlapping time segments such that their distribution varies across segments, indexed by $e \in \mathcal{E}$. In the basic case, the log-density of the jth latent source in segment e is assumed to follow an exponential family distribution such that:

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

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. (Exponential families with more than one sufficient statistic are also allowed.) 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.

TCL proceeds by defining a multinomial classification task, where we consider each original data point $\mathbf{X}(i)$ 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 recover $\mathbf{f}^{-1}$ 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 by $\mathbf{h}(\mathbf{X}(i); \theta)$ and parameterised by θ, together with a final linear layer. 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 segments* $e = 1, \ldots, E$ *and* $j = 1, \ldots, d$ *has full rank.*

Then in the limit of infinite data, 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 associated with each observation, followed by linear ICA on the hidden representations, $\mathbf{h}(\mathbf{X}; \theta)$. This theorem provides identifiability of this particular 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 above.

While Theorem 1 provides identifiability for a particular non-linear ICA model, it requires a final linear unmixing of sources (i.e., via linear ICA). However, when sources follow the piece-wise stationary distribution detailed in equation (3), traditional linear ICA methods may not be appropriate as sources will only be independent conditional on the segment. For example, it is possible that exponential family parameters, $\lambda_j(e)$, are dependent across sources (e.g., they may be correlated). This problem will be particularly pertinent when data is only collected over a reduced number of segments. As such, alternative linear ICA algorithms are required to effectively employ TCL in such a setting, as addressed in Section 3.2.

3 NON-LINEAR CAUSAL DISCOVERY VIA NON-LINEAR ICA

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 begin by providing an intuition for the proposed method in Section 3.1, which is based on the connection between non-linear ICA and non-linear SEMs. In Section 3.2 we propose a novel linear ICA algorithm which complements TCL for the purpose of causal discovery, particularly in the presence of observational data with few segments. Our method is formally detailed in Section 3.3, which also contains a proof of identifiability. Finally in Section 3.4 we present an alternative measure of causal direction based on asymptotic approximations to the likelihood ratio of non-linear causal models.

3.1 RELATING SEM TO ICA

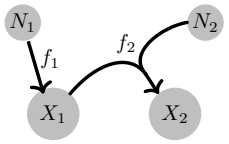
We assume we observe bivariate data $\mathbf{X}(i) \in \mathbb{R}^2$ and write $X_1(i)$ and $X_2(i)$ to denote the first and second entries of $\mathbf{X}(i)$ respectively. We will omit the i index whenever it is clear from context. Following the notation of Peters et al. (2016), we further assume data is available over a set of distinct environmental conditions $\mathcal{E} = \{1, \ldots, E\}$. As such, each $\mathbf{X}(i)$ is allocated to an experimental condition denoted by $C_i \in \mathcal{E}$. Let n_e be the number of observations within each experimental condition such that $n_{tot} = \sum_{e \in \mathcal{E}} n_e$.

The objective of the proposed method is to uncover the causal direction between X_1 and X_2. Suppose 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 distributions are also assumed to vary across experimental conditions. The DAG associated with equations (4) and (5) is shown in Figure 1. Fundamentally, the proposed NonSENS algorithm exploits the correspondence between the non-linear ICA model described in Section 2.2 and non-linear SEMs. This correspondence is formally stated as follows: observations generated according to the (possibly non-linear) SEM detailed in equations (4) and (5) will follow a non-linear ICA model where each disturbance variable, N_j, corresponds to a latent source, $S_{\pi(j)}$.



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

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

Moreover, structural equations f_1 and f_2 jointly define a bivariate non-linear mapping from sources to observations as in non-linear ICA. However, the mixing function $\mathbf{f}$ in non-linear ICA is not exactly the same as f_1 and f_2 (see Supplementary Material B). We note that due to the permutation indeterminacy present in ICA, each disturbance variable, N_j, will only be identifiable up to some permutation π of the set $\{1, 2\}$.

The proposed method consists of a two-step procedure. First, it seeks to recover latent disturbances via non-linear ICA. Given the estimated latent disturbances, the following property highlights how we may employ statistical independencies between observations and estimated sources in order to infer the causal structure:

Property 1 *Assume the true causal structure follows equations (4) and (5), as depicted in Figure 1. Then, assuming each observed variable is statistically dependent on its latent disturbance (thus avoiding degenerate cases), 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* $X_2 \not\perp\!\!\!\perp N_2$. [1]

Property 1 highlights the relationship between observations $\mathbf{X}$ and latent sources, $\mathbf{N}$, and provides some insight into how a non-linear ICA method, together with independence testing, could be employed to perform bivariate causal discovery. This is formalized in Section 3.3.

3.2 A LINEAR ICA ALGORITHM FOR PIECE-WISE STATIONARY SOURCES

Before proceeding, we have to improve the non-linear ICA theory of Hyvärinen and Morioka (2016). Assumptions 1–3 of Theorem 1 for TCL guarantee that the feature extractor, $\mathbf{h}(\mathbf{X}; \theta)$, will recover a linear mixture of latent independent sources (up to element-wise transformation by q). As a result, applying a linear unmixing method to the final representations, $\mathbf{h}(\mathbf{X}; \theta)$, will allow us to recover latent disturbances. However, the use of ordinary linear ICA to unmix $\mathbf{h}(\mathbf{X}; \theta)$ is premised on the assumption that latent sources are independent. This is not necessarily guaranteed when sources follow the ICA model presented in equation (3) with a fixed number of segments. For example, it is possible that parameters $\lambda_j(e)$ are correlated across segments. We note that this is not a problem when the number of segments increases asymptotically and parameters $\lambda_j(e)$ are assumed to be randomly generated, as stated in Corollary 1 of Hyvärinen and Morioka (2016).

In order to address this issue, we propose an alternative linear ICA algorithm to be employed in the final stage of TCL, through which to accurately recover latent sources in the presence of a small number of segments.

The proposed linear ICA algorithm explicitly models latent sources as following the piece-wise stationary distribution specified in equation (3). We write $\mathbf{Z}(i) \in \mathbb{R}^d$ to denote the ith observation, generated as a linear mixtures of sources: $\mathbf{Z}(i) = \mathbf{A}\mathbf{S}(i)$, where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a square mixing matrix. Estimation of parameters proceeds via score matching (Hyvärinen, 2005), which yields an objective function of the following form:

$$J = \sum_{e \in \mathcal{E}} \sum_{j=1}^{d} \lambda_j(e) \frac{1}{n_e} \sum_{C_i = e} q_j''(\mathbf{w}_j^T \mathbf{Z}(i))$$
$$+ \frac{1}{2} \sum_{e \in \mathcal{E}} \sum_{j,k=1}^{d} \lambda_k(e) \lambda_j(e) \mathbf{w}_k^T \mathbf{w}_j \frac{1}{n_e} \sum_{C_i = e} q_k'(\mathbf{w}_k^T \mathbf{Z}(i)) q_j'(\mathbf{w}_j^T \mathbf{Z}(i)),$$

where $\mathbf{W} \in \mathbb{R}^{d \times d}$ denotes the unmixing matrix and q_j' and q_j'' denote the first and second derivatives of the non-linear scalar functions introduced in equation (3). Details and results are provided in Supplementary C, where the proposed method is shown to outperform both FastICA and Infomax ICA, as well as the joint diagonalization method of Pham and Cardoso (2001), which is explicitly tailored for non-stationary sources.

3.3 CAUSAL DISCOVERY USING INDEPENDENCE TESTS

Now we give the outline of NonSENS. NonSENS performs causal discovery by combining Property 1 with a non-linear ICA algorithm. Notably, we employ TCL, described in Section 2.2, with the important addition that the final linear unmixing of the hidden representations, $\mathbf{h}(\mathbf{X}; \theta)$, is performed using the objective given in Section 3.2. The proposed method is summarized as follows:

1. (a) Using TCL, train a deep neural network with feature extractor $\mathbf{h}(\mathbf{X}(i); \theta)$ to accurately classify each observation $\mathbf{X}(i)$ according to its segment label C_i.
   (b) Perform linear unmixing of $\mathbf{h}(\mathbf{X}; \theta)$ using the algorithm presented in Section 3.2.

[1] We note that the property that effect is dependent on its direct causes typically holds, although one may construct specific examples (with discrete variables or continuous variables with complex causal relations) in which effect is independent from its direct causes. In particular, if faithfulness is assumed (Spirtes et al., 2000), the above property clearly holds.

2. Perform the four tests listed in Property 1, and conclude a cause-effect relationship in the case where there is evidence to reject the null hypothesis in three of the tests and only one of the tests fails to reject the null. The variable for which the null hypothesis was not rejected is considered the cause.

Each test is run at a pre-specified significance level, α, and Bonferroni corrected in order to control the family-wise error rate. Throughout this work we employ HSIC as a test for statistical independence (Gretton et al., 2005). Pseudo-code is provided in Supplementary G. Theorem 2 formally states the assumptions and identifiability properties of the proposed method.

Theorem 2 *Assume the following conditions hold:*

1. *We observe bivariate data* $\boldsymbol{X}$ *which has been generated from a non-linear SEM with smooth non-linearities and no hidden confounders.*
2. *Data is available over at least three distinct experimental conditions and latent disturbances,* N_j*, are generated according to equation (3).*
3. *We employ TCL, with a sufficiently deep neural network as the feature extractor, followed by linear ICA (as described in Section 3.2) on hidden representations to recover the latent sources.*
4. *We employ an independence test which can capture any type of departure from independence, for example HSIC, with Bonferroni correction and significance level* α*.*

Then in the limit of infinite data the proposed method will identify the cause variable with probability $1 - \alpha$*.*

See Supplementary D for a proof. Theorem 2 extends previous identifiability results relying on constraints on functional classes (e.g., ANM in Hoyer et al. (2009)) to the domain of arbitrary non-linear models, under further assumptions on nonstationarity of the given data.

3.4 LIKELIHOOD RATIO-BASED MEASURES OF CAUSAL DIRECTION

While independence tests are widely used in causal discovery, they may not be statistically optimal for deciding causal direction. In this section, we further propose a novel measure of causal direction which is based on the likelihood ratio under non-linear causal models, and which thus is likely to be more efficient.

The proposed measure can be seen as the extension of linear measures of causal direction, such as those proposed by Hyvärinen and Smith (2013), to the domain of non-linear SEMs. Briefly, Hyvärinen and Smith (2013) consider the likelihood ratio between two candidate models of causal influence: $X_1 \to X_2$ or $X_2 \to X_1$. The log-likelihood ratio is then defined as the difference in log-likelihoods under each model:

$$R = L_{1\to 2} - L_{2\to 1} \tag{6}$$

where we write $L_{1\to 2}$ to denote the log-likelihood under the assumption that X_1 is the causal variable and $L_{2\to 1}$ for the alternative model. Under the assumption that $X_1 \to X_2$, it follows that the underlying SEM is of the form described in equations (4) and (5). The log-likelihood for a single data point may thus be written as

$$L_{1\to 2} = \log P_{X_1}(X_1) + \log P_{X_2|X_1}(X_2|X_1).$$

Furthermore, in the context of linear causal models we have that equations (4) and (5) define a bijection between N_2 and X_2 whose Jacobian has unit determinant, such that the log-likelihood can be expressed as:

$$L_{1\to 2} = \log P_{X_1}(X_1) + \log P_{N_2}(N_2).$$

In the asymptotic limit we can take the expectation of log-likelihood, and the log-likelihood converges to:

$$\mathbb{E}[L_{1\to 2}] = -H(X_1) - H(N_2) \tag{7}$$

where $H(\cdot)$ denotes the differential entropy. Hyvärinen and Smith (2013) note that the benefit of equation (7) is that only univariate approximations of the differential entropy are required. In this section we seek to derive equivalent measures for causal direction without the assumption of linear causal effects. Recall that after training via TCL, we obtain an estimate of $\mathbf{g} = \mathbf{f}^{-1}$ which is parameterized by a deep neural network.

In order to compute the log-likelihood, $L_{1\to 2}$, we consider the following change of variables:

$$\begin{pmatrix} X_1 \\ N_2 \end{pmatrix} = \tilde{\mathbf{g}} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ \mathbf{g}_2(X_1, X_2) \end{pmatrix}$$

where we note that $\mathbf{g}_2 : \mathbb{R}^2 \to \mathbb{R}$ refers to the second component of $\mathbf{g}$. Further, we note that the the mapping $\tilde{\mathbf{g}}$ only applies the identity to the first element, thereby leaving X_1 unchanged. Given such a change of variables, we may evaluate the log-likelihood as follows:

$$L_{1\to 2} = \log p_{X_1}(X_1) + \log p_{N_2}(N_2) + \log |\det \mathbf{J}\tilde{\mathbf{g}}|,$$

where $\mathbf{J}\tilde{\mathbf{g}}$ denotes the Jacobian of $\tilde{\mathbf{g}}$, as we have $X_1 \perp\!\!\!\perp N_2$ by construction under the assumption that $X_1 \to X_2$.

Due to the particular choice of $\tilde{\mathbf{g}}$, we are able to easily evaluate the Jacobian, which can be expressed as:

$$\mathbf{J}\tilde{\mathbf{g}} = \begin{pmatrix} \frac{\partial \tilde{\mathbf{g}}_1}{\partial X_1} & \frac{\partial \tilde{\mathbf{g}}_1}{\partial X_2} \\ \frac{\partial \tilde{\mathbf{g}}_2}{\partial X_1} & \frac{\partial \tilde{\mathbf{g}}_2}{\partial X_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{\partial \mathbf{g}_2}{\partial X_1} & \frac{\partial \mathbf{g}_2}{\partial X_2} \end{pmatrix}.$$

As a result, the determinant can be directly evaluated as $\frac{\partial \mathbf{g}_2}{\partial X_2}$. Furthermore, since $\mathbf{g}_2$ is parameterized by a deep network, we can directly evaluate its derivative with respect to X_2. This allows us to directly evaluate the log-likelihood of X_1 being the causal variable as:

$$L_{1\to 2} = \log p_{X_1}(X_1) + \log p_{N_2}(N_2) + \log \left| \frac{\partial \mathbf{g}_2}{\partial X_2} \right|.$$

Finally, we consider the asymptotic limit and obtain the non-linear generalization of equation (7) as:

$$\mathbb{E}[L_{1\to 2}] = -H(X_1) - H(N_2) + \mathbb{E}\left[\log \left| \frac{\partial \mathbf{g}_2}{\partial X_2} \right|\right].$$

In practice we use the sample mean instead of the expectation.

One remaining issue to address is the permutation invariance of estimated sources (note this this permutation is not about the causal order of the observed variables). We must consider both permutations π of the set $\{1, 2\}$. In order to resolve this issue, we note that if the true permutation is $\pi = (1, 2)$, then assuming $X_1 \to X_2$, we have $\frac{\partial \mathbf{g}_1}{\partial X_2} = 0$ while $\frac{\partial \mathbf{g}_2}{\partial X_2} \neq 0$. This is because $\mathbf{g}_1$ unmixes observations to return the latent disturbance for causal variable, X_1, and is therefore not a function of X_2. The converse is true if the permutation is $\pi = (2, 1)$. Similar reasoning can be employed for the reverse model: $X_2 \to X_1$. As such, we propose to select the permutation as follows:

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \left\{ \mathbb{E}\left[\log \left| \frac{\partial \mathbf{g}_{\pi(2)}}{\partial X_2} \right|\right] + \mathbb{E}\left[\log \left| \frac{\partial \mathbf{g}_{\pi(1)}}{\partial X_1} \right|\right] \right\}.$$

For a chosen permutation, π^*, we may therefore compute the likelihood ratio in equation (6) as:

$$\begin{aligned} R = &-H(X_1) - H(N_{\pi^*(2)}) + \mathbb{E}\left[\log \left| \frac{\partial \mathbf{g}_{\pi^*(2)}}{\partial X_2} \right|\right] \\ &+H(X_2) + H(N_{\pi^*(1)}) - \mathbb{E}\left[\log \left| \frac{\partial \mathbf{g}_{\pi^*(1)}}{\partial X_1} \right|\right]. \end{aligned}$$

If R is positive, we conclude that X_1 is the causal variable, whereas if R is negative X_2 is reported as the causal variable. When computing the differential entropy, we employ the approximations described in Kraskov et al. (2004). We note that such approximations require variables to be standardized; in the case of latent variables this can be achieved by defining a further change of variables corresponding to a standardization.

Finally, we note that the likelihood ratio presented above can be connected to the independence measures employed in Section 3.3 when mutual information is used a measure of statistical dependence. In particular, we have

$$R = -I(X_1, N_{\pi(2)}) + I(X_2, N_{\pi(1)}), \tag{8}$$

where $I(\cdot, \cdot)$ denotes the mutual information between two variables. We provide a full derivation in Supplementary E. This result serves to connect the proposed likelihood ratio to independence testing methods for causal discovery which use mutual information.

3.5 EXTENSION TO MULTIVARIATE DATA

It is not straightforward to extend NonSENS to multivariate cases. Due to the permutation invariance of sources, we would require d^2 independence tests, where d is the number of variables, leading to a significant drop in power after Bonferroni correction. Likewise, the likelihood ratio test inherently considers only two variables.

Instead, we propose to extend to proposed method to the domain of multivariate causal discovery by employing it in conjunction with a traditional constraint based method such as the PC algorithm, as in Zhang and Hyvärinen (2009). Formally, the PC algorithm is first employed to estimate the skeleton and orient as many edges as possible. Any remaining undirected edges are then directed using either proposed bivariate method.

3.6 RELATIONSHIP TO PREVIOUS METHODS

NonSENS is closely related to linear ICA-based methods as described in Shimizu et al. (2006). However, there are important differences: LiNGAM focuses exclusively on linear causal models whilst NonSENS is specifically designed to recover arbitrary non-linear causal structure. Moreover, the proposed method is mainly designed for bivariate causal discovery whereas the original LiNGAM method can easily perform multivariate causal discovery by permuting the estimated ICA unmixing matrix. In this sense NonSENS is more closely aligned to the Pairwise LiNGAM method (Hyvärinen and Smith, 2013).

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 essentially shares the same underlying idea as NonSENS, with the difference being that it estimates latent disturbances via non-linear regression, as opposed to via non-linear ICA. Related is the Regression Error Causal Inference (RECI) algorithm (Blöbaum et al., 2018), which proposes measures of causal direction based on the magnitude of (non-linear) regression errors. Importantly, both of those methods restrict the non-linear relations to have additive noise.

Recently several methods have been proposed which seek to exploit non-stationarity in order to perform causal discovery. Following Schölkopf et al. (2012), Peters et al. (2016) propose to leverage the invariance of causal models under covariate shift in order to recover the true causal structure. Their method, termed Invariant Causal

Prediction (ICP), is tailored to the setting where data is collected across a variety of experimental regimes, similar to ours. However, their main results, including identifiability are in the linear or additive noise settings.

Zhang et al. (2017) proposed a method, termed CD-NOD, for causal discovery from heterogeneous, multiple-domain data or non-stationary data, which allows for general non-linearities. Their method thus solves a problem similar to ours, although with a very different approach. Their method accounts for non-stationarity, which manifests itself via changes in the causal modules, via the introduction of an surrogate variable representing the domain or time index into the causal DAG. Conditional independence testing is employed to recover the skeleton over the augmented DAG, and their method does not produce an estimate of the SEM to represent the causal mechanism.

4 EXPERIMENTAL RESULTS

In order to demonstrate the capabilities of the proposed method we consider a series of experiments on synthetic data as well as real neuroimaging data.

4.1 SIMULATIONS ON ARTIFICIAL DATA

In the implementation of the proposed method we employed deep neural networks of varying depths as feature extractors. All networks were trained on cross-entropy loss using stochastic gradient descent. In the final linear unmixing required by TCL, we employ the linear ICA model described in Section 3.2. For independence testing, we employ HSIC with a Gaussian kernel. All tests are run at the $\alpha = 5\%$ level and Bonferroni corrected.

We benchmark the performance of the NonSENS algorithm against several state-of-the-art methods. As a measure of performance against linear methods we compare against LiNGAM. In particular, we compare performance to DirectLiNGAM (Shimizu et al., 2011). 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 the RESIT method of Peters et al. (2014). Here we employ Gaussian process regression to estimate non-linear effects and HSIC as a measure of statistical dependence. Finally, we also compare against the CD-NOD method of Zhang et al. (2017) as well as the RECI method presented in Blöbaum et al. (2018). For the latter, we employ Gaussian process regression and note that this method assumes the presence of a causal effect, and is therefore only included in some experiments. We provide a description of each of the methods in the Supplementary material F.

We generate synthetic data from the non-linear ICA model detailed in Section 2.2. Non-stationary disturbances, $\mathbf{N}$, were randomly generated by simulating Laplace random variables with distinct variances in each segment. For the non-linear mixing function we employ a deep neural network ("mixing-DNN") with randomly generated weights such that:

$$\mathbf{X}^{(1)} = \mathbf{A}^{(1)}\mathbf{N}, \tag{9}$$

$$\mathbf{X}^{(l)} = \mathbf{A}^{(l)} f\left(\mathbf{X}^{(l-1)}\right), \tag{10}$$

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. In the special case of multivariate causal discovery, we follow Peters et al. (2014) and include edges with a probability of $\frac{2}{d-1}$, implying that the expected number of edges is d. We present experiments for $d = 6$ dimensions. Note that equation (9) follows the LiNGAM. For depths $l \geq 2$, equation (10) generates data with non-linear causal structure.

Throughout experiments we vary the following factors: the number of distinct experimental conditions (i.e., distinct segments), the number of observations per segment, n_e, as well as the depth, l, of the mixing-DNN. In the context of bivariate causal discovery we measure how frequently each method is able to correctly identify the cause variable. For multivariate causal discovery we consider the F_1 score, which serves to quantify the agreement between estimated and true DAGs.

Figure 2 shows the results for bivariate causal discovery as the number of distinct experimental conditions, $|\mathcal{E}|$, 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 $l = 5$. The top panels show the proportion of times the correct cause variable was identified across 100 independent simulations. In particular, the first top panel corresponds to linear causal dependencies. As such, all methods are able to accurately recover the true cause variable. However, 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 NonSENS algorithm is able to out-perform all alternative methods.

The bottom panels of Figure 2 shows the results when no directed acyclic causal structure is present. Here data was generated such that $\mathbf{A}^{(l)}$ was not lower-triangular. In

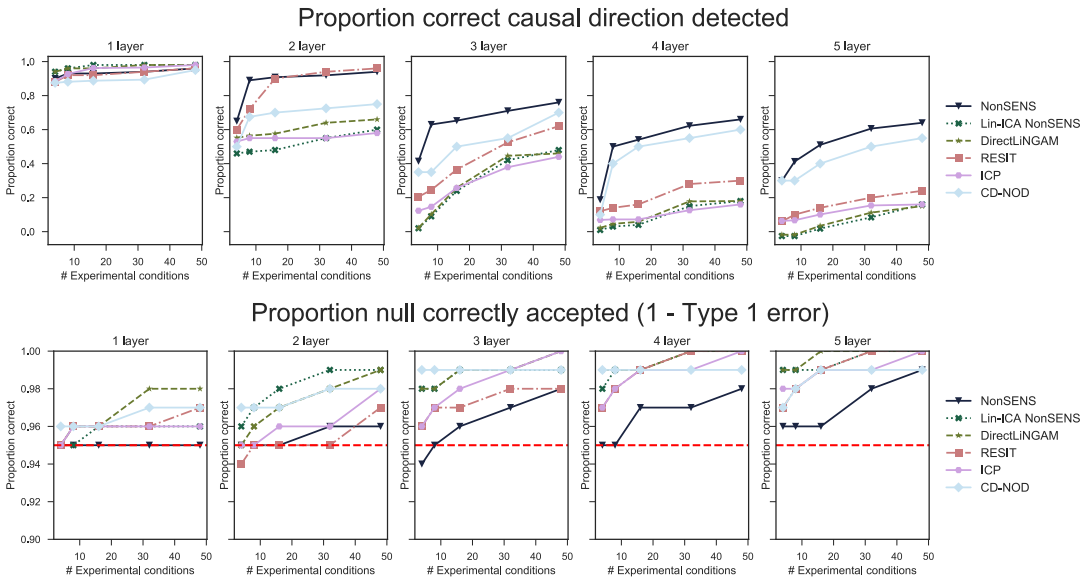


Figure 2: Experimental results indicating performance as we increase the number of experimental conditions, $|\mathcal{E}|$, whilst keeping the number of observation per condition fixed at $n_e = 512$. Each horizontal panel plots results for varying depths of the mixing-DNN, ranging from $l = 1, \dots 5$. The top panels show the proportion of times the correct cause variable is identified when a causal effect exists. The bottom panels considers data where no acyclic causal structure exists ($\mathbf{A}^{(l)}$ are not lower-triangular) and reports the proportion of times no causal effect is correctly reported. The dashed, horizontal red line indicates the theoretical $(1 - \alpha)\%$ true negative rate. For clarity we omit the standard errors, but we note that they were small in magnitude (approximately $2 - 5\%$).

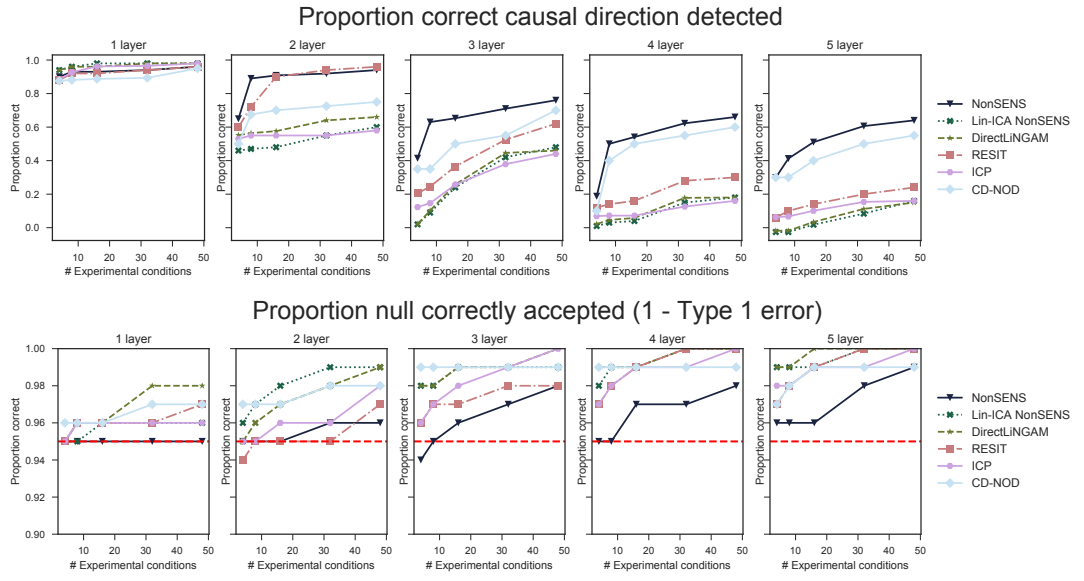


Figure 3: Experimental results visualizing performance under the assumption that a causal effect exists. This reduces the bivariate causal discovery problem to recovering the causal ordering over X_1 and X_2. The top panel considers an increasing number of experimental conditions whilst the bottom panel shows results when we vary the number of observations within a fixed number of experimental conditions, $|\mathcal{E}| = 10$. Each horizontal plane plots results for varying depths of the mixing-DNN, ranging from $l = 1, \dots, 5$.

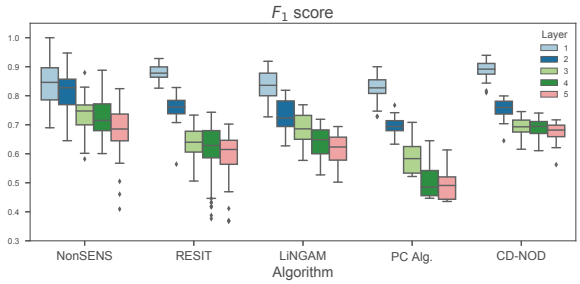


Figure 4: F_1 score for multivariate causal discovery over 6-dimensional data. For each algorithm, we plot the F_1 scores as we vary the depth of the mixing-DNN from $l = 1, \ldots, 5$. Higher F_1 scores indicate better performance.

particular, we set the off-diagonal entries of $\mathbf{A}^{(l)}$ to be identical and non-zero, resulting in cyclic causal structure. In the context of such data, we would expect all methods to report that the causal structure is inconclusive 95% of the time, as all tests are Bonferroni corrected at the $\alpha = 5\%$ level. The bottom panel of Figure 2 shows the proportion of times the causal structure is correctly reported as inconclusive. The results indicate that all methods are overly conservative in their testing, and become increasingly conservative as the depth, l, increases. 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, in Supplementary H.

Furthermore, we also consider the scenario where a causal effect is assumed to exist. In such a scenario, we consider both the likelihood ratio approach described in Section 3.4, termed NonSENS LR, and a heuristic approach of comparing the p-values of independence tests, termed NonSENS p-val. In the case of algorithms such as RESIT we compare p-values in order to determine direction. The results for these experiments are shown in Figure 3. The top panels show results as the number of experimental conditions, $|\mathcal{E}|$, increases. As before, we fix the number of observations per condition to $n_e = 512$. The bottom panels show results for a fixed number of experimental conditions $|\mathcal{E}| = 10$, as we increase the number of observations per condition. We note that the proposed measure of causal direction is shown to outperform alternative algorithms. Performance in Figure 3 appears significantly higher than that shown in Figure 2 due to that the fact that a causal effect is known to exist; this reduces the bivariate causal discovery problem to recovering the causal ordering over X_1 and X_2. The CD-NOD algorithm cannot easily be extended to assume the existence of a causal effect and is therefore not included in these experiments.

Finally, the results for multivariate causal discovery are

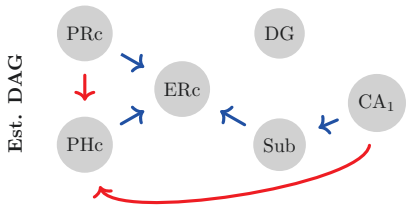


Figure 5: Estimated causal DAG on fMRI Hippocampal data by the proposed method. Blue edges are feasible given anatomical connectivity; red edges are not.

presented in Figure 4, where we plot the F_1 score between the true and inferred DAGs as the depth of the mixing-DNN increases. The proposed method is competitive across all depths. In particular, the proposed method outperforms the PC algorithm, indicating that its use to resolve undirected edges is beneficial.

4.2 HIPPOCAMPAL FMRI DATA

As a real-data application, the proposed method was applied to resting state fMRI data collected from six distinct brain regions as part of the MyConnectome project (Poldrack et al., 2015). Data was collected from a single subject over 84 successive days. Further details are provided in Supplementary Material I. We treated each day as a distinct experimental condition and employed the multivariate extension of the proposed method. For each unresolved edge, we employed NonSENS as described in Section 3.3 with a 5-layer network. The results are shown in Figure 5. While there is no ground truth available, we highlight in blue all estimated edges which are feasible due to anatomical connectivity between the regions and in red estimated edges which are not feasible (Bird and Burgess, 2008). We note that the proposed method recovers feasible directed connectivity structures for the entorhinal cortex (ERc), which is known to play an prominent role within the hippocampus.

5 CONCLUSION

We present a method to perform causal discovery in the context of general non-linear SEMs in the presence of non-stationarities or different conditions. This is in contrast to alternative methods which often require restrictions on the functional form of the SEMs. The proposed method exploits the correspondence between non-linear ICA and non-linear SEMs, as originally considered in the linear setting by Shimizu et al. (2006). Notably, we established the identifiability of causal direction from a completely different angle, by making use of non-stationarity instead of constraining functional classes. Developing computationally more efficient methods for the multivariate case is one line of our future work.

References

Anthony Bell and Terrence Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.*, 7(6):1129–1159, 1995.

Chris M. Bird and Neil Burgess. The hippocampus and memory: Insights from spatial processing. *Nat. Rev. Neurosci.*, 9(3):182–194, 2008.

Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-Effect Inference by Comparing Regression Errors. *AISTATS*, 2018.

Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. *Proc. Twenty-First Conf. Uncertain. Artif. Intell.*, pages 178–184, 2005.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. *Int. Conf. Algorithmic Learn. Theory*, pages 63–77, 2005.

Patrik O Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Neural Inf. Process. Syst.*, pages 689–696, 2009.

Aapo Hyvärinen. Fast and robust fixed-point algorithm for independent component analysis. *IEEE Trans. Neural Networks Learn. Syst.*, 10(3):626–634, 1999.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–708, 2005.

Aapo Hyvärinen. Some extensions of score matching. *Comput. Stat. Data Anal.*, 51(5):2499–2512, 2007.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. *Neural Inf. Process. Syst.*, 2016.

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

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

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6):16, 2004.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. *Neural Inf. Process. Syst.*, 2017.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters, J Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, 15:2009–2053, 2014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. Ser. B*, pages 947–1012, 2016.

Dinh Tuan Pham and Jean-Francois Cardoso. Blind Separation of Instantaneous Mixtures of Non Stationary Sources. *IEEE Trans. Signal Process.*, 49(9):1837–1848, 2001.

Russell A Poldrack et al. Long-term neural and physiological phenotyping of a single human. *Nat. Commun.*, 6, 2015. ISSN 20411723.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On Causal and Anticausal Learning. In *Int. Conf. Mach. Learn.*, pages 1255–1262, 2012.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *J. Mach. Learn. Res.*, 7:2003–2030, 2006.

Shohei Shimizu et al. DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *J. Mach. Learn. Res.*, 12:1225–1248, 2011.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Appl. Informatics*, 2016.

Peter Spirtes, Clark Glymour, Richard Scheines, David Heckerman, Christopher Meek, and Thomas Richardson. *Causation, Prediction and Search*. MIT Press, 2000.

David Van Essen et al. The Human Connectome Project: A data acquisition perspective. *NeuroImage*, 62(4): 2222–2231, 2012.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. *Proc. Twenty-Fifth Conf. Uncertain. Artif. Intell.*, pages 647–655, 2009.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. *Proc. 30th Int. Conf. Mach. Learn.*, 28:819–827, 2013. ISSN 1938-7228.

Kun Zhang, Biwei Huangy, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from Non-stationary/heterogeneous data: Skeleton estimation and orientation determination. In *Int. Jt. Conf. Artif. Intell.*, pages 1347–1353, 2017.