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 or latent disturbances. 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 both necessary and 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 discovery 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). As a result, a wide range of methods have been proposed (Shimizu et al., 2006; Pearl, 2009; 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 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 approach exploits a correspondence between a non-linear ICA model and non-linear causal models, and is specif-

ically tailored for observational data which is 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 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 (Pearl, 2009).

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

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

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:

$$\mathcal{S}_j : \quad X_j = f_j(\mathbf{PA}_j) + N_j, \quad j = 1, \ldots, d,$$

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, then 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 is 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

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

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:

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

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 $\mathbf{h}(\mathbf{X}(i); \theta)$ 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 $\mathbf{h}(X(i); \theta)$ be the same as $\mathbf{X}(i)$.*

3. *The matrix $\mathbf{L}$ with elements $\mathbf{L}_{e,i} = \lambda_i(e) - \lambda_i(1)$ for $e = 1, \ldots, E$ and $i = 1, \ldots, d$ 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, 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 condition 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 **SEM 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 casual 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 and denote by $C_i \in \mathcal{E}$ the experimental condition under which the ith observation was generated. 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. Without loss of generality, we explain the basic logic. 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 distribution is 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 piece-wise
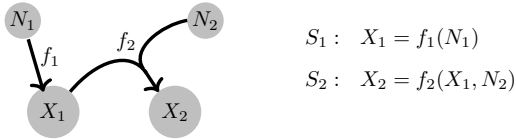
$$S_1 : \quad X_1 = f_1(N_1)$$
$$S_2 : \quad X_2 = f_2(X_1, N_2)$$

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

stationary non-linear ICA model of equations (2) and (3) will follow a (possibly non-linear) SEM where each disturbance variance, N_j, corresponds to a latent source, $S_{\pi(j)}$, and each structural equation, f_j, will correspond to an entry in the smooth, invertible funciton $\mathbf{f}$. 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 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. 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

---

[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.

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 B, 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 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.

2. Perform the four tests listed in Property 1, and conclude a causal effect 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 the cause variable.

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). 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 $\mathbf{X}$ 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 C 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: $X_1 \rightarrow X_2$ or $X_2 \rightarrow X_1$. The log-likelihood ratio is then defined as the difference in log-likelihoods under each model:

$$R = L_{1\rightarrow 2} - L_{2\rightarrow 1} \qquad (6)$$

where we write $L_{1\rightarrow 2}$ to denote the log-likelihood under the assumption that X_1 is the causal variable and $L_{2\rightarrow 1}$ for the alternative model. Under the assumption that $X_1 \rightarrow X_2$, 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

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

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:

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

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

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

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, $L_{1\rightarrow 2}$, 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 $\mathbf{g}_2 : \mathbb{R}^2 \rightarrow \mathbb{R}$ 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:

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

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 $X_1 \rightarrow X_2$.

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:

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

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

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

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

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

where $I(\cdot, \cdot)$ denotes the mutual information between two variables. We provide a full derivation in Sup-

plementary D. 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, independence testing would require d^2 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).

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 to the RESIT algorithm 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. For example, 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 tai-

lored 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.

In contrast, 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 very 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 E.

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 randomly include edges with a probability of $2/d-1$, implying that the expected number of edges is d. We present experiments for $d = 6$ dimensional data. Note that equation (9) is a 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, $|\mathcal{E}|$), 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 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
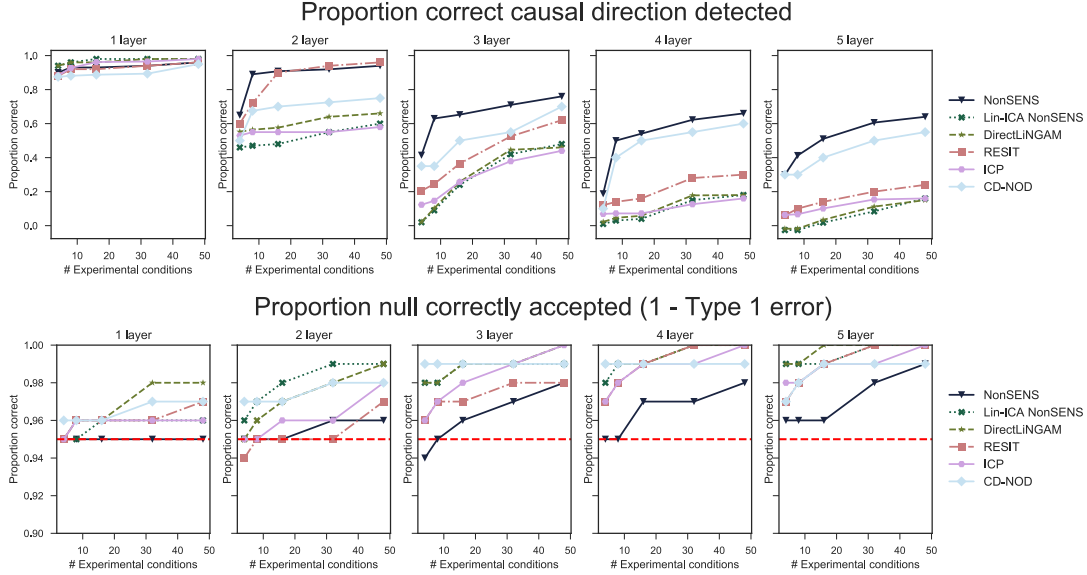
Figure 2: Experimental results indicating performance as we increase the number of experimental conditions, $|\mathcal{E}|$, whilst keeping the number of observation per condition fixed at $n_e = 512$. Each horizontal panel plots results for varying depths of the mixing-DNN, ranging from $l = 1, \ldots 5$. 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 $2 - 5\%$).
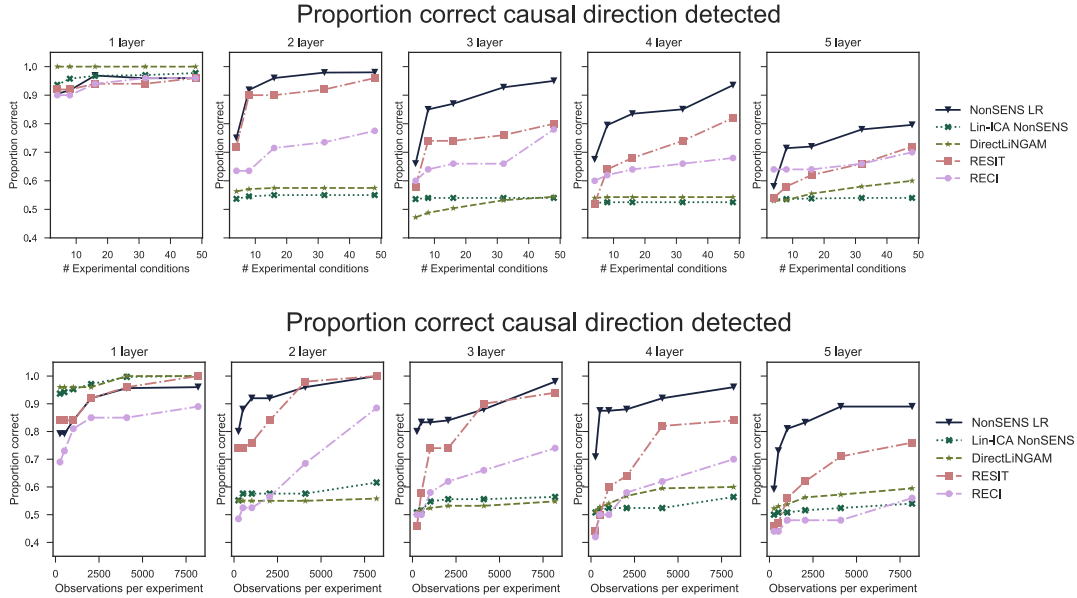

Figure 3: Experimental results visualizing performance under the assumption that a causal effect exists. This reduces the bivariate causal discovery problem to recovering the causal ordering over X_1 and X_2. The top panel considers an increasing number of experimental conditions whilst the bottom panel 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 $l = 1, \ldots, 5$.
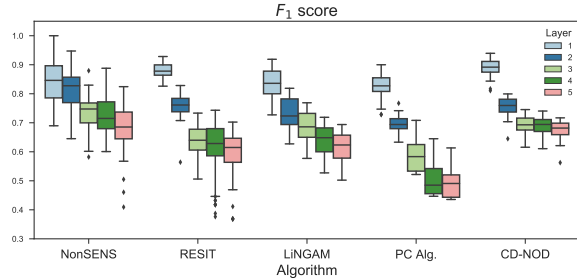
Figure 4: F_1 score for multivariate causal discovery over 6-dimensional data. For each algorithm, we plot the F_1 scores as we vary the depth of the mixing-DNN from $l = 1, \ldots, 5$. Higher F_1 scores indicate better performance.

particular, we set the off-diagonal entries of $\mathbf{A}^{(l)}$ to be equal 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 show 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. These results, presented in Supplementary F, demonstrate that the proposed method continues to perform competitively in such a scenario.

Furthermore, we also consider the scenario where a causal effect is assumed to exist. In such a scenario, we employ the likelihood ratio approach described in Section 3.4. 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 assumed 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.

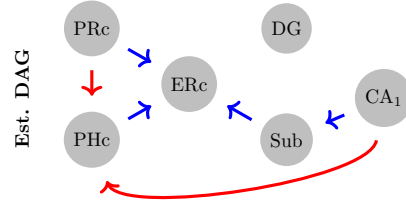Finally, the results for multivariate causal discovery are

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 out-performs 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 G. 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 to 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. ISSN 1471003X. doi: 10.1038/nrn2335.

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.

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

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 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 (Statistical Methodol.*, 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, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. 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.

D. C. Van Essen et al. The Human Connectome Project: A data acquisition perspective. *Neuroimage*, 62(4):2222–2231, oct 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.

Supplementary Material

A   Review: Non-linear ICA via time contrastive learning

Hwew we provide a more detailed review of the theory of TCL. For further details we refer readers to Hyvärinen and Morioka (2016) and we note that the results presented in this section are adapted from Hyvärinen and Morioka (2016). We begin by showing that an optimally discriminative feature extractor combined with a linear multinomial classification layer learns to model the non-stationary probability density of observations within each experimental condition.

Recall that we observe d-dimensional data, $\mathbf{X}(i) = \mathbf{f}(\mathbf{S}(i))$, which is generated via a smooth and invertible non-linear mixture, $\mathbf{f}$, of d independent latent variables, $\mathbf{S}(i)$. As in linear ICA, the latent variables are assumed to be mutually independent. However, we also assume they are non-stationary. In particular, we assume the distribution of latent variables to be piece-wises stationary such that we may associate a label, $C_i \in \mathcal{E}$, with each $\mathbf{S}(i)$ indicating the piece-wise stationary segment from which it was generated. In this manner, it is assumed that the distribution of latent variables varies across segments, as shown in Figure S.1. As such, we write $C_i \in \mathcal{E}$ to denote the segment of the ith observation where $\mathcal{E} = \{1, \ldots, E\}$ is the set of all distinct segments. For example, each segment may correspond to a distinct experimental condition. As the function $\mathbf{f}$ is smooth and invertible, it follows that the distribution of each $\mathbf{X}(i)$ will also vary across segments.

We may therefore consider the task of classifying observed data into the various segments as a multinomial classification task consisting of features, $\mathbf{X}(i)$, and categorical labels, C_i. For any observation, $\mathbf{X}(i)$, associated with true label $C_i \in \mathcal{E}$, we have:

$$p(C_i = \tau | \mathbf{X}(i), \theta, \mathbf{W}, \mathbf{b}) = \frac{\exp\left(\mathbf{w}_\tau^T \, \mathbf{h}(\mathbf{X}(i); \theta) \, + b_\tau\right)}{1 + \sum_{e=2}^{E} \exp\left(\mathbf{w}_e^T \, \mathbf{h}(\mathbf{X}(i); \theta) \, + b_e\right)}, \tag{S.1}$$

where θ are parameters for the neural network feature extractor and the weight matrix, $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_E]$, and bias vector, $\mathbf{b} = [b_1, \ldots, b_E]$, parameterize the final multinomial layer. We note that the sum in the denominator goes from $e = 2, \ldots, E$. This is because we fix $\mathbf{w}_1 = 0$ and $b_1 = 0$ in order to avoid indeterminacy in the softmax function.

Conversely, we can derive the true posterior distribution over the label C_i as:

$$p(C_i = \tau | \mathbf{X}(i)) = \frac{p(\mathbf{X}(i)|C_i = \tau) \, p(C_i = \tau)}{\sum_{e=1}^{E} p(\mathbf{X}(i)|C_i = e) \, p(C_i = e)}. \tag{S.2}$$

If we assume that the feature extractor has a universal function approximation capacity and that we have infinite data then the multinomial logistic classifier based on features $\mathbf{h}(\mathbf{X}; \theta)$ will converge to the optimal classifier, implying that equation (S.1) will equal equation (S.2) for all $\tau \in \mathcal{E}$. We may then consider the following ratio:

$$\frac{p(C_i = \tau | \mathbf{X}(i), \theta, \mathbf{W}, b)}{p(C_i = 1 | \mathbf{X}(i), \theta, \mathbf{W}, b)} = \frac{p(C_i = \tau | \mathbf{X}(i))}{p(C_i = 1 | \mathbf{X}(i))}, \tag{S.3}$$

which after expanding and taking logarithms yields:

$$\mathbf{w}_\tau^T \, \mathbf{h}(\mathbf{X}(i); \theta) + b_\tau = \log \, p(\mathbf{X}(i)|C_i = \tau) - \log \, p(\mathbf{X}(i)|C_i = 1) + \log \frac{p(C_i = \tau)}{p(C_i = 1)}, \tag{S.4}$$

indicating that the optimal feature extractor computes the log probability density function of the data within each experimental condition (relative to some pivot segment, in this case the first condition). We note that this condition holds for all $\tau \in \mathcal{E}$.

If we further assume the data were generated according to a non-stationary ICA model as described in equation (3), then equation (S.4) yields:

$$W_\tau^T \, \mathbf{h}(\mathbf{X}; \theta) - k_1(\mathbf{X}(i)) = \sum_{j=1}^{d} \lambda_j(\tau) \, q(S_j) - k_2(\tau), \tag{S.5}$$

where the sum is taken over each independent source, $S_1, \ldots, S_d$. Equation (S.5) follows from the change of variable from $\mathbf{S}$ to $\mathbf{X}$, noting that the Jacobians required by such a transformation cancel out because of the subtraction in the

right hand side of equation (S.4). As a result, by modeling the log probability densities with respect to some pivot segment (in this case segment 1), we do not need explicity compute the Jacobians. We note that k_1 is a function which does not depend on τ and k_2 is a function which does not depend on either $\mathbf{X}$ or $\mathbf{S}$. As a result, it follows that both $\mathbf{h}(\mathbf{X}; \theta)$ and $q(\mathbf{S})$ must span the same linear space, implying that we may compute latent sources up to some non-linearity, $q(\mathbf{S})$, by first learning a feature extractor based on TCL and subsequently applying linear ICA on estimated features, $\mathbf{h}(\mathbf{X}; \theta)$. Figure S.1 summarizes the relationship between the generative model and TCL.
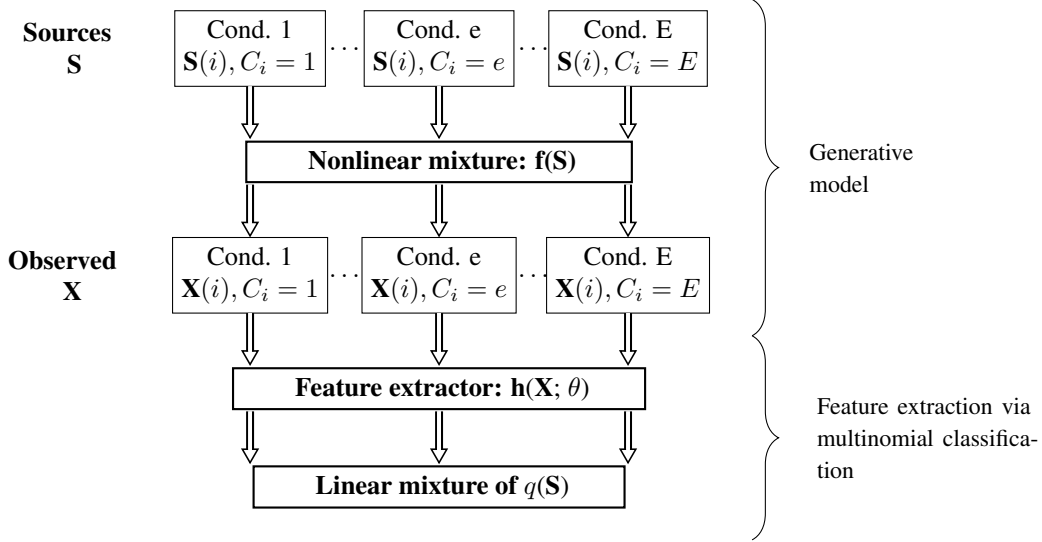

Figure S.1: A cartoon visualization which highlights the relationship between features learnt using TCL and a non-linear ICA model. The non-linear ICA model assumes observations, $\mathbf{X}$, are generated based on non-stationary latent sources whose distribution varies according the distinct experimental conditions, $e \in \mathcal{E}$.

B   A linear ICA method for piece-wise stationary sources

Formally, assumptions 1-3 of Theorem 1 guarantee that TCL, as presented in Section 2.2, will recover a linear mixture of latent independent sources up to point-wise transformation. This implies that the hidden representations obtained satisfy:

$$\mathbf{h}(\mathbf{X}; \theta) = \mathbf{A}q(\mathbf{N}), \tag{S.6}$$

for some linear mixing matrix, $\mathbf{A}$. Equation (S.6) suggests that applying ordinary linear ICA to hidden representations, $\mathbf{h}(\mathbf{X}; \theta)$, will allow us to recover $q(\mathbf{N})$. However, the use of linear ICA is premised on the assumption that latent variables are independent. This is not necessarily guaranteed under the generative model presented in equation (3) in the context of a fixed number of segments. While it is certainly true that N_1 and N_2 are conditionally independent given segment labels, it may be the case that they are not marginally independent over all segments; for example it is possible that their variances are somehow related (e.g., they may be monotonically increasing). We note that this is not an issue when the number of segments grows asymptotically and we assume exponential family parameters, $\lambda_j(e)$, are randomly generated, as stated in Corollary 1 of Hyvärinen and Morioka (2016).

Here we seek to address this issue by proposing an alternative linear ICA algorithm which explicitly models the piece-wise stationary nature of the data. In particular, the proposed linear ICA algorithm explicitly incorporates equation (3) as the generative model for latent sources. As such, it can be used to accurately unmix sources in the final stage of TCL, especially when the number of segments in small.

To set notation, we assume we observe data which corresponds to a linear mixture of sources:

$$\mathbf{Z} = \mathbf{A}\mathbf{S}$$

where $\mathbf{Z}, \mathbf{S} \in \mathbb{R}^d$ and $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a square mixing matrix. Popular ICA algorithms, such as FastICA, estimate the unmixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ by maximizing the log-likelihood under the assumptions that sources are independent and

non-Gaussian. The objective function for FastICA is therefore of the form:

$$\log p(\mathbf{Z}) = \sum_{j=1}^{d} q(\mathbf{w}_j^T \mathbf{Z}) - Z(W), \tag{S.7}$$

where $Z(W) = \log |\det W|$ is the normalization constant and we write $\mathbf{w}_k$ to denote the kth row of $\mathbf{W}$. Typically a parametric form is assumed for $q(\cdot)$, popular examples being exponential or log-cosh. In the context of the generative model for sources specified in equation (3), the FastICA algorithm therefore proceeds under the assumption that $\lambda_j(e)$ is constant across all segments $e \in \mathcal{E}$.

In order to address this issue we consider an alternative model for the density of latent sources. In particular, we seek to directly employ the piece-wise stationary log-density described in equation (3). As such, we log-density of an observation within segment $e \in \mathcal{E}$ is defined as:

$$\log p_e(\mathbf{Z}) = \sum_{j=1}^{d} \lambda_j(e) q_j(\mathbf{w}_j^T \mathbf{Z}) - Z(W, \lambda(e)). \tag{S.8}$$

In contrast to equation (S.7), the log-density of each observation depends on both the segment, e, the exponential family parameters, $\boldsymbol{\lambda} = \{\lambda_j(e) : e \in \mathcal{E}, j = 1, \ldots, n\}$, as well as the unmixing matrix, $\mathbf{W}$.

In other to recover latent sources we propose to estimate parameters, corresponding to unmixing matrix as well as exponential family parameters, via score matching (Hyvärinen, 2005). This avoids the need to estimate the normalization parameter, which may not be available analytically when sources follow unnormalized distributions. The score matching objective for the ICA model defiend in equation (S.8) is defined as:

$$\begin{aligned}
\tilde{J} = &\sum_{e \in \mathcal{E}} \sum_{j=1}^{d} \lambda_j(e) \frac{1}{n_e} \sum_{i, C_i = e} q_j''(\mathbf{w}_j^T \mathbf{Z}(i)) \\
&+ \frac{1}{2} \sum_{e \in \mathcal{E}} \sum_{j,k=1}^{d} \lambda_k(e) \lambda_j(e) \mathbf{w}_k^T \mathbf{w}_j \frac{1}{n_e} \sum_{i, C_i = e} q_k'(\mathbf{w}_k^T \mathbf{Z}(i)) q_j'(\mathbf{w}_j^T \mathbf{Z}(i)),
\end{aligned} \tag{S.9}$$

where we write q_j' and q_j'' to denote the first and second derivatives of q_j with respect to observations, $\mathbf{Z}$. We propose to minimize equation (S.9) via block gradient descent, conditionally updating the mixing matrix $\mathbf{W}$ and exponential family parameters $\boldsymbol{\lambda}$. This has the important benefit that conditional on $\mathbf{W}$, there is a closed form update for $\boldsymbol{\lambda}$ (Hyvärinen, 2007).

Experimental results

In order to assess the performance of the proposed linear ICA algorithm, we generated bivariate data following the piece-wise stationary distribution described in equation (3). We compare the performance of the proposal algorithm against the following popular linear ICA algorithms: FastICA (Hyvärinen, 1999), Infomax ICA (Bell and Sejnowski, 1995) and Joint Diagonalization method proposed by Pham and Cardoso (2001) which also accommodates non-stationary sources.

In order to assess the performance of the proposed method we consider two scenarios:

- The exponential family parameters are deliberately generated such that there is a statistical dependence structure across segments. In particular, we generate bivariate data where we explicitly enforce $\lambda_j(e)$to be monotonically increasing in e. As a concrete example, when sources follow a Laplace distribution this implies that $q(S) = |S|$ and in turn $\lambda_j(e)$ corresponds to the variance of the jth source in segment e. In such a setting, we generate piece-wise stationary Laplace sources with where the variances are correlated across segments.

- As a baseline, we also generate data where the exponential family parameters are generated at random. This removes any systematic, higher-order dependence between latent sources.
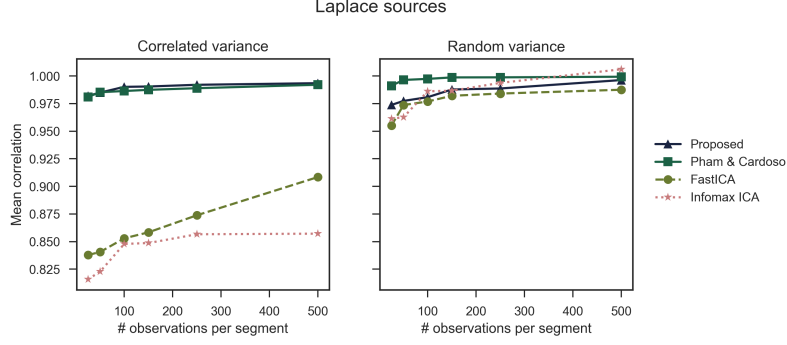
Figure S.2: Performance of various linear ICA algorithms for piece-wise stationary Laplace sources. The left panel shows performance when the variance of latent sources are positive correlated, introducing second order dependence in the data. As expected, FastICA and Infomax ICA perform poorly in this context. The right panel shows similar data were variances are no longer correlated, resulting in good performance for all algorithms.

We begin by generating bivariate data were sources follow a piece-wise stationary Laplace distribution. This implies that sources follow the log-density specified in equation (3) where $q_j(S_j) = |S_j|$ and each term $\lambda_j(e)$ denotes the variance of the jth source in segment e. Results when data is generated over five segments are provided in Figure S.2. The left panel shows the case were the variances of each latent source are correlated across segments. We note that when this is the case, methods such FastICA and Infomax ICA perform poorly. This is in contrast to the proposed method and the joint diagonalization approach of Pham and Cardoso (2001), who explicitly model the non-stationary nature of the data. The right panel of Figure S.2 shows equivalent results when variances are randomly generated, thereby removing second order dependence between latent variables. As expected, in this setting all methods perform well.
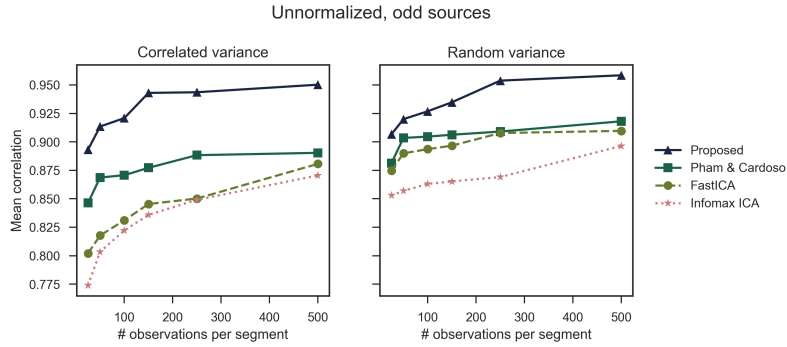

Figure S.3: Performance of various linear ICA algorithms when latent sources follow the log-density detailed in equation (S.10). The left panel shows performance when exponential family parameters, $\lambda_i(e)$, are positive correlated, introducing higher order dependence in the data. The right panel shows similar data were parameters, $\lambda_i(e)$ are randomly generated.

In order to further probe the differences between the proposed method and the approach of Pham and Cardoso (2001), we consider latent sources with an unnormalized distribution. In particular, we generate sources such that the log-density within a particular segment is as follows:

$$\log p_e(S_j) = \begin{cases} -3\lambda_j(e)|S_j| - \frac{1}{2}S_j^2, & \text{if } S_j \geq 0 \\ -\lambda_j(e)|S_j| - \frac{1}{2}S_j^2, & \text{otherwise.} \end{cases} \tag{S.10}$$

Such a density is both unnormalized but also odd. As such, we expect the joint diagonalization algorithm of Pham and Cardoso (2001) to perform poorly in this setting as it exclusively studies covariance structure and therefore cannot

model skewed distributions. Figure S.3 visualizes the results for these experiments. As expected, the Joint Diagonalization algorithm of Pham and Cardoso (2001) suffers a drop in performance. However, it continues to outperform FastICA and Infomax ICA, especially when there are dependencies over the exponential family parameters. We note that the proposed model, where parameters are estimated using score matching, is shown to be more robust.

C   Proof of Theorem 2

The proof of Theorem 2 follows from a combination of the presented assumptions together with Property 1. Formally, assumptions 1–3 guarantee that the TCL, as presented in Section 2.2, will recover a linear mixture of latent independent sources up to point-wise transformation. This, combined with the novel linear ICA algorithm described in Section 3.2, imply that latent disturbances can be recovered. We note that in practice we recover the latent disturbances up to point-wise transformation, $q(\mathbf{N})$, as opposed to $\mathbf{N}$. This is not a problem when a general test of statistical independence, such as HSIC which can capture arbitrary (i.e., non-linear) dependencies, is employed. Finally, Assumption 1 further states there is no latent confounder is present, implying that running all possible pairwise tests using a sufficiently flexible independence test, as required by assumption 4, will allow us to determine causal structure.

D   Relationship between likelihood ratio and measures of independence

In this section we derive the result presented in equation (8), for some permutation π of latent disturbances. We begin by considering the mutual information between X_1 and $N_{\pi(2)}$:

$$I(X_1, N_{\pi(2)}) = H(X_1) + H(N_{\pi(2)}) - H(X_1, N_{\pi(2)})$$

$$= H(X_1) + H(N_{\pi(2)}) - H(X_1, X_2) - \mathbb{E}\left[\log\left|\frac{\partial \mathbf{g}_{\pi(2)}}{\partial X_2}\right|\right]$$

where we employ the same change of variable, whose Jacobian can be easily evaluated, as in Section 3.4. In particular, we have used the property that:

$$H(X_1, N_{\pi(2)}) = H(X_1, X_2) + \log|\det \mathbf{J}\tilde{\mathbf{g}}|$$

and noted that the particular choice of $\tilde{\mathbf{g}}$ allows us to directly compute the Jacobian as $\frac{\partial \mathbf{g}_{\pi(2)}}{\partial X_2}$. We may therefore compute the difference in mutual information between each observed variable and the relevant latent disturbance, yielding:

$$I(X_1, N_{\pi(2)}) - I(X_2, N_{\pi(1)}) = H(X_1) + H(N_{\pi(2)}) - \mathbb{E}\left[\log\left|\frac{\partial \mathbf{g}_{\pi(2)}}{\partial X_2}\right|\right]$$

$$- H(X_2) - H(N_{\pi(1)}) + \mathbb{E}\left[\log\left|\frac{\partial \mathbf{g}_{\pi(1)}}{\partial X_1}\right|\right]$$

which is precisely the negative of likelihood ratio presented in Section 3.4.

E Baseline methods

In this section we briefly overview and provide pseudo-code for alternative methods which are presented as baselines in the manuscript.

- **DirectLiNGAM**: The DirectLiNGAM method of Shimizu et al. (2011) is based on the property that within a linear non-Gaussian acyclic model (LiNGAM), if we regress out the parents of any variable, then the residuals also follow a LiNGAM. Based on this property, the authors propose an iterative algorithm through which to iteratively uncover exogenous variables. oFurther, if variables follow a LiNGAM, then due to the additive nature of noise in such models, the residuals will be independent when we regress the parent on its children. As a result, we may infer the causal structure by studying the statistical independences between variables and residuals. Pseudo-code for the bivariate DirectLiNGAM method is provided in Algorithm 1.

Algorithm 1: Bivariate causal discovery using DirectLiNGAM (Shimizu et al., 2011)

Input : Bivariate data, $\mathbf{X}$, and significance level α.

1 **for** $i \in \{1, 2\}$ **do**
2   Linearly regress X_i on $X_{\{1,2\}\setminus i}$ and compute the residual $\hat{N}_i$.
3   Evaluate the test:

$$H_{X_i, \hat{N}_i, 0} : \mathbf{P}_{X_i, \hat{N}_j} = \mathbf{P}_{X_i} \mathbf{P}_{\hat{N}_j} \ \text{ against } \ H_{X_i, \hat{N}_j, 1} : \mathbf{P}_{X_i, N_j} \neq \mathbf{P}_{X_i} \mathbf{P}_{\hat{N}_j} \ \text{ at the } \tfrac{\alpha}{2} \text{ level.}$$

4 **end**
5 **if** *we fail to reject the null hypothesis only once* **then**
6   Variable i' such that we fail to reject $H_{X_{i'}, \hat{N}_j, 0}$ is the cause variable
7 **else**
8   The causal dependence structure is inconclusive
9 **end**

- **RESIT**: The RESIT method, first proposed by Hoyer et al. (2009) and subsequently extended by Peters et al. (2014), can be seen as a non-linear extension of the DirectLiNGAM algorithm. The RESIT algorithm is premised on the assumption of an additive noise model (ANM), which implies that each structural equation are of the form $X_j = f_j(\mathbf{PA}_j) + N_j$. Given the ANM assumption, RESIT is able to recover the causal structure by testing for dependence between variables and residuals. Gaussian process regression is employed in order to accommodate for non-linear additive causal relations.

Algorithm 2: Bivariate causal discovery using RESIT (Hoyer et al., 2009; Peters et al., 2014)

Input : Bivariate data, $\mathbf{X}$, and significance level α.

1 **for** $i \in \{1, 2\}$ **do**
2   Regress X_i on $X_{\{1,2\}\setminus i}$ and compute the residual $\hat{N}_i$.   `// Gaussian process regression is employed`
3   Evaluate the test:

$$H_{X_i, \hat{N}_i, 0} : \mathbf{P}_{X_i, \hat{N}_j} = \mathbf{P}_{X_i} \mathbf{P}_{\hat{N}_j} \ \text{ against } \ H_{X_i, \hat{N}_j, 1} : \mathbf{P}_{X_i, N_j} \neq \mathbf{P}_{X_i} \mathbf{P}_{\hat{N}_j} \ \text{ at the } \tfrac{\alpha}{2} \text{ level.}$$

4 **end**
5 **if** *We fail to reject the null hypothesis only once* **then**
6   Variable i' such that we fail to reject $H_{X_{i'}, \hat{N}_j, 0}$ is the cause variable
7 **else**
8   The causal dependence structure is inconclusive
9 **end**

- **Non-linear ICP**: The ICP method proposed by Peters et al. (2016) proposes a fundamentally different approach to causal discovery. The underlying principle of the ICP algorithm is that the direct causal predictors of a given variable must remain invariant across distribution shifts induced by various experimental conditions. In the context of bivariate data, the non-linear ICP algorithm, as described in Section 6.1 of Peters et al. (2016) therefore corresponds to fitting a non-linear regression model on the data across all experimental conditions and testing whether the distribution of residuals is the same within each condition. We note that such an approach assumes an additive noise model, as this greatly simplifies testing for invariance.

Algorithm 3: Bivariate causal discovery via non-linear ICP $\hfill$ (Peters et al., 2016)

Input : Bivariate data, $\mathbf{X}(i)$, labels $C_i \in \{1, \dots, E\}$ and significance level α.

1 **for** $i \in \{1, 2\}$ **do**

2 $\quad$ Regress X_i on $X_{\{1,2\}\backslash i}$ and compute the residual, $\hat{N}_i$
$\quad$ `// Note that Gaussian process regression is employed and we aggregate data`
$\qquad$ `across all experimental conditions`

3 $\quad$ Evaluate the test:

$$H_{\hat{N}_i,0} : \mathbf{P}_{\hat{N}_i,e} = \mathbf{P}_N \text{ for all } e \in \mathcal{E} \ \text{ against } \ H_{\hat{N}_i,1} : \mathbf{P}_{\hat{N}_i,e} \neq \mathbf{P}_N \text{ for some } e \in \mathcal{E}$$

$\quad$ where $\mathbf{P}_N$ is some arbitrary distribution. `// The Kolmogorov-Smirnov test is employed`

4 **end**

5 **if** *We fail to reject the null hyp. only once* **then**

6 $\quad$ Variable i such that we fail to reject $H_{\hat{N}_i,0}$ is the cause variable

7 **else**

8 $\quad$ The causal dependence structure is inconclusive

9 **end**

- **RECI**: Blöbaum et al. (2018) propose a method for inferring the causal relation between two variables by comparing the regression errors in each possible causal direction. Under some mild assumptions, they are able to prove that the magnitude of residual errors should be smaller in the causal direction. This suggests a straightforward causal discovery algorithm, which we outline below. We note that while any non-linear regression method may be employed, our implementation used Gaussian process regression.

Algorithm 4: Bivariate causal discovery using RECI $\hfill$ (Blöbaum et al., 2018)

Input : Bivariate data, $\mathbf{X}$.

1 Standardize data such that $\mathbf{X}_t$ is zero mean and unit variance.

2 **for** $i \in \{1, 2\}$ **do**

3 $\quad$ Regress X_i on $X_{\{1,2\}\backslash i}$ and evaluate the mean-squared error, MSE_i `// Gaussian process`
$\qquad$ `regression is employed`

4 **end**

5 **if** $MSE_1 < MSE_2$ **then**

6 $\quad$ Variable 1 is the cause variable

7 **else**

8 $\quad$ Variable 2 is the cause variable

9 **end**

- **CD-NOD**: Zhang et al. (2017) propose a causal discovery algorithm which explicitly accounts for non-stationarity or heterogeneity over observed variables. The CD-NOD algorithm accounts for non-stationarity, which may manifest itself as changes in the causal modules, by introducing an additional 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. Their method can find causal direction by making use of not only invariance, but also independent changes of causal models, as an extended notion of invariance. We also note that CD-NOD is a non-parametric method, implying it is able to accommodate non-linear causal dependencies.

Algorithm 5: Bivariate causal discovery using CD-NOD	(Zhang et al., 2017)

Input : Bivariate data, **X**.

1 Build an augmented dataset consisting of **X** and C, the observed variable representing time or domain index.
   `// Detection of changing modules`
2 **for** $i \in \{1, 2\}$ **do**
3 | Test for marginal and conditional dependence between X_i and C
4 | If they are conditionally independent given $X_{\{1,2\}\backslash i}$ then we remove the edge between X_i and C in the augmented DAG
5 **end**
   `// Recover causal skeleton`
6 **if** $X_1 \perp\!\!\!\perp X_2 \mid C$ **then**
7 | Remove the edge between X_1 and $X_2 \Rightarrow$ no causal relation between X_1 and X_2
8 **else**
9 | **if** *Only one of X_1 and X_2 is marginally or conditionally dependent on C* **then**
10 | | Dependent variable is reported as the cause
11 | **else**
12 | | Determine cause variable by comparing mutual information
13 | **end**
14 **end**

F   Further experimental results

In this section of the supplementary material we present further experimental results. In particular, in Section F.1 we present results for bivariate causal discovery in the context of a fixed number of experimental conditions, $|\mathcal{E}| = 10$, and increasing observations per segment. In Section F.2 we provide addition results in the context of multivariate causal discovery. In particular, we report the Hamming distance between true and estimated DAGs.

F.1   Additional bivariate causal discovery experiments

We 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. The results are presented in Figure S.4, where we repeat each experiment 100 times. We note that all algorithms are able to accurately identify causal structure in the presence of LiNGAMs (corresponding to a 1 layer mixing-DNN). However, as the causal structure becomes increasingly non-linear, the performance of all methods declines. In particular, we note that the proposed method has comparable performance with alternative methods such as RESIT and CD-NOD when the number of samples is small. However, as the number of observations increases the proposed method is able to out-perform alternative algorithms.

F.2   Multivariate causal discovery results

In this section we provide additional performance metrics in the context of multivariate causal discovery. While Figure 4 reported the F_1 score, we further provide results for the Hamming distance between true and estimated DAGs in Figure S.5.
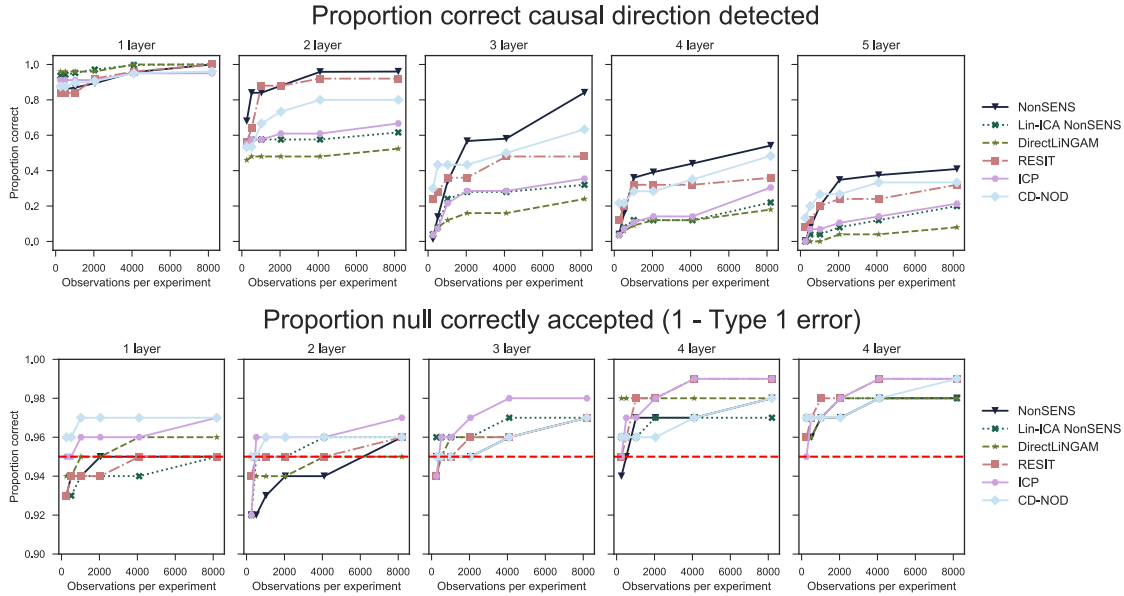
.

Figure S.4: 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 horizontal plane plots results for varying depths of the mixing-DNN, ranging from $l = 1, \ldots, 5$. The top panel plots the proportion of times the correct cause variable is identified when a causal effect exists. The bottom panel considers data where no acyclic causal structure exists ($\mathbf{A}^{(l)}$ is 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 $2 - 5\%$).
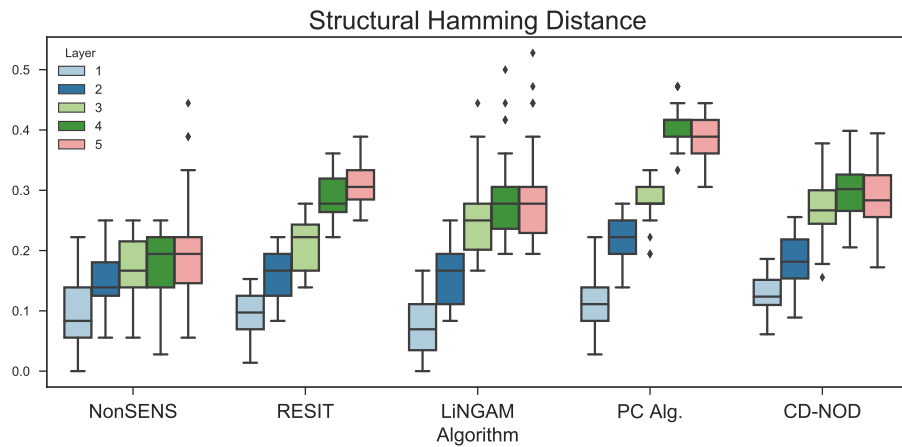

Figure S.5: Hamming distance results for multivariate causal discovery with 6-dimensional data. For each algorithm, we plot the structural Hamming distance as we vary the depth of the mixing-DNN from $l = 1, \ldots, 5$. Lower scores indicate better performance.

.

G   Hippocampal functional MRI data

In this section we provide further details of the Hippocampal fMRI data employed in Section 4.2. The data was collected as part of the MyConnectome project, presented in Poldrack et al. (2015), which involved daily fMRI scans for a single individual (Caucasian male, aged 45). The data may be freely downloaded from `https://openneuro.org/datasets/ds000031/`.

We focus only on the resting-state fMRI data taken from this project, noting that future work may also wish to study the other modalities of data collected.

Data was collected from the same subject over a series of 84 successive days, allowing us to consider data collected on distinct days as a distinct experimental condition. Full details of the data acquisition pipelines are provided in Poldrack et al. (2015). For each day, we observe 518 BOLD measurements. After preprocessing, data was collected from the following brain regions: perirhinal cortex (PRC), parahippocampal cortex (PHC), entorhinal cortex (ERC), subiculum (Sub), CA1, and CA3/Dentate Gyrus (DG). This resulted in $d = 6$ dimensional data. As such, data employed consists of $n_e = 518$ observations per experimental condition and $|\mathcal{E}| = 84$ distinct conditions.