Causal Autoregressive Flows

Ilyes Khemakhem* Gatsby Unit, UCL
Ricardo P. Monti* Gatsby Unit, UCL
Robert Leech King's College London
Aapo Hyvärinen University of Helsinki

Key ideas of Flows: build complex distributions from simple distributions via a flow of successive (invertible) transformation

Abstract

Two apparently unrelated fields — normalizing flows and causality — have recently received considerable attention in the machine learning community. In this work, we highlight an intrinsic correspondence between a simple family of autoregressive normalizing flows and identifiable causal models. We exploit the fact that autoregressive flow architectures define an ordering over variables, analogous to a causal ordering, to show that they are well-suited to performing a range of causal inference tasks, ranging from causal discovery to making interventional and counterfactual predictions. First, we show that causal models derived from both affine and additive autoregressive flows with fixed orderings over variables are identifiable, i.e. the true direction of causal influence can be recovered. This provides a generalization of the additive noise model well-known in causal discovery. Second, we derive a bivariate measure of causal direction based on likelihood ratios, leveraging the fact that flow models can estimate normalized log-densities of data. Third, we demonstrate that flows naturally allow for direct evaluation of both interventional and counterfactual queries, the latter case being possible due to the invertible nature of flows. Finally, throughout a series of experiments on synthetic and real data, the proposed method is shown to outperform current approaches for causal discovery as well as making accurate interventional and counterfactual predictions.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.
*Equal contribution.

1 INTRODUCTION

Causal models play a fundamental role in modern scientific endeavour (Spirtes et al., 2000; Pearl, 2009b). Many of the questions which drive scientific research are not associational but rather causal in nature. While randomized controlled studies are the gold standard for understanding the underlying causal mechanisms of a system, such experiments are often unethical, too expensive, or technically impossible. In the absence of randomized controlled trials, the framework of *structural equation models* (SEMs) can be used to encapsulate causal knowledge as well as to answer interventional and counterfactual queries (Pearl, 2009a). At a fundamental level, SEMs define a generative model for data based on causal relationships, and contain strictly more information than their corresponding causal graph and law.

The first step in performing causal inference is to determine the underlying causal graph. Whilst this can be achieved in several ways (e.g., randomized study, expert judgement), data driven approaches using purely observational data, termed *causal discovery*, are often employed. The challenge for causal discovery algorithms is that given a (typically empirical) data distribution one can write many different SEMs that could generate such distribution (Zhang et al., 2015a; Spirtes and Zhang, 2016). In other words, the causal structure is unidentifiable in the absence of any constraints.

Causal discovery algorithms typically take one of two approaches to achieve identifiability. The first approach is to introduce constraints over the family of functions present in the SEM, for example assuming all causal dependencies are linear or that disturbances are additive (Shimizu et al., 2006, 2011; Hoyer et al., 2009; Peters et al., 2014; Bloebaum et al., 2018; Zheng et al., 2018). While such approaches have been subsequently extended to allow for bijective transformations (Zhang and Hyvarinen, 2009), they often introduce unverifiable assumptions over the true underlying SEM. An alternative approach is to consider unconstrained causal models whilst introducing further assumptions over the data distribution. These methods often introduce non-

stationarity constraints on the distribution of latent variables (Peters et al., 2016; Monti et al., 2019) or assume exogeneous variables are present (Zhang et al., 2017).

In the present contribution, we consider the first approach, i.e. constraining the functions defining the causal relationships, and combine it with the framework of *normalizing flows* recently developed in deep learning literature.

Normalizing flows (Papamakarios et al., 2019; Kobyzev et al., 2020) provide a general way of constructing flexible generative models with tractable distributions, where both sampling and density estimation are efficient and exact. Flows model the data as an invertible transformation of some noise variable, whose distribution is often chosen to be simple, and make use of the change of variable formula in order to express the data density. This formula requires the evaluation of the Jacobian determinant of the transformation.

Autoregressive normalizing flows (Kingma et al., 2016; Papamakarios et al., 2018; Huang et al., 2018) purposefully yield a triangular Jacobian matrix, and the Jacobian determinant can be computed in linear time. Importantly for our purposes, the autoregressive structure in such flows is specified by an ordering on the input variables, and each output variable is only a function of the input variables that precede it in the ordering. Different architectures for autoregressive flows have been proposed, ranging from simple additive and affine transformations (Dinh et al., 2014, 2016), to more complex cubic and neural spline transformations (Durkan et al., 2019a,b). Flows have been increasingly popular, with applications in density estimation Dinh et al. (2016); Papamakarios et al. (2018), variational inference (Rezende and Mohamed, 2015; Kingma et al., 2016) and image generation (Kingma and Dhariwal, 2018; Durkan et al., 2019b), to name a few. Active research is conducted in order to increase the expressivity and flexibility of flow models, while maintaining the invertibility and sampling efficiency.

In this work, we consider the ordering of variables in an autoregressive flow model from a causal perspective, and highlight the similarities between SEMs and autoregressive flows. We show that under some constraints, autoregressive flow models are well suited to performing a variety of causal inference tasks. As a first contribution, we focus on the class of affine normalizing flows, and show that it defines an identifiable causal model. This causal model is a new generalization of the well-known additive noise model, and the proof of its identifiability constitutes the main theoretical result of this manuscript. We then leverage the properties of flows to perform causal discovery and inference in such a models. First, we use the fact that flows can efficiently evaluate exact likelihoods to propose a nonlinear measure of causal direction based on likelihood ratios, with ensuing optimality properties. Second, we show that when autoregressive flow models are conditioned upon the correct causal ordering, they can be employed to accurately answer interventional and counterfactual queries. Finally, we show that our method performs favourably on a range of experiments, both on synthetic and real data, when compared to previous methods.

2 PRELIMINARIES

2.1 Structural Equation Models

Suppose we observe d-dimensional random variables $\mathbf{x} = (x_1, \dots, x_d)$ with joint distribution $\mathbb{P}_{\mathbf{x}}$. A structural equation model (SEM) is here defined as a tuple $\mathcal{S} = (\mathbf{S}, \mathbb{P}_{\mathbf{n}})$ of a collection $\mathbf{S}$ of d structural equations:

$$S_j : \quad x_j = f_j(\mathbf{pa}_j, n_j), \quad j = 1, \dots, d \tag{1}$$

together with a joint distribution, $\mathbb{P}_{\mathbf{n}}$, over latent 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 SEM defines the *observational* distribution of the random vector $\mathbf{x}$: sampling from $\mathbb{P}_{\mathbf{x}}$ is equivalent to sampling from $\mathbb{P}_{\mathbf{n}}$ and propagating the samples through $\mathbf{S}$. The causal graph $\mathcal{G}$, associated with an SEM (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).

It is well known that for a DAG, there exists a causal ordering (or permutation) π of the nodes, such that $\pi(i) < \pi(j)$ if the variable x_i precedes the variable x_j in the DAG (but such an ordering is not necessarily unique). Thus, given the causal ordering of the associated DAG we may re-write equation (1) as

$$x_j = f_j\left(\mathbf{x}_{<\pi(j)}, n_j\right), \quad j = 1, \dots, d \tag{2}$$

where $\mathbf{x}_{<\pi(j)} = \{x_i : \pi(i) < \pi(j)\}$ denotes all variables before x_j in the causal ordering. Moreover, in the above definition of SEMs we allow f_j to be any (possibly nonlinear) function. Zhang et al. (2015b) proved that the causal direction of the general SEM (1) is not identifiable without constraints. To this end, the causal discovery community has focused on specific special cases in order to obtain identifiability results as well as provide practical algorithms. In particular, the additive noise model (Hoyer et al., 2009, ANM), which assumes the noise is additive, is of interest to us in the rest of this manuscript, and its SEM has the form

$$x_j = f_j(\mathbf{x}_{<\pi(j)}) + n_j, \quad j = 1, \dots, d \tag{3}$$

2.2 Autoregressive Normalizing Flows

Normalizing flow models seek to express the log-density of observations $\mathbf{x} \in \mathbb{R}^d$ as an invertible and differentiable transformation $\mathbf{T}$ of latent variables, $\mathbf{z} \in \mathbb{R}^d$, which follow a simple (typically factorial) base distribution that has density $p_{\mathbf{z}}(\mathbf{z})$. This allows for the density of $\mathbf{x}$ to be obtained via a change of variables as follows:

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{z}}(\mathbf{T}^{-1}(\mathbf{x}))\,|\det J_{\mathbf{T}^{-1}}(\mathbf{x})| \tag{3}$$

Typically, $\mathbf{T}$ or $\mathbf{T}^{-1}$ will be implemented with neural networks. Very often, normalizing flow models are obtained by chaining together different transformations $\mathbf{T}_1, \dots, \mathbf{T}_k$ from the same family to obtain $\mathbf{T} = \mathbf{T}_1 \circ \cdots \circ \mathbf{T}_k$, while remaining invertible and differentiable. The Jacobian determinant of $\mathbf{T}$ can simply be computed from the Jacobian determinants of the sub-transformations $\mathbf{T}_l$. As such, an important consideration is ensuring the Jacobian determinant of each of the sub-transformations to be efficiently calculated.

Autoregressive flows use transformations that are designed precisely to enable simple Jacobian computation by restricting their Jacobian matrices to be lower triangular (Huang et al., 2018). In this case, the transformation $\mathbf{T}$ has the form:

$$x_j = \tau_j(z_j, \mathbf{x}_{<\pi(j)}) \tag{4}$$

where π is a permutation that specifies an autoregressive structure on $\mathbf{x}$ and the functions τ_j (called *transformers*) are invertible with respect to their first arguments and are parametrized by their second argument.

3 CAUSAL AUTOREGRESSIVE FLOW MODEL

The ideas presented in this manuscript highlight the similarities between equations (2) and (4). In particular, both models explicitly define an ordering over variables and both models assume the latent variables (denoted by $\mathbf{n}$ or $\mathbf{z}$ respectively) follow simple, factorial distributions. Throughout the remainder of this paper, we will look to build upon these similarities in order to employ autoregressive flow models for causal inference. First, we explicit in Section 3.1 the general conditions under which such correspondence is possible. Then, we consider bivariate *affine* flows in Section 3.2, and show that they define a causal model which is identifiable, and which generalizes existing models, in particular additive noise models. In Section 3.3, we present our measure of causal direction based on the ratio of the likelihoods under two alternative flow models corresponding to different causal orderings. Finally, Section 3.4 presents an extension to the multivariate case. The causal model as well as the flow-based likelihood ratio measure of causal direction constitute the **c**ausal **a**uto**re**gressive **f**low (CAREFL) model.

3.1 From Autoregressive Flow models to SEMs

There are some constraints we need to make on how we define autoregressive normalizing flows so that they remain compatible with causal models:

(**I**) **Fixed ordering:** When chaining together different autoregressive transformations $\mathbf{T}_1, \dots, \mathbf{T}_k$ into $\mathbf{T} = \mathbf{T}_1 \circ \cdots \circ \mathbf{T}_k$, the ordering π of the input variables should be the same for all sub-transformations.

(**II**) **Affine/additive transformations:** The transformers τ_j in (4) take what is called an *affine* form:

$$\tau_j(u, \mathbf{v}) = e^{s_j(\mathbf{v})}u + t_j(\mathbf{v}) \tag{5}$$

where an *additive* transformation is a special case with $s_j = 0$.

Constraint (**I**) ensures that composing transformations maintains the autoregressive structure of the flow, so as to respect the correspondence with a SEM (2). In fact, if all sub-transformations $\mathbf{T}_l$ are autoregressive and follow the same ordering π, then $\mathbf{T}$ is also autoregressive and follows π (see Appendix C for a proof). We emphasize this point here because it is contrary to the common practice of changing the ordering π throughout the flow to make sure all input variables interact with each other (Germain et al., 2015; Dinh et al., 2016; Kingma and Dhariwal, 2018).

Constraint (**II**) ensures that the flow model is not too flexible, and in particular cannot approximate any density. In fact, the causal ordering of autoregressive flows with universal approximation capability is not identifiable. A proof can be found using the theory of non-linear ICA (Hyvärinen and Pajunen, 1999): we can autoregressively and *in any order* transform any random vector into independent components with simple distributions. In other words, for any two variables x_1 and x_2, we can construct another variable z_2 such that $z_2 \perp\!\!\!\perp x_1$. Such construction is invertible for x_2, meaning that we can write x_2 as a function of (x_1, z_2). Similarly, the same treatment can be done in the reverse order, to construct a variable z_1 that is independent of x_2, such that x_1 is a function of (x_2, z_1). That is, any two variables would be symmetric according to the SEM. This is in contradiction with the definition of identifiability of a causal model, which states that the transformation $\mathbf{T}$ from noise $\mathbf{z}$ to observed variable $\mathbf{x}$ has a unique causal ordering. Fortunately, flows based

on *additive* and *affine* transformations, as defined above (based on Dinh et al. (2016)), are not universal density approximators (see Appendix B for a proof).

Finally, note that constraints (**I**) and (**II**) only limit the expressivity of flows as universal *density* approximators. In contrast, the coefficients s_j and t_j of the affine transformer (5), when parametrized as neural networks, can be universal *function* approximators. This property of universal approximation of the functional relationships is preserved when stacking flows (see Appendix D for a proof).

3.2 Model Definition and Identifiability

Suppose we observe bivariate data $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$. Underlying the data, there is a causal ordering described by a permutation π of the set $\{1, 2\}$, where $\pi = (1, 2)$ if $x_1 \to x_2$ and $\pi = (2, 1)$ otherwise.

As per Constraints (**I**) and (**II**), let $\mathbf{T}_1, \ldots, \mathbf{T}_k$ be $k \geq 1$ *affine* autoregressive transformations—*i.e.* of the form (4) where the transformers τ_j are affine functions (5)—with ordering π, and let $\mathbf{T} = \mathbf{T}_1 \circ \cdots \circ \mathbf{T}_k$. Then $\mathbf{T}$ is also an affine transformation (see Appendix C for a proof). As mentioned earlier, such *composability* is a central and well-known property of affine flows: the ordering stays the same and the composition is still an affine flow.

The flow $\mathbf{T}$ defines the following SEM on the observations $\mathbf{x}$:

$$x_j = e^{s_j(x_{<\pi(j)})} z_j + t_j(x_{<\pi(j)}), \quad j = 1, 2 \qquad (6)$$

where z_1, z_2 are statistically independent latent noise variables, and $s_j(x_{<\pi(j)})$ and $t_j(x_{<\pi(j)})$ are defined constant (with respect to $\mathbf{x}$) for $\pi(j) = 1$. Equation (6) defines our proposed causal model where the noise is not merely added to some function of the cause (as typical in existing models), but also modulated by the cause.

As a special case, if the transformations $\mathbf{T}_l$, $l = 1, \ldots, k$ are additive (in the sense defined above), then the flow $\mathbf{T}$ is also additive, and $s_1 = s_2 = 0$. In such a special case, Equation (6) is part of the additive noise model family (3), which was proven to be identifiable by Hoyer et al. (2009).

We present next a non-technical Theorem which states that the more general affine causal model (6) is also identifiable, when the noise variable $\mathbf{z}$ is Gaussian. A more rigorous treatment as well as the proof of a more general case can be found in Appendix A.

Theorem 1 (Identifiability). *Assume* $\mathbf{x} = (x_1, x_2)$ *follows the model described by equation* (6)*, with* z_1, z_2 *statistically independent, and the function* t_j *linking cause to effect is non-linear and invertible. If* z_1 *and* z_2 *are Gaussian, the model is identifiable (i.e.,* π *is uniquely defined). Alternatively (Hoyer et al., 2009), if* $s_1 = s_2 = 0$*, the model is identifiable for any (factorial) distribution of the noise variables* z_1 *and* z_2*.*

Note that while the main result in Theorem 1 assumes Gaussian noise, we believe that the identifiability result also holds for general noise. We show that empirically in Section 5.

3.3 Choosing Causal Direction using Likelihood Ratio

Next, we use our flow-based framework to develop a concrete method for estimating the causal direction, i.e. π. We follow Hyvärinen and Smith (2013) and pose causal discovery as a statistical testing problem which we solve by likelihood ratio testing. We seek to compare two candidate models which can be seen as corresponding to two hypotheses: $x_1 \to x_2$ against $x_1 \leftarrow x_2$. Likelihood ratios are, in general, an attractive way to deciding between alternative hypotheses (models) because they have been proven to be uniformly most powerful, at least when testing "simple" hypotheses (Neyman and Pearson, 1933). However, in our special case, the framework in fact reduces to simply choosing the causal direction which has a higher likelihood.

Normalizing flows allow for easy and exact evaluation of the likelihoods. If we assume the causal ordering $\pi = (1, 2)$, then the likelihood of an affine autoregressive flow is:

$$\log L_{\pi=(1,2)}(\mathbf{x}) = \log p_{z_1}\left(e^{-s_1}(x_1 - t_1)\right) + \log p_{z_2}\left(e^{-s_2(x_1)}(x_2 - t_2(x_1))\right) - s_1 - s_2(x_1)$$

We propose to fit two affine autoregressive flow models (6), each conditioned on a distinct causal order over variables: $\pi = (1, 2)$ or $\pi = (2, 1)$. For each candidate model we train parameters for each flow via maximum likelihood. In order to avoid overfitting we look to evaluate log-likelihood for each model over a held out testing dataset. As such, the proposed measure of causal direction is defined as:

$$R = \mathbb{E}\left[\log L_{\pi=(1,2)}(\mathbf{x}_{test}; \mathbf{x}_{train})\right] - \mathbb{E}\left[\log L_{\pi=(2,1)}(\mathbf{x}_{test}; \mathbf{x}_{train})\right] \qquad (7)$$

where $\mathbb{E}\left[\log L_{\pi=(1,2)}(\mathbf{x}_{test}; \mathbf{x}_{train})\right]$ is the empirical expectation of the estimated log-likelihood evaluated on unseen test data $\mathbf{x}_{test}$. If R is positive we conclude that x_1 is the causal variable and if R is negative we conclude that x_2 is the causal variable.

3.4 Extension to Multivariate Data

We can generalize the likelihood ratio measure developed in section 3.3 to the multivariate case by computing the log-likelihood $\log L_\pi$ for each ordering π, and accept the ordering with highest log-likelihood as the true causal ordering of the data. This procedure is only feasible for small values of d, since the numbers of permutations of $[\![1, d]\!]$ grows exponentially with d. An alternative approach is to employ the bivariate likelihood ratio (7) in conjunction with a traditional constraint based method such as the PC algorithm, similarly to Zhang and Hyvarinen (2009). The PC algorithm is first used to estimate the skeleton of the DAG G that describes the causal structure of the data, and orient as many edges as possible. Then, the remaining edges are oriented using the likelihood ratio measure.

We can also extend the likelihood ratio measure in a different way: we can identify the causal direction between pairs of multivariate variables. More specifically, consider two random vectors $(\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^{2d}$, and suppose that $\mathbf{x}_1 \to \mathbf{x}_2$. Then they can be described by the following SEM:

$$\mathbf{x}_1 = e^{\mathbf{s}_1} \cdot \mathbf{z}_1 + \mathbf{t}_1$$
$$\mathbf{x}_2 = e^{\mathbf{s}_2(\mathbf{x}_1)} \cdot \mathbf{z}_2 + \mathbf{t}_2(\mathbf{x}_1)$$

where $(\mathbf{z}_1, \mathbf{z}_2)$ is the vector of latent noise variables that are supposed independent, $\mathbf{s}_i$ and $\mathbf{t}_i$ are vector-valued instead of scalar-valued, and $\cdot$ denotes the element-wise product. The likelihood ratio measure (7) can be used straightforwardly here to find the correct causal direction between $\mathbf{x}_1$ and $\mathbf{x}_2$. Note that while the identifiability theory was developed for the bivariate case, our experiments in Section 5 show that it also holds for this case of two multivariate $\mathbf{x}_i$. To the best of our knowledge, this is the first model that can readily perform causal discovery over groups of multivariate variables.

4 CAUSAL INFERENCE USING AUTOREGRESSIVE FLOWS

In this section we demonstrate how flow architectures may be employed to perform both interventional and counterfactual queries. We assume that the true causal ordering over variables has been resolved (e.g., as the result of expert judgement or obtained via the method described in Section 3). Interventional queries involve marginalization over latent variables and thus can be evaluated by propagating forward the structural equations. However, counterfactual queries require us to condition, as opposed to marginalize, over latent variables. This requires us to first infer the posterior distribution of latent variables, termed *abduction* by Pearl (2009a). In many causal inference models this is challenging, often requiring complex inference algorithms. However, the invertible nature of flows means that the posterior of latent variables given observations can be readily obtained.

4.1 Interventions

It is possible to manipulate an SEM $\mathcal{S}$ to create interventional distributions over $\mathbf{x}$. As described in Pearl (2009b), intervention on a given variable x_i defines a new *mutilated* generative model where the structural equation associated with variable x_i is replaced by the interventional value, while keeping the rest of the equations (4) fixed. Interventions are very useful in understanding causal relationships. If, under the assumption of faithfulness, intervening on a variable x_i changes the marginal distribution of another variable x_j, then it is likely that x_i has some causal effect on x_j. Conversely, if intervening on x_j doesn't change the marginal distribution of x_i, then the latter is not a descendant of x_j. We follow Pearl (2009b) and denote by $do(x_i = \alpha)$ the interventions that puts a point mass on x_i.

Autoregressive flow modelling allows us to answer interventional queries easily. After fitting a flow model (4) conditioned on the right causal ordering (assumed known) to the data, we change the structural equation for variable x_i from $x_i = \tau_i(z_i, \mathbf{x}_{<\pi(i)})$ to $x_i = \alpha$. This breaks the edges from $x_{<\pi(i)}$ to x_i, and puts a point mass on the latent variable z_i. Thereafter, we can directly draw samples from the distribution $\prod_{j \neq i} p_{z_j}$ for all remaining latent variables $z_{j \neq i}$. Finally, we obtain a sample for $\mathbf{x}^{do(x_i=\alpha)}$ by passing these samples through the flow, which allows us to compute empirical estimates of the interventional distribution. This is described in Appendix E.1.

4.2 Counterfactuals

A counterfactual query seeks to quantify statements of the form: what would the value for variable x_i have been if variable x_j had taken value α, *given that we have observed* $\mathbf{x} = \mathbf{x}^{obs}$? The fundamental difference between an interventional and counterfactual query is that the former seeks to marginalize over latent variables, whereas the latter conditions on them.

Given a set of structural equations and an observation $\mathbf{x}^{obs}$, we follow the notation of Pearl (2009b) and write $x_{i,x_j \leftarrow \alpha}(\mathbf{z})$ to denote the value of x_i under the counterfactual that $x_j \leftarrow \alpha$. As detailed by Pearl (2009b), counterfactual inference involves three steps: *abduction, action* and *prediction*. The first step involves, after fitting the flow to the data, evaluating

the posterior distribution over latent variables given observations $\mathbf{x}^{obs}$. This is non-trivial for most causal models. However, since flow models readily give access to both forward and backward transformation between observations and latent variables (Papamakarios et al., 2018; Kingma et al., 2016; Durkan et al., 2019b), this first step can be readily evaluated.

The remaining two steps mirror those taken when making interventional predictions: the structural equation for the counterfactual variable is fixed at α and the structural equations are propagated forward. The only difference here is that the latent samples are drawn from their new distribution: in fact, conditioning on $\mathbf{x} = \mathbf{x}^{obs}$ changes the distribution of the latent variables by putting a point mass on $\mathbf{z} = \mathbf{T}^{-1}(\mathbf{x}^{obs})$. This is summarized in Appendix E.2.

5 EXPERIMENTS

5.1 Causal Discovery

We compare the performance of CAREFL on a range of synthetic and real world data, against several alternative methods: the linear likelihood ratio method of Hyvärinen and Smith (2013), the additive noise model (Hoyer et al., 2009; Peters et al., 2014, ANM), and the Regression Error Causal Inference (RECI) method of Bloebaum et al. (2018). For CAREFL, we considered the more general affine flows, as well as the special case of additive flows (denoted CAREFL-NS, for "non-scaled"), where $s_j = 0$ in (6). For ANM, we considered both a Gaussian process and a neural network as the regression class. Experimental details can be found in Appendix F. Code to reproduce the experiments is available here.

5.1.1 Synthetic data

We consider a series of synthetic experiments where the underlying causal model is known. Data was generated according to the following SEM:

$$x_1 = z_1 \quad \text{and} \quad x_2 = f(x_1, z_2)$$

where z_1, z_2 follow a standard Laplace distribution. We consider three distinct forms for f: (i) linear, where $f(x_1, z_2) = \alpha x_1 + z_2$; (ii) non-linear with additive noise, where $f(x_1, z_2) = x_1 + \alpha x_1^3 + z_2$; (iii) non-linear with modulated noise, where $f(x_1, z_2) = \sigma(x_1) + \frac{1}{2}x_1^2 + \sigma(x_1)z_2$; (iv) non-linear with non-linear noise, where $f(x_1, z_2) = \sigma\left(\sigma\left(\alpha x_1\right) + z_2\right)$. We write σ to denote the sigmoid non-linearity. We also consider a high dimensional SEM:

$$\mathbf{x}_1 = \mathbf{z}_1 \in \mathbb{R}^{10} \quad \text{and} \quad \mathbf{x}_2 = \mathbf{g}(\mathbf{x}_1, \mathbf{z}_2) \in \mathbb{R}^{10}$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ follow standard Laplace distribution, and for each $i \in [\![1, 10]\!]$, g_i has one of the following forms, picked at random: (i) a function of all inputs $g_i(\mathbf{x}_1, \mathbf{z}_2) = \sigma(\sigma(\sum_j x_{1,j}) + z_i)$; (ii) a function of the first half of the input $g_i(\mathbf{x}_1, \mathbf{z}_2) = \sigma(\sigma(\sum_{j\leq 5} x_{1,j}) + z_i)$; (iii) a function of the second half of the input $g_i(\mathbf{x}_1, \mathbf{z}_2) = \sigma(\sum_{j>5} \sigma(x_{1,j})^{j-5} + z_i)$.

For each distinct class of SEMs, we consider the performance of each algorithm under various distinct sample sizes ranging from $N = 25$ to $N = 500$ samples. Furthermore, each experiment is repeated 250 times. For each repetition, the causal ordering is selected at random. We implemented CAREFL by stacking two affine flows (6), where s_j and t_j are feed-forward networks with one hidden layer of dimension 10.

Results are presented in Figure 1. Only CAREFL is able to consistently uncover the true causal direction in all situations. We note that the same architecture and training parameters were employed throughout all experiments, highlighting the fact that the proposed method is agnostic to the nature of the true underlying causal relationship.

We note that while the identifiability results of Theorem 1 are premised on Gaussian noise variables, the simulations used a Laplace distribution instead. This proves that the Gaussianity assumption is sufficient but not necessary for identifiability to hold.

Robustness to prior misspecification In the simulations above, the prior distribution of the flow was chosen to be a Laplace distribution, matching the noise distribution. To investigate CAREFL's robustness to prior mismatch, we run additional simulations where the flow prior is still a Laplace distribution, but the noise distribution is changed. The remaining of the architectural parameters are kept the same as the simulations above. The results are shown in Figure 2. We see that the performance stays the same. We also note that in the next subsection, we will consider real world datasets where we did not set the underlying (unknown) noise distribution while maintaining better performance when compared to alternative methods.

5.1.2 Real data

Cause effect pairs data We also consider performance of the proposed method on cause-effect pairs benchmark dataset (Mooij et al., 2016). This benchmark consists of 108 distinct bivariate datasets where the objective is to distinguish between cause and effect. For each dataset, two separate autoregressive flow models were trained conditional on $\pi = (1, 2)$ or $\pi = (2, 1)$ and the log-likelihood ratio was evaluated as in equation (7) to determine the causal variable. Results are pre-

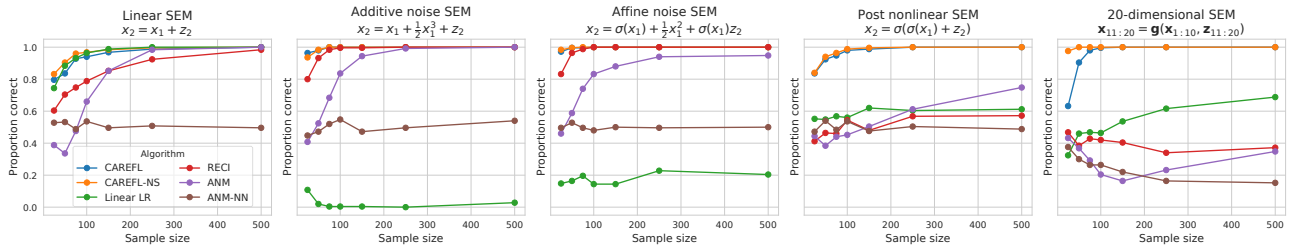

Figure 1: Performance on synthetic data generated under distinct SEMs. We note that for all five SEMs CAREFL performs competitively and is able to robustly identify the underlying causal direction.

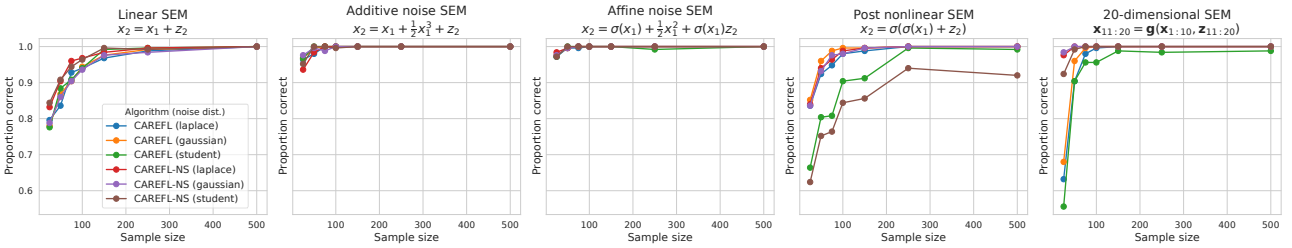

Figure 2: Impact of prior mismatch on the performace of CAREFL. The prior of each flow is fixed to a Laplace distribution, while the noise distribution is chosen to be either a Laplace, Student-t or Gaussian distribution.

Table 1: Percentage of correct causal variables identified over 108 pairs from the Cause Effect Pairs benchmark.

CAREFL	Linear LR	ANM	RECI
73 %	66%	69 %	69%

sented in Table 1. We note that the proposed method performs better than alternative algorithms.

Arrow of time on EEG data Finally, we consider the performance of CAREFL in inferring the arrow of time from open-access electroencephalogram (EEG) time series (Dornhege et al., 2004). The data consists of 118 EEG channels for one subject. We only consider the first n time points, where $n \in \{150, 500\}$, after which each of the channels is randomly reversed. More details on the preprocessing can be found in Appendix F.3. The goal is to correctly infer whether $x_t \to x_{t+1}$ or $x_{t+1} \to x_t$ for each channel. This is a useful test case for causal methods since the true direction is known to be from the past to the future. We report in Figure 3 the accuracy as a function of the percentage of channels considered, sorted from highest to lowest confidence (*i.e.* by how high the amplitude of the output of each algorithm is). For average to high confidence, CAREFL is comparable in performance to the baseline methods, but performs better in the low confidence regime. We also note that the performance of CAREFL improves by increasing the sample-size, which is to be expected from a method based on deep learning.

5.2 Interventions

To demonstrate that CAREFL can answer interventional queries, we will consider both a synthetic controlled example, as well as real fMRI data.

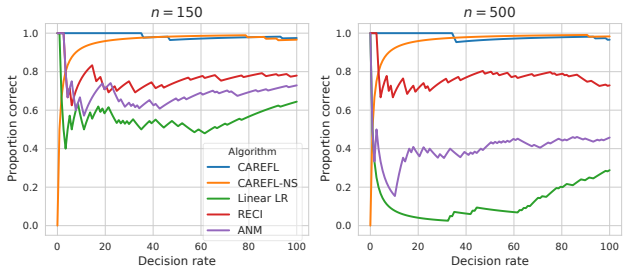

Figure 3: Performance on finding the arrow of time of EEG data, as a function of decision rate (percentage of channels — sorted by decreasing confidence — we have to classify).

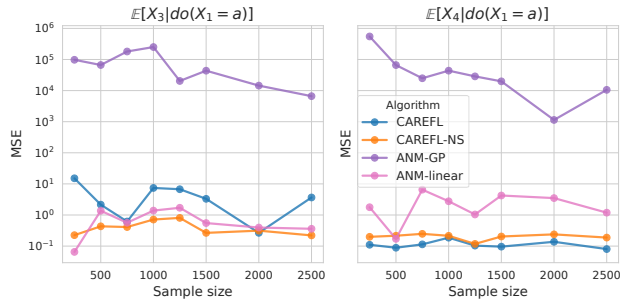

Figure 4: Mean square error for interventional predictions on simulated data, generated using equation (8). The left and right panels consider linear and non-linear interventional distributions.

Synthetic data Consider four-dimensional data generated as

$$\begin{aligned} x_1 &= z_1 & x_3 &= x_1 + c_1 x_2^3 + z_3 \\ x_2 &= z_2 & x_4 &= c_2 x_1^2 - x_2 + z_4 \end{aligned} \tag{8}$$

where each z_i is drawn independently from a standard Laplace distribution, and (c_1, c_2) are random coefficients. From the SEM above we can derive the expectations for x_3 and x_4 under an intervention $do(X_1 = \alpha)$ as being α and $c_2\alpha^2$ respectively.

We compare CAREFL against the regression function from an ANM (Hoyer et al., 2009), where the regression is either linear or a Gaussian process. Figure 4 visualizes the expected mean squared error between predicted expectations for x_3 and x_4 under the intervention $do(X_1 = \alpha)$ for the proposed method, and the true expectations. We note that CAREFL is able to better infer the nature of the true interventional distributions when compared to the baseline.

Interventional fMRI data In order to validate the performance on interventional real-data we applied CAREFL to open-access electrical stimulation fMRI (Thompson et al., 2020). Data was collected across 26 patients with medically refractory epilepsy, which required surgically implanting intracranial electrodes in cortical and subcortical locations. FMRI data was then collected during rest as well as while electrodes were being stimulated. Whilst each patient had electrodes implanted in slightly different locations, we identified 16 patients with electrodes in or near the Cingulate Gyrus and studied these patients exclusively. We further restricted ourselves to studying the data from the Cingulate Gyrus (CG) and Heschl's Gyrus (HG), resulting in bivariate time-series per patient. Full data preprocessing and preparation is described in Appendix F.4.

We compared CAREFL with both linear and additive noise models. Throughout these experiments we assumed the underlying causal structure between regions was known (with CG→HG) and trained each model using the resting-state data. Given the trained model, sessions where the CG was stimulated were treated as interventional sessions, with the task being to predict fMRI activation in HG given CG activity. Whilst the true underlying DAG will be certainly be more complex than the simple bivariate structure considered here, these experiments nonetheless serve as a real dataset benchmark through which to compare various causal inference algorithms. The results are provided in Table 2, where CAREFL is shown to out-perform alternative causal models.

Table 2: Median absolute error for interventional predictions in electrical stimulation fMRI data.

Algorithm	Median abs error (std. dev.)
CAREFL	**0.586 (0.048)**
ANM	0.655 (0.057)
Linear SEM	0.643 (0.044)

5.3 Counterfactuals

We continue with the simple 4 dimensional structural equation model described in equation (8). We assume we observe $\mathbf{x}^{obs} = (2.00, 1.50, 0.81, -0.28)$ and consider the counterfactual values under two distinct scenarios: (i) the expected counterfactual value of x_3 if $x_2 = \alpha$ instead of $x_1 = 2$; (ii) the expected counterfactual value of x_4 if $x_1 = \alpha$ instead of $x_1 = 2$. Counterfactual predictions require us to infer the values of latent variables, called *abduction* step by Pearl (2009a). This is non-trivial for most causal models, but can be easily achieved with CAREFL due to the invertibility of flow models. Figure 5 demonstrates that CAREFL can indeed make accurate counterfactual predictions.

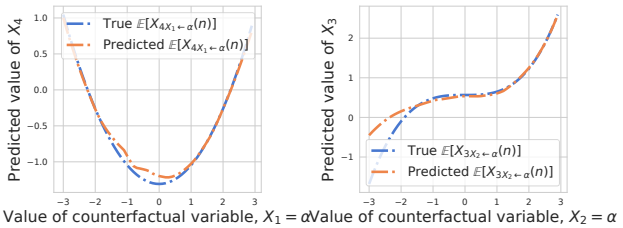

Figure 5: Counterfactual predictions for variables x_3 and x_4. Note that flow is able to obtain accurate counterfactual predictions for a range of values of α.

6 DISCUSSION

Existing identifiability results on causal models other than additive noise models are limited. To our knowledge, the other notable and identifiable non-additive noise models are the post-non-linear model (Zhang and Hyvarinen, 2009, PNL) and the non-stationary non-linear SEM model (Monti et al., 2019, NonSENS). The PNL model assumes that the cause x and the effect y are related through the equation $y = f_2(f_1(x) + n)$, where n is a noise variable independent of x. In contrast to affine flows, the function f_2 is fixed (in the sense of not modulated by the cause x), while being non-linear as opposed to affine. By applying its inverse f_2^{-1} to y, we actually end up with an additive noise model.

In our model, in stark contrast to the PNL model, it is not possible to apply a fixed (as in not a function of the cause) transformation to the effect to revert back to an additive noise model. This is the main reason why the existing identifiability theory doesn't cover our causal model (6). Theorem 1 thus presents a novel identifiability result in the context of non-additive noise models, and the proposed estimation algorithm benefits from it, as was shown in our experiments.

The NonSENS framework allows for general non-linear relationships between cause, noise and effect. Assuming access to non-stationary data, it is identifiable even in such a general case by leveraging recent results in the theory of non-linear ICA (Hyvärinen and Morioka, 2016; Hyvärinen et al., 2019; Khemakhem et al., 2020b,a). In contrast, the proposed model does restrict the nature of non-linear relationships but places no assumptions of nonstationarity, so our model can be applied in more general scenarios. Our work follows a recent trend of combining flexible generative models (such as autoregressive flows and VAEs) with structural causal models (Pawlowski et al., 2020; Wehenkel and Louppe, 2020; Louizos et al., 2017).

In the context of additive noise models, the estimation methods by Hoyer et al. (2009, ANM) and Bloebaum et al. (2018, RECI) require least-squares regressions in both directions. RECI then compares the magnitudes of the residuals, while ANM depends on independence tests between residuals and causes. Choosing the right regression model in both these methods is difficult. As stated by Bloebaum et al. (2018), a very good regression function can reduce the performance of ANM and RECI because it decreases the confidence of the independence tests. We have observed this in our experiments when using neural networks as the regression class, as seen in Figure 1. Importantly, if the additive noise assumption fails to hold, both approaches will fail regardless of the regression class.

CAREFL is specifically leveraging the recent developments in deep learning with the promise of finding computationally efficient methods, as well as improving the statistical efficiency (power) by using likelihood ratios. Furthermore, both ANM and RECI were solely designed for causal discovery, and the invertibility of the system in order to perform interventions and counterfactuals wasn't discussed. So, it is plausible that our model might be preferable even the context of ANM's, in addition to generalizing them.

We note that the likelihood ratio approach by Hyvärinen and Smith (2013) was originally designed for LiNGAM, which is a linear model based on non-Gaussianity (Shimizu et al., 2006). An extension of likelihood ratios to non-linear ANM was also proposed by Hyvärinen and Smith (2013), together with a heuristic approximation which roughly amounts to RECI.

7 CONCLUSION

We argue that autoregressive flow models are well-suited to causal inference tasks, ranging from causal discovery to making counterfactual predictions. This is because we can interpret the ordering of variables in an autoregressive flow in the framework of SEMs.

We show that affine flows in particular define a new class of causal models, where the noise is modulated by the cause. For such models, we prove a completely new causal identifiability result which generalizes additive noise models. We show how to efficiently learn causal structure by selecting the ordering with the highest test log-likelihood and thus present a measure of causal direction based on the likelihood-ratio for non-linear SEMs.

Furthermore, by restricting ourselves to autoregressive flow models we are able to easily evaluate interventional queries by fixing the interventional variable whilst sampling from the flow. The invertible property of autoregressive flows further facilitates the evaluation of counterfactual queries.

In experiments on synthetic and real data, our method outperformed alternative methods in causal discovery as well as interventional and counterfactual predictions.

Acknowledgments

I.K. and R.P.M. were supported by the Gatsby Charitable Foundation. A.H. was supported by a Fellowship from CIFAR.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14.

Bloebaum, P., Janzing, D., Washio, T., Shimizu, S., and Schölkopf, B. (2018). Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pages 900–909.

Dinh, L., Krueger, D., and Bengio, Y. (2014). NICE: Non-linear Independent Components Estimation. *arXiv:1410.8516 [cs]*.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2016). Density estimation using Real NVP. *arXiv:1605.08803 [cs, stat]*.

Dornhege, G., Blankertz, B., Curio, G., and Müller, K.-R. (2004). Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. *IEEE transactions on biomedical engineering*, 51(6):993–1002.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. (2019a). Cubic-Spline Flows. *arXiv:1906.02145 [cs, stat]*.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. (2019b). Neural Spline Flows. *arXiv:1906.04032 [cs, stat]*.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., et al. (2019). fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. (2015). MADE: Masked Autoencoder for Distribution Estimation. *arXiv:1502.03509 [cs, stat]*.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., and Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, 5:13.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, pages 689–696.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. (2018). Neural Autoregressive Flows. *arXiv:1804.00779 [cs, stat]*.

Hyvärinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, pages 3765–3773.

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

Hyvärinen, A., Sasaki, H., and Turner, R. (2019). Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models. *Journal of Machine Learning Research*, 14(Jan):111–152.

Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. (2020a). Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *The 23rd International Conference on Artificial Intelligence and Statistics*.

Khemakhem, I., Monti, R. P., Kingma, D. P., and Hyvärinen, A. (2020b). ICE-BeeM: Identifiable Conditional Energy-Based Deep Models. *arXiv:2002.11537 [cs, stat]*.

Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative Flow with Invertible 1x1 Convolutions. *arXiv:1807.03039 [cs, stat]*.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improving Variational Inference with Inverse Autoregressive Flow. *arXiv:1606.04934 [cs, stat]*.

Kobyzev, I., Prince, S. J. D., and Brubaker, M. A. (2020). Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821*.

Monti, R. P., Zhang, K., and Hyvarinen, A. (2019). Causal discovery with general non-linear relationships using non-linear ICA. In *35th Conference on Uncertainty in Artificial Intelligence, UAI 2019*, volume 35.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016). Distinguishing cause from effect using observational data: Methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204.

Neyman, J. and Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2019). Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*.

Papamakarios, G., Pavlakou, T., and Murray, I. (2018). Masked Autoregressive Flow for Density Estimation. *arXiv:1705.07057 [cs, stat]*.

Pawlowski, N., Castro, D. C., and Glocker, B. (2020). Deep structural causal models for tractable counterfactual inference. *arXiv preprint arXiv:2006.06485*.

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

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

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012.

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2014). Causal Discovery with Continuous Additive Noise Models. *arXiv:1309.6779 [stat]*.

Rezende, D. J. and Mohamed, S. (2015). Variational Inference with Normalizing Flows. *arXiv:1505.05770 [cs, stat]*.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12(Apr):1225–1248.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, Prediction, and Search*. MIT press.

Spirtes, P. and Zhang, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, 3(1):3.

Thompson, W. H., Nair, R., Oya, H., Esteban, O., Shine, J. M., Petkov, C., Poldrack, R. A., Howard, M., and Adolphs, R. (2020). Human esfMRI Resource: Concurrent deep-brain stimulation and whole-brain functional MRI. *bioRxiv*, page 2020.05.18.102657.

Vogt, B. A. (2019). *Cingulate Cortex*. Elsevier.

Wehenkel, A. and Louppe, G. (2020). Graphical normalizing flows. *arXiv preprint arXiv:2006.02548*.

Zhang, K., Huang, B., Zhang, J., Glymour, C., and Schölkopf, B. (2017). Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI: Proceedings of the Conference*, volume 2017, page 1347. NIH Public Access.

Zhang, K. and Hyvarinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, volume 35.

Zhang, K., Wang, Z., Zhang, J., and Schölkopf, B. (2015a). On Estimation of Functional Causal Models: General Results and Application to the Post-Nonlinear Causal Model. *ACM Transactions on Intelligent Systems and Technology*, 7(2):13:1–13:22.

Zhang, K., Zhang, J., and Schölkopf, B. (2015b). Distinguishing Cause from Effect Based on Exogeneity. *arXiv:1504.05651 [cs, stat]*.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483.