Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Nonlinear Causal Discovery with Confounders

Chunlin Li, Xiaotong Shen & Wei Pan

To cite this article: Chunlin Li, Xiaotong Shen & Wei Pan (2023): Nonlinear Causal Discovery with Confounders, Journal of the American Statistical Association, DOI: 10.1080/01621459.2023.2179490

To link to this article: https://doi.org/10.1080/01621459.2023.2179490

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–10: Theory and Methods
https://doi.org/10.1080/01621459.2023.2179490

Nonlinear Causal Discovery with Confounders

Chunlin Li [a], Xiaotong Shen [a], and Wei Pan [b]

[a]School of Statistics, University of Minnesota, Minneapolis, MN; [b]Division of Biostatistics, University of Minnesota, Minneapolis, MN

ABSTRACT

This article introduces a causal discovery method to learn nonlinear relationships in a directed acyclic graph with correlated Gaussian errors due to confounding. First, we derive model identifiability under the sublinear growth assumption. Then, we propose a novel method, named the Deconfounded Functional Structure Estimation (DeFuSE), consisting of a deconfounding adjustment to remove the confounding effects and a sequential procedure to estimate the causal order of variables. We implement DeFuSE via feedforward neural networks for scalable computation. Moreover, we establish the consistency of DeFuSE under an assumption called the strong causal minimality. In simulations, DeFuSE compares favorably against state-of-the-art competitors that ignore confounding or nonlinearity. Finally, we demonstrate the utility and effectiveness of the proposed approach with an application to gene regulatory network analysis. The Python implementation is available at *https://github.com/chunlinli/defuse*. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2021
Accepted February 2023

KEYWORDS
Deconfounding; Directed acyclic graph; Gene regulatory networks; Neural networks; Variable selection

1. Introduction

Causal relationships are fundamental to understanding the mechanisms of complex systems and the consequences of actions in natural and social sciences. Causal discovery, namely to learn a Directed Acyclic Graph (DAG) representing causal relationships, arises in many applications. In gene network analysis, scientists explore gene-to-gene regulatory relationships to unravel the genetic underpinnings of a disease (Sachs et al. 2005). In such a situation, latent confounders such as environmental or lifestyle factors could introduce spurious associations or mask causal relationships in observed gene expression levels, making causal discovery more challenging. Currently, causal discovery from observational data is an important research topic as randomized experiments are often unethical, expensive, or infeasible. In this article, we concentrate on the discovery of causal relationships in the presence of latent confounders.

Linear causal discovery without confounders has been extensively studied (Spirtes, Glymour, and Scheines 2000; Chickering 2002; Tsamardinos, Brown, and Aliferis 2006; Shimizu et al. 2006; de Campos 2006; Jaakkola et al. 2010; de Campos and Ji 2011; Zheng et al. 2018; Gu, Fu, and Zhou 2019; Yuan et al. 2019; Li, Shen, and Pan 2020). However, in practice, many causal relations are nonlinear, raising concerns about using a linear model (Voorman, Shojaie, and Witten 2014). For nonlinear causal models without confounders, three major approaches include (a) nonlinear independent component analysis (Zhang and Hyvärinen 2009; Monti, Zhang, and Hyvärinen 2020), (b) combinatorial search for the causal order (Mooij et al. 2009; Bühlmann, Peters, and Ernest 2014), and (c) continuous constrained optimization for causal structure learning (Zheng et al. 2020). The first estimates the functional relations through the mutual independence of errors. The second determines the causal order based on a certain criterion. For example, the Causal Additive Model (CAM) (Bühlmann, Peters, and Ernest 2014) assumes the nonlinear functions are of additive form and estimates the causal order that maximizes the likelihood. The third approach directly optimizes an objective function subject to a smooth constraint characterizing acyclicity. The most representative example is NOTEARS (Zheng et al. 2020). The reader may consult Peters, Janzing, and Scholkopf (2017) and Glymour, Zhang, and Spirtes (2019) for excellent surveys of nonlinear causal discovery.

In the presence of latent confounders, several methods are available for linear causal discovery. As extensions of the PC algorithm, FCI (Spirtes, Glymour, and Scheines 2000) and its variant RFCI (Colombo et al. 2012) address latent confounders by producing a Partial Ancestral Graph (PAG) instead of a completed partially DAG (CPDAG). Another approach (Frot, Nandy, and Maathuis 2019; Shah et al. 2020) assumes the confounding is pervasive (Chandrasekaran, Parrilo, and Willsky 2012; Wang and Blei 2019) and recovers the CPDAG in two steps. For example, LRpS-GES (Frot, Nandy, and Maathuis 2019) uses the low-rank plus sparse estimator (Chandrasekaran, Parrilo, and Willsky 2012) to remove confounding, followed by the GES algorithm (Chickering 2002) to perform causal structure estimation. Besides, the instrumental variable estimation is a well-known approach but requires the availability of valid instruments (Chen et al. 2018; Li, Shen, and Pan 2021).

Despite the foregoing progress, nonlinear causal discovery with confounders remains largely unexplored. In a bivariate

CONTACT Chunlin Li *li000007@umn.edu* School of Statistics, University of Minnesota, Minneapolis, MN 55455.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

case, the work of Janzing et al. (2009) estimates the confounding effect by minimizing the L_2-distance between data points and a curve evaluated at the estimated values of the confounder. For a multivariate case, it remains unclear whether nonlinearity can help causal discovery with confounding, although third-order differentiability suffices for the identifiability of nonlinear causal discovery without confounders (Peters et al. 2014). Moreover, major computational and theoretical challenges arise when we confront the curse of dimensionality in learning a nonparametric DAG. During the review process, a preprint by Agrawal et al. (2021) proposes a two-step procedure for nonlinear causal discovery in the presence of pervasive confounders. However, for consistent estimation, their method requires that the sample size grows slower than the quadratic graph size, $n \ll p^2$, which may be restrictive, especially for nonparametric estimation.

This article contributes to the following areas. First, we derive a new condition, called the *sublinear growth assumption*, for model identifiability in the presence of latent confounders. Second, we propose a novel approach for causal discovery, called the Deconfounded Functional Structure Estimation (DeFuSE), comprising a deconfounding adjustment and an iterative procedure to reconstruct the topological order of the variables. Third, we implement DeFuSE through feedforward neural networks without assuming additive functional relationships while allowing efficient computation for a reasonable graph size p, say $p = 100$. This is in contrast to traditional nonparametric methods that suffer from inefficiency in high dimensions, such as tensor-product B-splines (Hastie, Tibshirani, and Friedman 2009). Fourth, we develop a novel theory for DeFuSE, establishing its consistency for discovering the underlying DAG structure. DeFuSE requires an assumption for consistent causal discovery, called the *strong causal minimality*, which is an analogy of the strong faithfulness (Uhler et al. 2013) and the beta-min condition (Meinshausen and Bühlmann 2006). A central message of this article is that nonlinearity plays an important role in causal discovery, permitting the separation of the nonlinear causal effects from linear confounding effects.

The rest of the article is structured as follows. Section 2 introduces the DAG model with hidden confounders and the proposed method DeFuSE. Section 3 implements DeFuSE based on feedforward neural networks for scalable computation. Section 4 provides a theoretical guarantee of DeFuSE for consistent discovery. Section 5 presents some numerical examples and compares DeFuSE with CAM, NOTEARS, RFCI, and LRpS-GES, followed by a discussion in Section 6. The Appendix contains additional theoretical results and implementation details, and the supplementary materials contain the technical proofs.

2. Directed Acyclic Graph with Confounders

Consider a random vector $Y = (Y_1, \ldots, Y_p)$ generated from a nonlinear structural equation model with additive confounders and noises,

$$Y_j = f_j\left(Y_{\mathrm{PA}(j)}\right) + \eta_j + e_j, \quad j \in V = \{1, \ldots, p\}, \tag{1}$$

where f_j maps the subvector $Y_{\mathrm{PA}(j)} = (Y_k)_{k\in\mathrm{PA}(j)}$ to a real number, $\mathrm{PA}(j) \subseteq V \setminus \{j\}$ is an index subset, $\eta = (\eta_1, \ldots, \eta_p) \sim N_p(0, \Sigma_\eta)$ is a vector of hidden confounders and is independent of random errors $e = (e_1, \ldots, e_p) \sim N_p(0, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2))$, Σ_η is an unknown covariance matrix, and $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ is an unknown diagonal matrix. Then (1) is associated with a directed graph $G = (V, E)$ such that $E = \{k \to j : k \in \mathrm{PA}(j),\ j \in V\}$. In this situation, $\mathrm{PA}(j)$ denotes the set of parents of j. Throughout this article, we assume that G is a DAG in that no directed path $j \to \cdots \to j$ exists in G. As a result, (1) generalizes the nonlinear DAG without unmeasured confounders (Hoyer et al. 2008; Peters et al. 2014) and the linear DAG (Peters and Bühlmann 2014).

In (1), we assume the *causal minimality* to ensure that the effect of each parent is nonvanishing. In other words, we require $\mathrm{PA}(j) = \mathrm{ARG}(f_j)$; $j = 1, \ldots, p$, where $\mathrm{ARG}(f_j)$ denotes the minimal argument set $B \subseteq \mathrm{PA}(j)$ such that the value of f_j only depends on $Y_B = (Y_k)_{k\in B}$. In particular, if f_j is a constant function, we have $\mathrm{PA}(j) = \mathrm{ARG}(f_j) = \emptyset$. When $\eta \equiv 0$ (no confounder), this definition agrees with the usual causal minimality condition (Pearl 2009), requiring that the probability distribution of Y is not Markov to any proper subgraph of G. The *causal minimality*, as a form of causal faithfulness (Spirtes, Glymour, and Scheines 2000), ensures that the problem of nonlinear causal discovery is well-defined.

Equivalently, we rewrite (1) by letting $\varepsilon_j = \eta_j + e_j$,

$$Y_j = f_j\left(Y_{\mathrm{PA}(j)}\right) + \varepsilon_j, \quad j \in V = \{1, \ldots, p\}, \tag{2}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p) \sim N(0, \Sigma)$ and $\Sigma = \Sigma_\eta + \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Whereas (1) has a clear causal interpretation, (2) is simpler for the subsequent discussion. Our goal is to discover the causal relations between variables $Y_1, \ldots, Y_p$ by identifying $\{f_j\}_{1\le j\le p}$ and $\{\mathrm{PA}(j)\}_{1\le j\le p}$. One major challenge is that the error ε_j may be correlated with $Y_{\mathrm{PA}(j)}$ due to unmeasured confounders.

2.1. Model Identifiability

This section establishes the identifiability conditions for (2). First, we introduce the concept of *topological depth* for a DAG $G = (V, E)$ with nodes $V = \{1, \ldots, p\}$ and directed edges $E \subseteq V \times V$. A node j is a *root* if it has no parent, that is, $\mathrm{PA}(j) = \emptyset$. If there exists a directed path $k \to \cdots \to j$, then node k is an ancestor of j and j is a descendant of k. The *topological depth* d_j of node $j \in V$ is the maximal length of a directed path from a root to j. Clearly, a root node has depth zero, and we have $0 \le d_j \le d_{\max} \le p - 1$ for $j \in V$, where $d_{\max}$ is the length of the longest directed path in G. Let $V(d) = \{j : d_j < d\}$ be the set of nodes with topological depth less than d, where $1 \le d \le d_{\max} + 1$. Then $\emptyset \equiv V(0) \subseteq V(1) \subseteq \cdots \subseteq V(d_{\max} + 1) = V$ and $V(d_j)$ contains all the ancestors (and hence all the parents) of Y_j but contains no descendant of Y_j. See Figure 1 for an illustration.

Next, we present a new condition for $\{f_j\}_{1\le j\le p}$ and $\{\mathrm{PA}(j)\}_{1\le j\le p}$ in (2) to be identifiable. For continuous function $f : \mathbb{R}^m \to \mathbb{R}$, f is of *sublinear growth* if $\lim_{\|x\|\to\infty} f(x)/\|x\| = 0$, where $\|\cdot\|$ is the Euclidean norm.

Condition 1. Assume that $\{f_j\}_{1\le j\le p}$ are of sublinear growth.

For example, Condition 1 is satisfied if $\{f_j\}_{1\le j\le p}$ are continuous and bounded. This sublinear growth assumption imposes

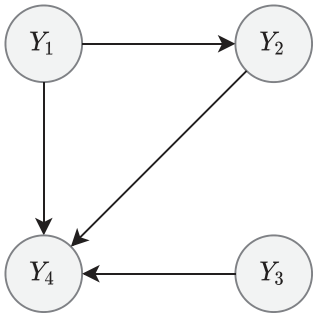


Figure 1. Topological depth: $d_1 = d_3 = 0$ (nodes 1 and 3 are root nodes), $d_2 = 1$, $d_4 = 2$. Here $V(1) = \{1, 3\}$, $V(2) = \{1, 2, 3\}$, and $V(3) = V = \{1, 2, 3, 4\}$.

restrictions on the nonlinearity of $\{f_j\}_{1\le j\le p}$, in contrast to the third-order differentiability condition for DAGs without confounders (Hoyer et al. 2008; Peters et al. 2014).

Theorem 1 (Identifiability). Assume Condition 1 is satisfied.

(A) The sets $V(1) \subseteq \cdots \subseteq V(d_{\max})$ are uniquely identifiable for almost every positive definite Σ with respect to the Lebesgue measure, where the set of such Σ is denoted as Ψ. Moreover, for $\Sigma \in \Psi$, if $d_j = d$, then $Y_j - \mathrm{E}\left(Y_j \mid Y_{V(d)}\right)$ is normally distributed with mean zero and constant variance $\mathrm{var}\left(Y_j \mid Y_{V(d)}\right)$; if $d_j > d$, then $Y_j - \mathrm{E}\left(Y_j \mid Y_{V(d)}\right)$ is not normally distributed; $j = 1, \ldots, p$.

(B) Given $V(1) \subseteq \cdots \subseteq V(d_{\max})$, we have $\{f_j\}_{1\le j\le p}$ and $\{\mathrm{PA}(j)\}_{1\le j\le p}$ are well-defined and identifiable from the distribution of Y.

By Theorem 1, model (2) is generically identifiable under Condition 1. Different from Frot, Nandy, and Maathuis (2019), Theorem 1 does not require pervasive confounding. The sublinear growth assumption (Condition 1) allows us to separate the linear confounding effect from nonlinear causal relationships.

2.2. DeFuSE

This section proposes the causal discovery method Deconfounded Functional Structure Estimation (DeFuSE). We commence with least squares regressions of $\{Y_j\}_{j\notin V(d)}$ on $Y_{V(d)}$,

$$Y_j = \underbrace{\mathrm{E}(Y_j \mid Y_{V(d)})}_{\text{(i)}} + \underbrace{Y_j - \mathrm{E}(Y_j \mid Y_{V(d)})}_{\text{(ii)}},$$

where (i) is the regression function and (ii) is the residual of the regression. By Theorem 1, (ii) is normally distributed if and only if $d_j = d$, suggesting that normality tests (e.g., the Anderson-Darling test (Anderson and Darling 1952)) for $\{Y_j - \mathrm{E}(Y_j \mid Y_{V(d)})\}_{j\notin V(d)}$ can be used to identify $V(d+1)$. Further, if $d_j = d$, then (i) becomes

$$\mathrm{E}(Y_j \mid Y_{V(d)}) = f_j(Y_{\mathrm{PA}(j)}) + \mathrm{E}(\varepsilon_j \mid Y_{V(d)}),$$

where $\mathrm{E}(\varepsilon_j \mid Y_{V(d)})$ is the bias arising from hidden confounding. Theorem 2 allows us to estimate $\{f_j\}_{j\in V(d+1)}$ and $\{\mathrm{PA}(j)\}_{j\in V(d+1)}$ by regressions with deconfounding adjustment.

Theorem 2. In (2), if $d_j = d$, then

$$\mathrm{E}(Y_j \mid Y_{V(d)}) = f_j(Y_{\mathrm{PA}(j)}) + \langle \xi_{V(d)}, \beta_j \rangle, \tag{3}$$

where $\xi_{V(d)} \equiv (Y_k - \mathrm{E}(Y_k \mid Y_{V(d_k)}))_{k\in V(d)}$, β_j is a parameter vector, $\langle \cdot, \cdot \rangle$ is the Euclidean inner product, and we define $\langle \xi_{V(d)}, \beta_j \rangle \equiv 0$ whenever $V(d) = \emptyset$.

Now, we develop an algorithm that iteratively estimates $V(d+1)$, $\xi_{V(d+1)}$, $\{f_j\}_{j\in V(d+1)}$, and $\{\mathrm{PA}(j)\}_{j\in V(d+1)}$, given $V(d)$ and $\xi_{V(d)}$ as input. To proceed, suppose an independent sample $\{(Y_1^{(i)}, \ldots, Y_p^{(i)})\}_{1\le i\le n}$ from model (2) is given. Let $\widehat{\xi}_{V(d)}^{(i)} = (Y_k^{(i)} - \widehat{Y}_k^{(i)})_{k\in V(d)}$ be the estimated residual vector for the ith observation, where $\widehat{Y}_k^{(i)} = \widehat{f}_k\big(Y_{V(d_k)}^{(i)}\big) + \langle \widehat{\xi}_{V(d_k)}^{(i)}, \widehat{\beta}_j \rangle$ for $k \in V(d)$. Based on (3), we regress each variable in $\{Y_j\}_{j\notin V(d)}$ on $\big(Y_{V(d)}, \xi_{V(d)}\big)$,

$$(\widehat{f}_j, \widehat{\beta}_j) = \underset{\{(f_j,\beta_j): f_j\in\mathcal{F}_j\}}{\arg\min} \sum_{i=1}^{n} \Big(Y_j^{(i)} - f_j\big(Y_{V(d)}^{(i)}\big) - \langle \widehat{\xi}_{V(d)}^{(i)}, \beta_j \rangle\Big)^2 \quad \text{s.t.} \quad |\mathrm{ARG}(f_j)| \le \kappa_j, \tag{4}$$

where $|\mathrm{ARG}(f_j)|$ is the effective input dimension of f_j, $\kappa_j \ge 0$ is an integer-valued hyperparameter and is estimated via a standalone validation set (see Section A.3), and $\mathcal{F}_j$ is a function space consisting of sublinear growth continuous functions. Then we perform normality tests for $\{(\widehat{\xi}_j^{(1)}, \ldots, \widehat{\xi}_j^{(n)})\}_{j\notin V(d)}$, and estimate $V(d+1)$ by including $V(d)$ and all the indices failing to reject the tests. Finally, we estimate $\{\widehat{\mathrm{PA}}(j)\}_{j\in V(d+1)}$ by $\widehat{\mathrm{PA}}(j) = \mathrm{ARG}(\widehat{f}_j)$.

We summarize the procedure in Algorithm 1, where a boldface letter denotes a data vector/matrix of sample size n.

Algorithm 1: DeFuSE

Input: An $n \times p$ data matrix $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p)$;
Parameters: significance level α for normality test; hyperparameters $\{\kappa_j\}_{1\le j\le p}$;

Let $V(0) \leftarrow \emptyset$ and $d \leftarrow 0$;
while $V(d) \neq V$ **do**
  Regress $\{\boldsymbol{Y}_j\}_{j\notin V(d)}$ on $(\boldsymbol{Y}_{V(d)}, \widehat{\boldsymbol{\xi}}_{V(d)})$ based on (4);
  Update $\{\widehat{\boldsymbol{\xi}}_j \leftarrow \boldsymbol{Y}_j - \widehat{\boldsymbol{Y}}_j\}_{j\notin V(d)}$;
  Let $V(d+1) \leftarrow V(d) \cup \{j \notin V(d) :$ $\widehat{\boldsymbol{\xi}}_j$ fails to reject the normality test$\}$;
  Let $\{\widehat{\mathrm{PA}}(j) \leftarrow \mathrm{ARG}(\widehat{f}_j)\}_{j\in V(d+1)}$ and $d \leftarrow d+1$;
end
Output: $\{\widehat{f}_j\}_{1\le j\le p}$ and $\{\widehat{\mathrm{PA}}(j)\}_{1\le j\le p}$;

Remark 1 (Normality test and the choice of α). For implementation, we use the Anderson-Darling test (Anderson and Darling 1952) to examine the null hypotheses

$$\mathcal{H}_0^{(j,d)} : Y_j - \mathrm{E}(Y_j \mid Y_{V(d)}) \text{ is normal; } j \notin V(d),\ 0 \le d \le d_{\max}.$$

Other tests or metrics, such as the Wasserstein distance, can also be used. Moreover, the normality test can be combined with a goodness of fit measure to further improve performance. The significance level $0 < \alpha < 1$ is a hyperparameter similar to that in the PC algorithm (Kalisch and Bühlman 2007). To choose α, denoting by $\mathcal{T}$ the set of true null hypotheses, then $P\Big(\text{some } \mathcal{H}_0^{(j,d)} \in \mathcal{T} \text{ is rejected}\Big) \le \sum_{\mathcal{H}_0^{(j,d)}\in\mathcal{T}} P\Big(\mathcal{H}_0^{(j,d)} \text{ is rejected}\Big) \approx |\mathcal{T}|\alpha$. For $1 \le d \le d_{\max}+1$,

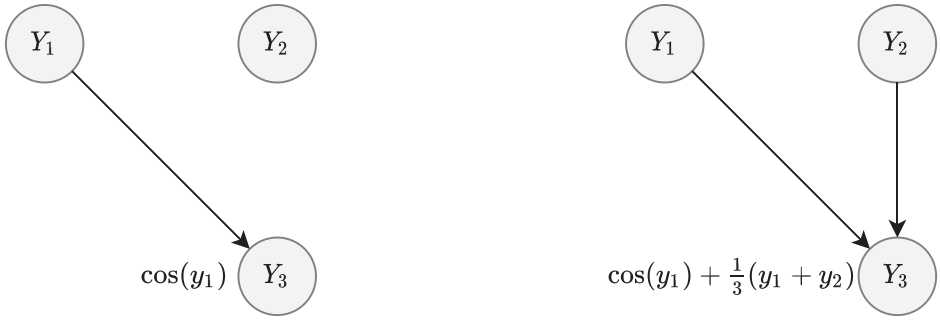


Figure 2. Display of the directed acyclic graph in Example 1.

identifying $V(d)$ requires $p - |V(d-1)|$ tests, among which $|V(d)| - |V(d-1)|$ null hypotheses are true and $p - |V(d)|$ are not. Thus, $|\mathcal{T}| = \sum_{d=1}^{d_{\max}+1}(|V(d)| - |V(d-1)|) = p$, suggesting an empirical rule $\alpha = o(1/p)$ so that $|\mathcal{T}|\alpha \to 0$.

Finally, Example 1 illustrates the importance of deconfounding for causal discovery.

Example 1. Consider a special case of (1) with three variables,

$$Y_1 = e_1 + \eta, \quad Y_2 = e_2 + \eta, \quad Y_3 = \cos(Y_1) + e_3 + \eta, \quad (5)$$

where $e_1, e_2, e_3, \eta \sim N(0,1)$ independently; see Figure 2. As a special case of (3), we have $\mathrm{E}(Y_3 \mid Y_1, Y_2) = \cos(Y_1) + \mathrm{E}(\eta \mid Y_1, Y_2) = \cos(Y_1) + Y_1/3 + Y_2/3$, where $d_3 = 1$, $V(1) = \{1, 2\}$, $\xi_{V(1)} = (\xi_1, \xi_2) = (e_1 + \eta, e_2 + \eta)$, and $\xi_{V(2)} = \xi_3 = e_3 + (\eta - e_1 - e_2)/3$. The presence of $Y_2/3$ is due to the confounder η. If we have regressed Y_3 on Y_1 and Y_2 to identify the parent variables of Y_3, then the regression would yield a true discovery $Y_1 \to Y_3$ and a false discovery $Y_2 \to Y_3$. Consequently, direct regression of Y_j on $Y_{V(d_j)}$ without any adjustment renders false discovery of functional causal relations.

3. DeFuSE via Neural Networks

Solving (4) is challenging for a large-scale problem due to fitting nonparametric functions. Existing nonparametric methods such as tensor-product splines and kernels are not scalable in a growing sample size and dimension. For example, tensor-product B-splines least squares regression suffers from exponential growth of time and space complexity with increasing dimensions. To overcome this difficulty, we solve (4) via a Feedforward Neural Network (FNN) together with stochastic gradient descent for scalable computation.

Specifically, for $d_j \geq d$, we approximate $f_j\left(Y_{V(d)}\right) + \langle \xi_{V(d)}, \beta_j \rangle$ by an FNN,

$$\begin{aligned} g_j\left(Y_{V(d)}, \xi_{V(d)}\right) &= f_j^L \circ \cdots \circ f_j^1\left(Y_{V(d)}\right) + \langle \xi_{V(d)}, \beta_j \rangle, \, f_j^l(\cdot) \\ &= \sigma^l\left(W^l(\cdot) + b^l\right); \; l = 1, \ldots, L, \end{aligned} \quad (6)$$

where $W^l \in \mathbb{R}^{h_l \times h_{l-1}}$ is the weight matrix of links from the $(l-1)$th to the lth layer, $b^l \in \mathbb{R}^{h_l}$ is the bias vector in the lth layer, h_l is the number of neurons in the lth layer with $h_l = h$; $l = 1, \ldots, L-1$, and $h_L = 1$, L is the number of layers, and $\sigma^l(\cdot)$ is an activation function. For $l = 1, \ldots, L-1$, we use the Rectifier Linear Unit (ReLU) activation $\sigma^l(z) = \max(0, z)$.

To solve (4), consider a FNN parameter vector $\theta_j = ((W_j^l, b_j^l)_{1 \leq l \leq L}, \beta_j)$ which belongs to a parameter space Θ_d. We impose constraints $\sum_{k \in V(d)} \min(\|W_k^1\|/\tau, 1) \leq \kappa_j$ on the kth column W_k^1 of the weight matrix W^1 at the first layer to enforce the constraint $|\textsc{arg}(f_j)| \leq \kappa_j$ in (4), where $\min(|\cdot|/\tau, 1)$ is to approximate $\mathrm{I}(\cdot \neq 0)$ as $\tau \to 0^+$ (Shen, Pan, and Zhu 2012). As such, if $W_k^1 = 0$ then $g_j\left(Y_{V(d)}, \xi_{V(d)}\right)$ does not depend on Y_k. Finally, we regularize the FNN by an L_2-norm constraint $\|\theta_j\| \leq s$ on the model parameters θ_j for numerical consideration. This leads to the following regression for estimating (f_j, β_j),

$$\begin{aligned} \min_{\{\theta_j : \|\theta_j\| \leq s\}} \quad & \sum_{i=1}^{n} \left(Y_j^{(i)} - f_j\left(Y_{V(d)}^{(i)}\right) - \langle \widehat{\xi}_{V(d)}^{(i)}, \beta_j \rangle\right)^2, \\ \text{s.t.} \quad & \sum_{k \in V(d)} \min(\|W_k^1\|/\tau, 1) \leq \kappa_j, \\ & \sum_{k \in V(d)} \min(|\beta_{j,k}|/\tau, 1) \leq \varsigma_j, \end{aligned} \quad (7)$$

where $\tau > 0$, $0 \leq \kappa_j \leq |V(d)|$, $0 \leq \varsigma_j \leq |V(d)|$, and $s \geq 0$ are hyperparameters. See Section A.3 for more details on network training and hyperparameter tuning.

Remark 2. Algorithm 1 requires $O(d_{\max}(p-1))$ normality tests and regressions (4). Each regression (4), solved by (7) with stochastic gradient descent, requires $O(N_{\text{epoch}} n \dim(\theta))$ operations, where N_{epoch} is the number of epochs in training and one epoch means that each sample in training has an opportunity to update model parameters.

4. Learning Theory

This section develops a novel theory to quantify the finite-sample error of DeFuSE. In what follows, c_1-c_6 are positive constants and $^\circ$ decorates the truth. Let $\mathcal{G}_j$ be the function space of regression functions $g_j(\cdot, \star) = f_j(\cdot) + \left\langle \star, \beta_j^\circ \right\rangle$, and denote the true regression function by $g_j^\circ(\cdot, \star) = f_j^\circ(\cdot) + \left\langle \star, \beta_j^\circ \right\rangle$. By definition, $\textsc{pa}^\circ(j) = \textsc{arg}(f_j^\circ)$.

Condition 2. There exists an approximating function $g_j^*(\cdot, \star) = f_j^*(\cdot) + \left\langle \star, \beta_j^\circ \right\rangle \in \mathcal{G}_j$ such that $\|g_j^* - g_j^\circ\|_{L_2} = \|f_j^* - f_j^\circ\|_{L_2} \leq c_3 \epsilon_n$; $j = 1, \ldots, p$, where $\|\cdot\|_{L_2}$ is the L_2-norm with respect to measure P. Moreover, assume $\{f_j^\circ\}_{1 \leq j \leq p}$ are continuous and $\|f_j^\circ\|_\infty \leq c_1$, where $\|\cdot\|_\infty$ is the sup-norm.

To measure the signal strength, we define the degree of nonlinear separation as

$$D_{\min} = \min_{1\le j\le p} \inf \left\{ \frac{\|g_j - g_j^\circ\|_{L_2}^2}{|\text{PA}^\circ(j) \setminus \text{ARG}(f_j)|} : \begin{array}{ll} g_j \in \mathcal{G}_j, & \text{ARG}(f_j) \neq \text{PA}^\circ(j), \\ \|\beta_j\|_0 \le \varsigma^\circ, & |\text{ARG}(f_j)| \le |\text{PA}^\circ(j)| \end{array} \right\}.$$

Condition 3 (Strong causal minimality). Assume $D_{\min} \ge c_4 \max\left(4\epsilon_n^2, n^{-1}\log n, n^{-1}\log p\right)$, where $c_4 \ge 1$.

The strong causal minimality (Condition 3) requires that the signal strengths of parent variables are sufficiently strong so that the corresponding causal function is distinguishable from those supported on nonparent variables. It is a strong version of the causal minimality for nonlinear causal discovery from a finite sample, similar to the strong faithfulness (Uhler et al. 2013) for linear causal discovery and the beta-min condition (Meinshausen and Bühlmann 2006) for high-dimensional variable selection.

Theorem 3 (Error bounds for DeFuSE). Assume Conditions 1–3, Conditions A1–A2 in Section A.2 are met and $\Sigma \in \Psi$.

(A) The DAG recovery error is $P(\widehat{G} \neq G^\circ) \le c_6 \exp(-c_5 n\epsilon_n^2 - \log n) + \pi_\alpha(G^\circ)$, when the hyperparameters $\kappa_j = |\text{PA}^\circ(j)|$ and $\|\beta_j^\circ\|_0 \le \varsigma_j \le \varsigma^\circ$; $1 \le j \le p$, where $\pi_\alpha(G^\circ)$ is the normality test error given the true model. Consequently, $P(\widehat{G} \neq G^\circ) \to 0$ provided that $\pi_\alpha(G^\circ) \to 0$, as $n \to \infty$.

(B) The regression estimation error is $\max_{1\le j\le p} \|\widehat{g}_j - g_j^\circ\|_{L_2} = O_p(\epsilon_n)$. Suppose f_j° satisfies $\|f_j^\circ\|_\infty \le C$ and has bounded support; $1 \le j \le p$. Then the causal function estimation error is $\max_{1\le j\le p} \|\widehat{f}_j - f_j^\circ\|_{L_2} = O_p(\epsilon_n)$ provided that $\|\widehat{f}_j\|_\infty \le C'$ for $C' \ge C$.

Typically, we have $\pi_\alpha(G^\circ) \to 0$ when $\alpha = o(1/p)$ and the dimension p does not grow too fast. Moreover, Theorem 3 indicates that hyperparameter κ_j is critical to consistent discovery, while ς_j is less important provided that $\varsigma_j \ge \|\beta_j^\circ\|_0$ and is not too large; see also Section A.3.

Next, we apply Theorem 3 to the implementation via FNNs in (7). Before proceeding, we define $\mathcal{C}_j^r$, the space of functions with r-continuous derivatives over the domain $\mathbb{R}^{|\text{PA}^\circ(j)|}$. For any function $f_j \in \mathcal{C}_j^r$, the $\mathcal{C}_j^r$-norm of f_j is defined as

$$\|f_j\|_{\mathcal{C}_j^r} = \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<r} \|\partial^{\boldsymbol{\alpha}} f_j\|_\infty + \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor r\rfloor} \sup_{x_1\neq x_2} \frac{|\partial^{\boldsymbol{\alpha}} f(x_1) - \partial^{\boldsymbol{\alpha}} f(x_2)|}{\|x_1 - x_2\|_\infty^{r-\lfloor r\rfloor}},$$

where $\partial^{\boldsymbol{\alpha}} = \partial^{\alpha_1}\cdots\partial^{\alpha_{|\text{PA}(j)|}}$ with $\boldsymbol{\alpha} \in \mathbb{N}^{|\text{PA}(j)|}$ and $|\boldsymbol{\alpha}| = \sum_{k=1}^{|\text{PA}(j)|}\alpha_k$; $j = 1,\ldots,p$. In what follows, C_1–C_3 are positive constants that may depend on (κ°, r).

Condition 4. Assume $f_j^\circ \in \left\{f_j \in \mathcal{C}_j^r : \|f_j\|_{\mathcal{C}_j^r} \le C_1\right\}$, where r does not depend on (p, n).

Theorem 4 (Consistency of FNN-DeFuSE). Under Conditions 3–4, and A2 in Section A.2, DeFuSE implemented by FNNs in (7) consistently recovers all causal relations defined in (2) with $\epsilon_n^2 = C_3(n^{-r/(r+\kappa^\circ+\varsigma^\circ)}(\log n)^3 + n^{-1}(\kappa^\circ + \varsigma^\circ)\log p)$ in Theorem 3, provided that the width of the FNN $h = C_2\epsilon_n^{-\kappa^\circ/r}$ and its depth $L = C_2\log(1/\epsilon_n)$, the hyperparameters $s = C_2\epsilon_n^{-(\kappa^\circ+\varsigma^\circ)/r}\log(1/\epsilon_n)$, $\kappa_j = |\text{PA}^\circ(j)|$, $\|\beta_j^\circ\|_0 \le \varsigma_j \le \varsigma^\circ$; $j = 1,\ldots,p$. Here, the FNN function space $\mathcal{G}_j = \{g_j = g_j(\cdot;\theta) : \theta \in \Theta_j\}$ is associated with the FNN parameter space

$$\Theta_j = \left\{\theta = ((W^l, b^l)_{1\le l\le L}, \beta_j) : \max_{1\le l\le L} h_l \le h,\ \|\theta\| \le s\right\};\quad j = 1,\ldots,p.$$

It is worth noting that the rate $\epsilon_n^2 \asymp n^{-r/(r+\kappa^\circ+\varsigma^\circ)}(\log n)^3 + n^{-1}(\kappa^\circ+\varsigma^\circ)\log p$ for FNN relies on the approximation result of Schmidt-Hieber (2019) as well as the choice of L, h, and s. This rate agrees with Farrell, Liang, and Misra (2021) up to logarithm terms; however, it is slower than $n^{-r/(r+(\kappa^\circ+\varsigma^\circ)/2)}$ in view of Stone (1982) for nonparametric regression over $[0,1]^{\kappa^\circ+\varsigma^\circ}$, suggesting that it may be suboptimal. This may be due to the approximation, namely the use of non-differentiable ReLU FNNs to approximate smooth functions.

5. Numerical Examples

5.1. Simulations

This section examines the operating characteristics of DeFuSE and compares DeFuSE with CAM (Bühlmann, Peters, and Ernest 2014), NOTEARS (FNN version) (Zheng et al. 2020), LRpS-GES (Frot, Nandy, and Maathuis 2019), and RFCI (Colombo et al. 2012). We implement DeFuSE in Python. For competitors, we use R packages for CAM (`CAM`), RFCI (`pcalg`), and LRpS-GES (`lrpsadmm` and `pcalg`), and use a Python program for NOTEARS (`notears`).

In simulations, we consider two types of DAGs with hidden confounders. Define an adjacency matrix $\boldsymbol{U} = (U_{jk})_{p\times p}$ of a DAG as $U_{jk} = 1$ if $j \in \text{PA}(k)$ and 0 otherwise.

Random DAG. Consider a sparse graph where the edges are added independently with equal probability. In particular, an adjacency matrix $\boldsymbol{U} \in \{0,1\}^{p\times p}$ is randomly generated: $P(U_{jk} = 1) = s$ if $j < k$ and $P(U_{jk} = 1) = 0$ otherwise, where s controls the degree of sparseness of the DAG. In our simulation, we choose $s = 1/p$.

Hub DAG. Consider a sparse graph with a hub node. Let $\boldsymbol{U} \in \{0,1\}^{p\times p}$, where $U_{1k} = 1$ and $U_{jk} = 0$ otherwise. In this case, node 1 has a dense neighborhood, but the whole DAG remains sparse.

Simulated data. Given $\boldsymbol{U}$, we generate a random sample of size n from

$$Y_j = \alpha_0 Y_{k_1} Y_{k_2} + \sum_{k\in\text{PA}(j)} \alpha_{j,k} f_{j,k}(Y_k + \omega_{j,k}) + \varepsilon_j;\quad j = 1,\ldots,p, \quad (8)$$

where the function $f_{j,k}$ is randomly sampled from $\{x \mapsto x^2, x \mapsto \cos(x)\}$, the coefficients $\alpha_{j,k} \sim \text{Uniform}([-3,-2]\cup[2,3])$, $\omega_{j,k} \sim \text{Uniform}([-1,1])$, and

$$\begin{cases} \alpha_0 = 0, & |\text{PA}(j)| = 1, \\ \alpha_0 = 1,\ k_1, k_2 \text{ are randomly sampled from } \text{PA}(j), & |\text{PA}(j)| > 1. \end{cases}$$

For error terms, let $\varepsilon \sim N(0,\Sigma)$ with $\Sigma_{jj} = 2$ for $1 \le j \le p$, $\Sigma_{2k-1,2k} = \Sigma_{2k,2k-1} = 1$ for $1 \le k \le \lfloor p/2\rfloor$, and $\Sigma_{jj'} = 0$ otherwise. Of note, (8) violates Condition 1 as the functions $(y_1, y_2) \mapsto \alpha_0 y_1 y_2$ and $f_{j,k}$ may not be of sublinear growth.

Table 1. Averaged False Positive Rate (FPR), False Discovery Rate (FDR), True Positive Rate (TPR), Structural Hamming Distance (SHD), and their standard deviations in parenthesis, for five methods based on 50 replications.

Graph		Random				Hub			
(p,n)	Method	FPR	FDR	TPR	SHD	FPR	FDR	TPR	SHD
(30,500)	DeFuSE	0.00(.00)	0.12(0.06)	0.93(0.04)	2.6(1.2)	0.00(0.00)	0.06(0.06)	0.87(0.10)	5.3(4.6)
	DeFuSE*	0.00(0.00)	0.13(0.11)	0.93(0.07)	1.7(1.4)	0.00(0.00)	0.07(0.10)	0.91(0.16)	4.2(5.6)
	CAM	0.03(0.00)	0.52(0.02)	1.0(0.02)	14.2(1.0)	0.09(1.0)	0.69(0.05)	0.53(0.07)	48.2(6.9)
	NOTEARS	0.28(0.07)	0.91(0.02)	0.80(0.13)	120.2(31.6)	0.19(0.02)	0.84(0.05)	0.52(0.17)	94.3(12.8)
	RFCI	0.07(0.01)	0.89(0.03)	0.29(0.11)	26.8(1.2)	0.22(0.02)	0.95(0.01)	0.04(0.01)	74.4(3.7)
	LRpS-GES	0.07(0.01)	0.91(0.03)	0.21(0.07)	31.9(1.7)	0.08(0.01)	0.92(0.01)	0.06(0.01)	44.5(1.4)
(100,500)	DeFuSE	0.00(0.00)	0.03(0.03)	0.92(0.03)	4.0(1.7)	0.00(0.00)	0.05(0.03)	0.72(0.24)	31.4(23.7)
	DeFuSE*	0.00(0.00)	0.16(0.06)	0.85(0.06)	10.6(3.0)	0.00(0.00)	0.10(0.18)	0.71(0.27)	32.9(26.2)
	CAM	0.01(0.00)	0.61(0.01)	1.0(0.01)	57.4(2.5)	0.05(0.01)	0.94(0.01)	0.16(0.03)	306.3(13.0)
	NOTEARS	0.04(0.02)	0.93(0.04)	0.18(0.15)	130.6(24.8)	0.18(0.02)	0.96(0.01)	0.03(0.05)	992.6(65.4)
	RFCI	0.02(0.00)	0.95(0.02)	0.15(0.06)	83.5(1.1)	0.07(0.01)	0.99(0.01)	0.01(0.00)	268.6(6.7)
	LRpS-GES	0.02(0.00)	0.96(0.01)	0.10(0.04)	83.3(2.0)	–	–	–	–

NOTE: A smaller value of FPR, FDR, and SHD indicates higher accuracy, whereas a larger value of TPR means higher accuracy. For DeFuSE*, the data are standardized. For hub DAG, when $p = 100$ and $n = 500$, LRpS-GES fails to deliver the computational results after 96 hr.

Metrics. For evaluation, we consider four graph metrics: the False Discovery Rate (FDR), the False Positive Rate (FPR), the True Positive Rate (TPR), and the Structural Hamming distance (SHD). To compute the metrics, let TP, RE, and FP be the numbers of identified edges with correct directions, those with wrong directions, and estimated edges not in the skeleton of the true graph. Moreover, denote by PE the total number of estimated edges, TN the number of correctly identified non-edges, and FN the number of missing edges compared to the true skeleton. Then

$$\text{FDR} = (\text{RE} + \text{FP})/\text{PE}, \quad \text{FPR} = (\text{RE} + \text{FP})/(\text{FP} + \text{TN}),$$
$$\text{TPR} = \text{TP}/(\text{TP} + \text{FN}), \quad \text{SHD} = \text{FP} + \text{FN} + \text{RE}.$$

Note that LRpS-GES outputs a completed partially DAG (CPDAG) and RFCI outputs a partial ancestral graph (PAG). Both PAG and CPDAG may contain undirected edges, in which case they are evaluated favorably by assuming the correct directions for undirected edges whenever possible, similar to Zheng et al. (2020).

As suggested in Table 1, DeFuSE performs the best across all the situations in terms of FPR, FDR, TPR, and SHD. As expected, CAM and NOTEARS cannot treat unobserved confounders, whereas RFCI and LRpS-GES cannot deal with nonlinear causal relationships. It is worth noting that DeFuSE* takes standardized data as input and achieves comparable performance to DeFuSE, indicating that DeFuSE is insensitive to the degree of varsortability (Reisach, Seiler, and Weichwald 2021). Moreover, DeFuSE seems robust in the absence of Condition 1; see also Theorem A1 in Appendix and discussions there. Overall, nonlinearity helps identify causal relations, allowing for a separation of nonlinear causal effects from linear confounding effects.

Sensitivity to normality test significance level α. In the above experiments, we use the Anderson-Darling test (Anderson and Darling 1952) with $\alpha = 0.025$ as the default choice. Now, we assess the algorithmic sensitivity to different choices of $\alpha \in \{0.1, 0.05, 0.025, 0.01\}$.

As suggested in Table 2, the overall performance of DeFuSE seems insensitive to the choice of α, although the default choice $\alpha = 0.025$ may be sub-optimal. Based on our limited numerical experience, we suggest $\alpha = o(1/p)$ as an empirical rule to reduce the tuning cost of α; see also Remark 1.

Table 2. Sensitivity analysis: Averaged False Positive Rate (FPR), False Discovery Rate (FDR), True Positive Rate (TPR), Structural Hamming Distance (SHD), and their standard deviations in parenthesis, for different choices of α based on 50 replications.

Graph	α	FPR	FDR	TPR	SHD
Random	0.100	0.00(0.00)	0.12(0.08)	0.95(0.05)	2.4(1.7)
	0.050	0.00(0.00)	0.13(0.07)	0.96(0.04)	2.4(1.5)
	0.025	0.00(0.00)	0.12(0.06)	0.93(0.04)	2.6(1.2)
	0.010	0.00(0.00)	0.13(0.07)	0.92(0.07)	3.0(1.6)
Hub	0.100	0.00(0.00)	0.08(0.04)	0.91(0.04)	5.0(2.5)
	0.050	0.00(0.00)	0.05(0.04)	0.95(0.03)	3.0(2.0)
	0.025	0.00(0.00)	0.06(0.06)	0.87(0.10)	5.3(4.6)
	0.010	0.00(0.00)	0.03(0.02)	0.97(0.02)	1.8(1.5)

NOTE: A smaller value of FPR, FDR, and SHD indicates higher accuracy, whereas a larger value of TPR means higher accuracy. Here, $p = 30$ and $n = 500$.

5.2. Real Data Analysis

This section applies DeFuSE to reconstruct gene regulatory networks for the Alzheimer's Disease Neuroimaging Initiative (ADNI) data. In particular, we construct two gene networks respectively for Alzheimer's Disease (AD) and healthy subjects to highlight some gene–gene interactions differentiating patients with AD/cognitive impairments and healthy individuals.

The ADNI dataset (*http://adni.loni.usc.edu/*) includes gene expressions, whole-genome sequencing, and phenotypic data. After cleaning and merging, we obtain a sample of 712 subjects in four groups, Alzheimer's Disease (AD), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), and Cognitive Normal (CN). For our purpose, we treat 247 CN individuals as controls while the remaining 465 individuals as cases (AD-MCI). Previous studies suggest that the amyloid precursor protein, the presenilin proteins, and the tau protein may involve in AD (O'brien and Wong 2011; Kelleher III and Shen 2017; Palmqvist et al. 2020), so we focus on the metabolic pathways of these proteins. Specifically, we extract the reference pathways in *https://genome.jp/pathway/map05010* from the KEGG database (Kanehisa and Goto 2000), including $p = 20$ genes in the data.

For data analysis, we first regress the gene expression levels on five covariates, Gender, Handedness, Education level, Age, and Intracranial volume, then use the residuals as gene expressions

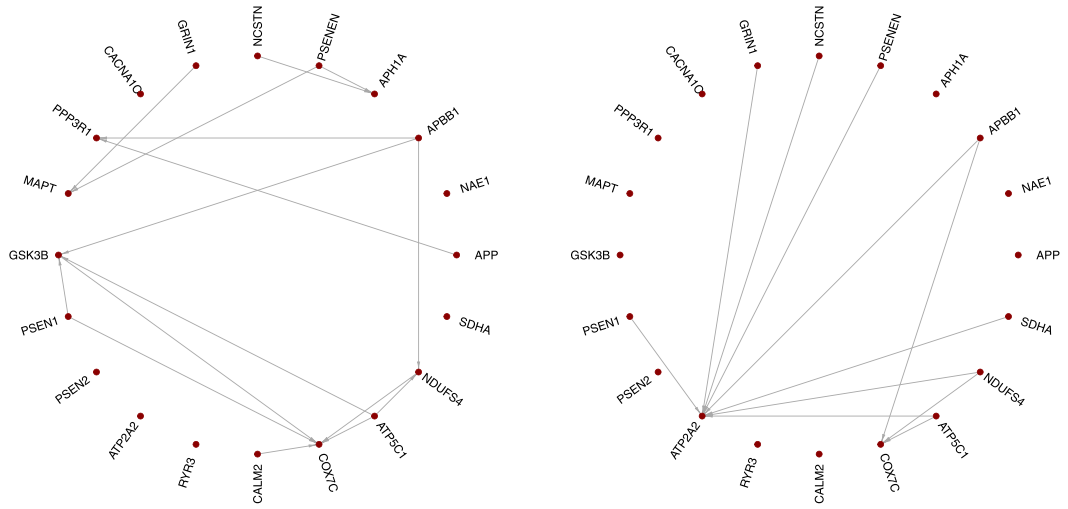


Figure 3. Reconstructed directed acyclic graphs for (a) AD-MCI and (b) CN groups.

Table 3. The AIC values for quadratic and linear models fitted for each non-root gene, as defined in (9).

Group	AD-MCI						CN	
Gene name	APH1A	PPP3R1	MAPT	GSK3B	COX7C	NDUFS4	ATP2A2	COX7C
Quadratic	0.717	0.656	0.528	0.620	0.356	0.606	0.572	0.304
Linear	0.701	0.732	0.567	0.695	0.395	0.657	0.656	0.349

NOTE: A smaller AIC value indicates better model fitting.

in the subsequent analysis. We normalize all gene expression levels and use the same FNN structure for fitting as in the simulation study. The normality test is conducted at a significance level $\alpha = 0.05$.

As displayed in Figure 3, the reconstructed DAGs exhibit some common and distinctive characteristics for the AD-MCI and CN groups. In the AD-MCI group, (a) directed edges GRIN1 $\rightarrow$ MAPT and PSEN1 $\rightarrow$ GSK3B agree with the reference pathways of the tau protein; (b) genes {APH1A, PSENEN, NCSTN, PPP3R1, APBB1, APP} have more directed connections, corresponding to the amyloid precursor protein. So do genes {PSEN1, GSK3B} for the presenilin proteins. By comparison, the genes participating in the amyloid precursor protein and tau protein metabolism have fewer connections in the CN group (O'brien and Wong 2011; Palmqvist et al. 2020). This observation seems consistent with previous studies that both genes may be involved in AD. Moreover, there are six and two non-root genes, respectively for the AD-MCI and CN groups.

For model diagnostics, we check the nonlinearity assumption on the gene expression levels. To this end, we compare a linear and a quadratic regression model for each non-root gene in the AD-MCI and CN groups in terms of their AIC values (Akaike 1992). These models are fitted on the estimated parents of DeFuSE, and the quadratic model includes additional quadratic terms $(Y_k^2)_{k\in\widehat{\mathrm{PA}}(j)}$ as covariates. For a linear or a quadratic model m for a non-root variable Y_j, the AIC value is defined as

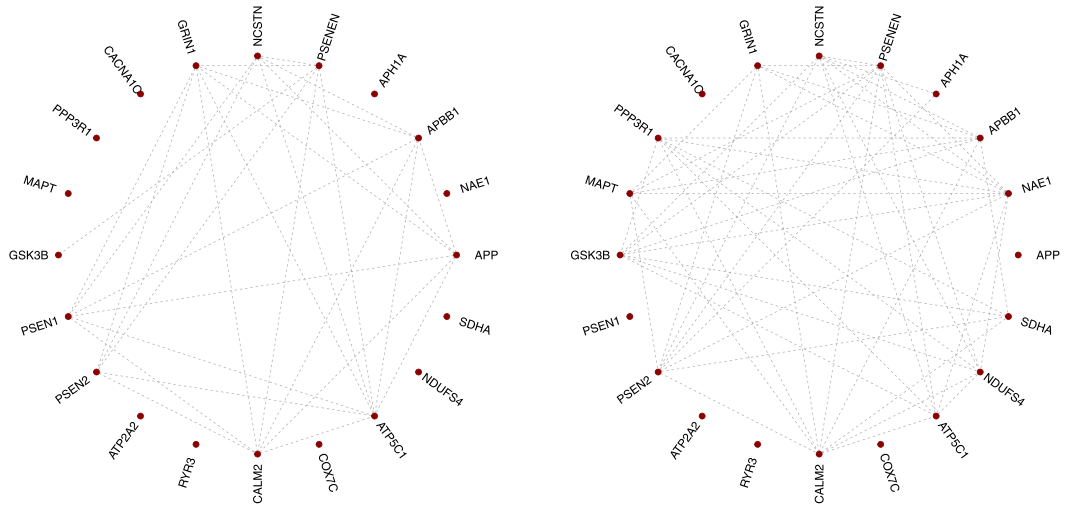
$$\mathrm{AIC}(\widehat{m}) = (n\widehat{\sigma}^2_{\mathrm{FNN}})^{-1}\sum_{i=1}^{n}(Y_j^{(i)} - \widehat{Y}_j^{(i)})^2 + 2n^{-1}\dim(\widehat{m}), \quad (9)$$

where $\widehat{m}$ and $\widehat{\sigma}^2_{\mathrm{FNN}}$ are the fitted model and the error variance estimated by FNN, $\widehat{Y}_j^{(i)}$ is the fitted values of $Y_j^{(i)}$, and $\dim(\widehat{m})$ denotes the number of parameters in model $\widehat{m}$. As suggested in Table 3, the quadratic model generally fits better than the corresponding linear model, as measured by AIC, suggesting that the nonlinearity assumption is approximately satisfied. Finally, the correlation plots of $\big(Y_j^{(i)} - \widehat{f}_j(Y^{(i)}_{\widehat{\mathrm{PA}}(j)})\big)_{j\in V}; i = 1, \ldots, n$ in Figure 4 exhibit the presence of (linear) hidden confounding as evident from the fact that many genes have multiple connections to other genes, indicating nonzero off-diagonals of Σ. This observation seems plausible due to the absence of some genes in the analysis.

6. Discussion

This article proposes a novel method for learning functional causal relations with additive confounders. For modeling, we establish identifiability under a sublinear growth condition on the functional relationships. On this basis, we propose a novel method called DeFuSE and implement it with feedforward neural networks for scalability. Theoretically, we show that the proposed method consistently reconstructs all nonlinear causal relations.

One central message is that nonlinearity permits the separation of the nonlinear causal relationships from the confounding effects in model (1) with observational data only. As nonlinear causal discovery with hidden confounding remains understudied, we hope the work could inspire further research in this direction.



(a) AD-MCI (b) CN

Figure 4. Undirected graph displaying the estimated residual correlations of $\widehat{\varepsilon} = \left(Y_j - \widehat{f}_j(Y_{\widehat{\mathrm{pa}}(j)})\right)_{j\in V}$, where a connection between two genes indicates the absolute value of residual correlation exceeds 0.15. Edge connections from one gene to other multiple genes suggest the presence of confounders or nonzero off-diagonal elements of the covariance matrix Σ.

Appendix

A.1. Additional Results on Identifiability

If $\Sigma \in \Psi$, the sublinear growth condition (Condition 1) is sufficient for identifying both $\{f_j\}_{1\le j\le p}$ and $\{\mathrm{PA}(j)\}_{1\le j\le p}$ in (1). When this condition is not satisfied, it is still possible to establish identifiability under an alternative assumption. Now, we consider model (2) with additive functions,

$$Y_j = \sum_{k\in \mathrm{PA}(j)} f_{j,k}(Y_k) + \varepsilon_j, \quad j \in V = \{1,\dots,p\}, \tag{A1}$$

where $\{f_{j,k}\}$ are nonlinear and $\varepsilon \sim N(0,\Sigma)$. Theorem A1 establishes the identifiability of $\{\mathrm{PA}(j)\}_{1\le j\le p}$ in (A1), without the sublinear growth condition.

Theorem A1. In (A1), assume that $Y_j - \mathrm{E}\left(Y_j \mid Y_{V(d)}\right)$ is not normally distributed for $d_j > d; 0 \le d \le d_{\max}$. For any univariate function f, we define its equivalence class

$$[f] = \{\widetilde{f} : \widetilde{f}(z) = f(z) + \gamma z, \gamma \in \mathbb{R}\}.$$

If

$$[f_{j,k}] \ne \sum_{j'\in V(d_j)} \gamma_{j'}[f_{j',k}] \quad \text{for all } \gamma_{j'} \in \mathbb{R}; j' \in V(d_j), j \in V = \{1,\dots,p\},$$

then $\{\mathrm{PA}(j)\}_{1\le j\le p}$ are uniquely identifiable.

The assumption that $Y_j - \mathrm{E}\left(Y_j \mid Y_{V(d)}\right)$ is not normal for $d_j > d$ imposes constraints on the compositions of nonlinear functions, which is automatically satisfied by sublinear growth functions when $\Sigma \in \Psi$ (Theorem 1). As suggested by the simulations in Section 5, DeFuSE continues to perform well in recovering the DAG even when Condition 1 and the additive function model (A1) are both violated.

A.2. Regularity Conditions

We impose the following regularity conditions to establish the consistency of DeFuSE.

Metric entropy. We define the bracketing L_2-metric entropy as a complexity measure of function spaces $\mathcal{G}_j = \left\{g_j : g_j(\cdot,\star) = f_j(\cdot) + \left\langle \star, \beta_j \right\rangle\right\}; j = 1,\dots,p$, where $\cdot$ and $\star$ represent a $|V(d_j)|$-dimensional vector, respectively. The bracketing L_2-metric entropy of $\mathcal{G}_j$ is the logarithm of the smallest u-bracket cardinality, $H(u,\mathcal{G}_j) = \log(\min\{m : \mathcal{S}(u,m)\})$, where a u-bracket $\mathcal{S}(u,m) = \{g_1^-, g_1^+, \dots, g_m^-, g_m^+\} \subseteq L_2(P)$ is a set of functions such that (i) $\max_{1\le k\le m} \|g_k^- - g_k^+\|_{L_2} \le u$ and (ii) for any $g \in \mathcal{G}_j$ there exists $g_k^- \le g \le g_k^+$ almost surely.

Condition A1. For some positive $\epsilon_n < 1/2$,

$$\max_{1\le j\le p} \max_{\{A:|A|\le|\mathrm{PA}^\circ(j)|\}} \int_{\epsilon_n^2/256}^{\sqrt{2}\epsilon_n} H^{1/2}(u/c_1, \mathcal{G}_j(A))du \le c_2\sqrt{n}\epsilon_n^2,$$

where $\mathcal{G}_j(A) = \left\{g_j \in \mathcal{F}_j : A = \mathrm{ARG}(f_j),\ \|g_j - g_j^\circ\|_2 \le 2\epsilon_n\right\}$ is the $2\epsilon_n$-neighborhood of g_j° on the index set of effective arguments A.

In view of Condition A1, the error rate ϵ_n is determined by solving the integral equation in ϵ_n. Such a condition has been used to quantify the convergence rate of sieve estimates (Wong and Shen 1995; van de Geer 2000). The entropy results are available for many function classes, such as the FNN in Theorem 4.

Sparsity and confounding. Next, we impose a regularity condition on sparsity and confounding structures, requiring the true support of g_j°, the maximum depth $d_{\max}$, and the error variance not to increase with the sample and graph sizes (n,p).

Condition A2. Assume $\kappa^\circ = \max_{1\le j\le p} |\mathrm{PA}^\circ(j)|$, $\varsigma^\circ = \max_{1\le j\le p} \|\beta_j^\circ\|_0$, $d_{\max} = \max_{1\le j\le p} d_j$, and $c_- \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le c_+$ are independent of (p,n), where $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ are the smallest and largest eigenvalues of $\Sigma \in \Psi$.

A.3. Implementation Details

The code is open-sourced at *https://github.com/chunlinli/defuse*.

Training and hyperparameter tuning for DeFuSE. Training and tuning a neural network requires intensive computation. Following the conventional practice of deep learning, we split the original sample into training and validation sets with a partition ratio 9:1, and use on-the-fly evaluation over the validation set for tuning during the training process.

To tune hyperparameters κ_j, ς_j in (7), we adopt a greedy strategy combined with an asynchronous-synchronous training technique since it is unnecessary to identify the exact value of ς_j, see, Theorem 3. We first optimize (7) in β_j with $\theta_j = 0$, subject to the sparsity constraint $\sum_{k\in V(d)} \min(|\beta_{j,k}|/\tau, 1) \le \varsigma_j$, followed by selecting $\varsigma_j \in \{0, 1, \ldots, |V(d)|\}$ that minimizes the mean squared error on the validation set. Throughout, we fix $\tau = 0.05$ as a signal-noise threshold. This stage intends to perform a sparsity-constrained linear regression, so it is very efficient in computing. Next, given the selected variable set $B = \{k : |\beta_{jk}| \ge \tau\}$ in (7), we estimate $(\theta_j, \beta_{j,B})$ with $\beta_{j,B^c} = 0$ by minimizing

$$\min_{\theta_j} \quad \sum_{i=1}^{n} \left(Y_j^{(i)} - f_j\big(Y_{V(d)}^{(i)}\big) - \big\langle \widehat{\xi}_{V(d)}^{(i)}, \beta_{j,B} \big\rangle \right)^2,$$

$$\text{s.t.} \quad \sum_{k\in V(d)} \min(\|W_k^1\|/\tau, 1) \le \kappa_j.$$

To leverage the automatic differentiation in modern deep learning libraries, we consider its regularized version with κ_j replaced by a hyperparameter $\lambda_j > 0$:

$$\min_{\theta_j} \quad \sum_{i=1}^{n} \left(Y_j^{(i)} - f_j\big(Y_{V(d)}^{(i)}\big) - \big\langle \widehat{\xi}_{V(d)}^{(i)}, \beta_{j,B} \big\rangle \right)^2 + \lambda_j \sum_{k\in V(d)} \min(\|W_k^1\|/\tau, 1).$$

where $\lambda_j > 0$ controls the degree of regularization. Then, after the regularized optimization is completed, we tune $\kappa_j \in \{0, 1, \ldots, |V(d)|\}$ using the top κ_j variables (sorted by weight $\|W_k^1\|$) among all variables and masking the rest.

In our experiments, we use an adaptive regularization approach for $\lambda_j > 0$ during training, similar to adaptive learning rate scheduling. Specifically, we consider three candidate values $\lambda_j \in \{0.0001, 0.001, 0.05\}$. The training process starts with $\lambda_j = 0.0001$ and gradually increases λ to achieve better validation performance by inducing more sparsity. Based on our limited experience, this adaptive regularization strategy is effective and can be combined with other deep learning techniques such as early stopping.

For network structure, we use an FNN with one hidden layer and 50 hidden neurons. For optimization, we use the Adam optimizer (Kingma and Ba 2014) with a learning rate 0.1 and various numbers of epochs $\{250, 500, \ldots, 4000\}$ in our experiments. Then we choose the best-performing model.

Other methods. The R packages `CAM`, `pcalg`, and `lrpsadmm` are available at *https://github.com/cran/CAM*, *https://github.com/cran/pcalg*, and *https://github.com/benjaminfrot/lrpsadmm*, respectively. The Python program `notears` is available at *https://github.com/xunzheng/notears*. We use their default settings for CAM, NPTEARS, LRpS-GES, and RFCI.

Supplementary Materials

The supplementary materials provide technical proofs of theorems.

Acknowledgments

The authors would like to thank the editor, the associate editor, and the anonymous referee for their helpful comments and suggestions.

Funding

The research is supported in part by NSF grant DMS-1952539, NIH grants R01GM113250, R01GM126002, R01AG065636, R01AG074858, R01AG069895, U01AG073079.

ORCID

Chunlin Li https://orcid.org/0000-0003-2989-8785
Xiaotong Shen https://orcid.org/0000-0003-1300-1451
Wei Pan https://orcid.org/0000-0002-1159-0582

References

Agrawal, R., Squires, C., Prasad, N., and Uhler, C. (2021), "The DeCAM-Founder: Non-linear Causal Discovery in the Presence of Hidden Variables," arXiv preprint arXiv:2102.07921. [2]

Akaike, H. (1992), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Breakthroughs in Statistics*, eds. S. Kotz and N. L. Johnson, pp. 610–624, New York: Springer. [7]

Anderson, T., and Darling, D. (1952), "Asymptotic Theory of Certain "Goodness of Fit" Criteria based on Stochastic Processes," *The Annals of Mathematical Statistics*, 23, 193–212. [3,6]

Bühlmann, P., Peters, J., and Ernest, J. (2014), "CAM: Causal Additive Models, High-Dimensional Order Search and Penalized Regression," *The Annals of Statistics*, 42, 2526–2556. [1,5]

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012), "Latent Variable Graphical Model Selection via Convex Optimization," *The Annals of Statistics*, 40, 1935–1967. [1]

Chen, C., Ren, M., Zhang, M., and Zhang, D. (2018), "A Two-Stage Penalized Least Squares Method for Constructing Large Systems of Structural Equations," *Journal of Machine Learning Research*, 19, 40–73. [1]

Chickering, D. M. (2002), "Optimal Structure Identification with Greedy Search," *Journal of Machine Learning Research*, 3, 507–554. [1]

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012), "Learning High-Dimensional Directed Acyclic Graphs with Latent and Selection Variables," *The Annals of Statistics*, 40, 294–321. [1,5]

de Campos, C. P., and Ji, Q. (2011), "Efficient Structure Learning of Bayesian Networks using Constraints," *Journal of Machine Learning Research*, 12, 663–689. [1]

de Campos, L. M. (2006), "A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests," *Journal of Machine Learning Research*, 7, 2149–2187. [1]

Farrell, M. H., Liang, T., and Misra, S. (2021), "Deep Neural Networks for Estimation and Inference," *Econometrica*, 89, 181–213. [5]

Frot, B., Nandy, P., and Maathuis, M. H. (2019), "Robust Causal Structure Learning with Some Hidden Variables," *Journal of the Royal Statistical Society*, Series B, 81, 459–487. [1,3,5]

Glymour, C., Zhang, K., and Spirtes, P. (2019), "Review of Causal Discovery Methods based on Graphical Models," *Frontiers in Genetics*, 10. [1]

Gu, J., Fu, F., and Zhou, Q. (2019), "Penalized Estimation of Directed Acyclic Graphs from Discrete Data," *Statistics and Computing*, 29, 161–176. [1]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer. [2]

Hoyer, P. O., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. (2008), "Nonlinear Causal Discovery with Additive Noise Models," in *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pp. 689–696. [2,3]

Jaakkola, T., Sontag, D., Globerson, A., and Meila, M. (2010), "Learning Bayesian Network Structure using LP Relaxations," in *International Conference on Artificial Intelligence and Statistics*, pp. 358–365. PMLR. [1]

Janzing, D., Peters, J., Mooij, J., and Schölkopf, B. (2009), "Identifying Confounders using Additive Noise Models," in *Conference on Uncertainty in Artificial Intelligence*, pp. 249–257. [2]

Kalisch, M., and Bühlman, P. (2007), "Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm," *Journal of Machine Learning Research*, 8. [3]

Kanehisa, M., and Goto, S. (2000), "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, 28, 27–30. [6]

Kelleher III, R. J., and Shen, J. (2017), "Presenilin-1 Mutations and Alzheimer's Disease," *Proceedings of the National Academy of Sciences*, 114, 629–631. [6]

Kingma, D. P., and Ba, J. (2014), "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980. [9]

Li, C., Shen, X., and Pan, W. (2020), "Likelihood Ratio Tests for a Large Directed Acyclic Graph," *Journal of the American Statistical Association*, 115, 1304–1319. [1]

Li, C., Shen, X., and Pan, W. (2021), "Inference for a Large Directed Acyclic Graph with Unspecified Interventions," arXiv preprint arXiv:2110.03805. [1]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection with the Lasso," *The Annals of Statistics*, 34, 1436–1462. [2,5]

Monti, R. P., Zhang, K., and Hyvärinen, A. (2020), "Causal Discovery with General Non-linear Relationships using Non-linear ICA," in *Conference on Uncertainty in Artificial Intelligence*, pp. 186–195. PMLR. [1]

Mooij, J., Janzing, D., Peters, J., and Schölkopf, B. (2009), "Regression by Dependence Minimization and its Application to Causal Inference in Additive Noise Models," in *International Conference on Machine Learning*, pp. 745–752. [1]

O'brien, R. J., and Wong, P. C. (2011), "Amyloid Precursor Protein Processing and Alzheimer's Disease," *Annual Review of Neuroscience*, 34, 185–204. [6,7]

Palmqvist, S., Janelidze, S., Quiroz, Y., Zetterberg, H., Lopera, F., Stomrud, E., Su, Y., Chen, Y., Serrano, G., Leuzy, A., et al. (2020), "Discriminative Accuracy of Plasma Phospho-tau217 for Alzheimer Disease vs Other Neurodegenerative Disorders," *JAMA*, 324, 772–781. [6,7]

Pearl, J. (2009), *Causality*, Cambridge: Cambridge University Press. [2]

Peters, J., and Bühlmann, P. (2014), "Identifiability of Gaussian Structural Equation Models with Equal Error Variances," *Biometrika*, 101, 219–228. [2]

Peters, J., Janzing, D., and Scholkopf, B. (2017), *Elements of Causal Inference*, Cambridge, MA: MIT Press. [1]

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014), "Causal Discovery with Continuous Additive Noise Models," *Journal of Machine Learning Research*, 15, 2009–2053. [2,3]

Reisach, A., Seiler, C., and Weichwald, S. (2021), "Beware of the Simulated DAG! Causal Discovery Benchmarks may be Easy to Game," in *Advances in Neural Information Processing Systems* (Vol. 34), pp. 27772–27784. [6]

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005), "Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data," *Science*, 308, 523–529. [1]

Schmidt-Hieber, J. (2019), "Deep ReLU Network Approximation of Functions on a Manifold," arXiv preprint arXiv:1908.00695. [5]

Shah, R. D., Frot, B., Thanei, G.-A., and Meinshausen, N. (2020), "Right Singular Vector Projection Graphs: Fast High Dimensional Covariance Matrix Estimation under Latent Confounding," *Journal of the Royal Statistical Society*, Series B, 82, 361–389. [1]

Shen, X., Pan, W., and Zhu, Y. (2012), "Likelihood-based Selection and Sharp Parameter Estimation," *Journal of the American Statistical Association*, 107, 223–232. [4]

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, 2003–2030. [1]

Spirtes, P., Glymour, C., and Scheines, R. (2000), *Causation, Prediction, and Search*, Cambridge, MA: MIT Press. [1,2]

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053. [5]

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006), "The max-min Hill-Climbing Bayesian Network Structure Learning Algorithm," *Machine Learning*, 65, 31–78. [1]

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. (2013), "Geometry of the Faithfulness Assumption in Causal Inference," *The Annals of Statistics*, 41, 436–463. [2,5]

van de Geer, S. A. (2000), *Empirical Processes in M-Estimation* (Vol. 6), Cambridge: Cambridge University Press. [8]

Voorman, A., Shojaie, A., and Witten, D. (2014), "Graph Estimation with Joint Additive Models," *Biometrika*, 101, 85–101. [1]

Wang, Y., and Blei, D. M. (2019), "The Blessings of Multiple Causes," *Journal of the American Statistical Association*, 114, 1574–1596. [1]

Wong, W. H., and Shen, X. (1995), "Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES," *The Annals of Statistics*, 23, 339–362. [8]

Yuan, Y., Shen, X., Pan, W., and Wang, Z. (2019), "Constrained Likelihood for Reconstructing a Directed Acyclic Gaussian Graph," *Biometrika*, 106, 109–125. [1]

Zhang, K., and Hyvärinen, A. (2009), "On the Identifiability of the Post-nonlinear Causal Model," in *Conference on Uncertainty in Artificial Intelligence*, pp. 647–655. [1]

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. (2018), "DAGs with NO TEARS: Continuous Optimization for Structure Learning," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 9492–9503. [1]

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. (2020), "Learning Sparse Nonparametric DAGs," in *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425, PMLR. [1,5,6]