Neural Causal Discovery and Social Science Research

Joshua Pierce

January 13, 2023

Abstract

This review explores the use of deep learning in causal discovery research, with a particular focus on how these methods may advance applications within the social science domain. Over the course of the review we find that deep learning is advancing causal discovery techniques by replacing the explicit priors and assumptions of earlier methods with the data-driven representational capacity of neural networks. This is especially relevant within the context of social science research, where complex and messy data with unknown ground truth causal structures pose a challenge for accurately navigating any large hyperparameter space and strict assumptions associated with other causal discovery methods. The advantage is most evident in recent supervised causal discovery techniques which employ transformer architectures to learn a mapping between synthesized training data and underlying causal structures. One drawback of this approach is the corresponding dependence on training data with ground truth causal structures, which is not readily available in the social science domain. We suggest that the development of a more complicated, realistic benchmark library for causal discovery research would be a promising avenue for advancing the current SOTA towards real-world applications.

1 Introduction

This review is motivated by an observation that recent progress in machine learning, such as the proliferation of neural networks, does not readily translate to breakthroughs in social science methodology. There seem to be barriers between implementations of the powerful functional representations of neural networks in fields such as computer vision and natural language processing, and the capacity to leverage our data-rich environment towards new understanding and policy mechanisms in fields such as economics and sociology.

This may be partially attributed to a difference in ambition: while many of the most striking examples of recent machine learning progress entails neural networks learning complex correlations from a large data distribution for predictive purposes, a lot of social science research is more interested in studying how these observed distributions would change at a causal level[41]. Social policy researchers often seek to estimate the impact of a policy intervention, and economists seek to elucidate the driving mechanisms behind market behavior. In comparison to many prediction-driven machine learning applications which may model the single joint distribution in the observed data, one can consider the distinction of causal learning as trying to model the data-generating process behind the observed distribution (ie, estimating different distributions for each causal intervention, see figure 1).

This distinction in ambition is embodied by causality research, which helps clarify the gap between inferential applications of machine learning techniques, such as neural networks, and the knowledge-driven focus of social science research. This motivates the paper’s focus on how advancements in deep learning methods are affecting state-of-the-art in causal discovery, with a particular attention to the challenges and barriers that hinder real-world applications in a social sciences environment.

Over the course of this review we consider the initial ‘unsupervised learning’ approaches to causal discovery, finding that the stricter assumptions and priors explicitly represented by the large hyperparameter space proves problematic for many practical settings. We then evaluate recent ‘supervised learning’ methods of neural causal discovery, which in contrast leverages neural networks to allow for a much more flexible mapping between observational data and causal structures. However, the ‘supervised’ nature of these methods are very dependent on the synthesized training data to provide ground truth causal structures.
Ultimately, we find that there is a role for deep learning to facilitate causal discovery in the social sciences by easing the strict priors and large hyperparameter space of previous unsupervised learning approaches. However, inherent lack of ‘ground truth’ causal structures within the social sciences hinders training and validation of these supervised learning methods and remains a substantial hurdle for developing and demonstrating sufficient robustness for real-world applications. We suggest that a promising avenue for progress would be the development of a causal discovery benchmark library consisting of more nuanced and naturalistic synthesized data. We also discuss the potential utility of alternative performance metrics for causal discovery that do not rely on a ground truth causal graph, such as predicting distributions that result from unobserved interventions.

This paper covers the following outline:

• A review of Causal Discovery (DAGs and SCM framework)
• Unsupervised Causal Discovery
• Supervised Causal Discovery
• Discussion

2 Review of DAGs and SCM framework

Rigorous consideration of causality plays a fundamental role in social science research, and has motivated efforts in the field to discern causal hypotheses from observational data[1], [8], [15], [28]. Pearl et al[15] offers the ‘Ladder of Causality’ as a helpful framework for causal research (figure 2). Each rung of the ladder represents a further step of causal understanding. The bottom rung, ‘association’, represents observing and predicting from a given data source. This would include many machine learning techniques that learn inference from correlations observed in the data. The second rung, ‘intervention’, is a further step in causal learning that allows one to estimate the effects of actions that may shift the underlying distribution. This is particularly relevant for social policy research. The highest rung is ‘counterfactuals’, which focuses on estimating a specific, unobservable alternative provided a given change (eg, ‘What would the health outcome have been if patient X was instead given the placebo?’). Counterfactuals may be particularly relevant to considerations of fairness in a social policy setting;

Figure 1: Causal vs statistical models: causal models can be seen as estimating distinct distributions for each causal variable, while predictive statistical models capture the joint distribution of the observed data. The causal model is inherently a set of assumptions about the generating process which cannot be expressed entirely from the properties of the observational data.[52]
Figure 2: Pearl's 'Ladder of Causality'. From figure 1 we can see that the causal models representing multiple distributions are not directly evident from the single observed distribution, instead relying on assumptions of the data-generating process. In other words, going from first rung to the second in the ladder of causality requires additional assumptions, or 'priors', applied to the observed distribution.
Figure 3: Note that a given DAG represents assumptions of a data-generating process rather than a single observed distribution, and can accordingly represents estimations of how multiple distributions will vary as result of changes in the node values. **DAG Factorization:**

\[
P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | X_{pa(i, G)})
\]

for example, would the hiring outcome for this candidate have been different if they were a different gender or race?

Pearl’s research introduced a helpful mathematical framework for considering causal relationships: Directed Acyclic Graphs (DAGs) and Structural Causal Models (SCM)[15]. DAGs offer a graphical representation of causal structures— that is, they encode the data-generating process that underlies resulting distributions of data. Each node represents a variable, a directed edge between nodes represents a causal relationship between the two variables, and the absences of edges suggest conditional independencies in the probability distribution of the variables. These conditional independencies provide for the corresponding statistical model of a DAG via DAG factorization: see figure 3. The *d-separation* criteria[2][15] identifies the variables that must be conditioned on to sever the causal path between any two nodes, and the *global Markov property* holds that d-separation in a DAG corresponds to conditional independence in the probability distribution: \((X \text{ [d-sep] } Y | Z) \Rightarrow (X \perp \perp Y | Z)\).

### 2.1 Causal Discovery

Ascertaining the underlying causal structure is a critical and challenging first step for subsequent downstream causal analysis, and often relies on domain knowledge as well as inference from observable data. This process of inferring a causal structure from observed data is called "Causal Discovery", and represents a primary challenge for applying causal learning techniques in real-world settings where the causal structure is unknown. Causal discovery can be described as a search across the space of potential DAGs for the correct causal structure model. Traditional efforts in this area can be broadly categorized into three groups:

- **Score-based searches** through all possible DAGs and ranks them based on some measure of data-fit-scoring-functions[9], [12], [17], [21], [23], [31], [33], [43]. Recently this approach has also been restructured as a continuous optimization problem[35], [38], [42] and a mix of continuous and discrete optimization. [36], [47], [49], [51]

- **Constraint-based approach** relies on estimated conditional independence tests between all variables, and uses results to winnow down viable DAGs that comply with the non-independence definition of causal relationships. [8], [13], [20], [40], [43]

- **The restricted structural equation modelling approach** makes assumptions about the functional relationships of variables (e.g., linear with non-gaussian error), and uses a corresponding asymmetry between cause and effect variables to infer the direction of causal relationships. [11], [14], [18], [19], [30], [34]

These approaches naturally face a challenge of scale: the search space of potential DAGs increases prohibitively with an increase in number of nodes. This paper reviews: i) how recent methods have
Figure 4: Note that the number of potential DAGs increases super-exponentially with the number of nodes, making discrete searches prohibitively costly for higher-dimensional distributions addressed this issue by reframing the DAG search as a continuous optimization problem; and ii) how this reframing facilitates utilizing deep learning approaches for a more flexible causal discovery method.

(See figure 4)

3 Unsupervised Causal Discovery

Recent causal discovery research has sought to reframe the challenging search task to an optimization problem that utilizes gradient descent. These ‘learning’ methods of causal discovery can be broadly categorized into ‘unsupervised’ and ‘supervised’. The ‘unsupervised’ approach entails defining metrics for causal graphs that provide a smooth, continuous space where distance meaningfully corresponds to a gap in preferential causal structures. These metrics can be combined and incorporated into a constrained ‘loss function’ that allows for gradient descent towards an estimated causal structure.

The first paper to propose a transformation of the causal discovery task to a continuous optimization problem was DAGs with NO TEARS[35], by Zheng et al. Zheng’s approach starts with a score-based search formulation, where a scoring function is defined to guide the discrete search across potential DAGs. For this approach, consider $X$ as a $n \times d$ matrix of values corresponding to $d$ random variables and $n$ independent samples. Now define a $d \times d$ adjacency matrix $A$ where $a_{ij} = 1$ indicates that $X_i$ is a causal parent of $X_j$, and the corresponding DAG would have a directed edge from variable $X_j$ to $X_i$. Assuming linearity, we can write the corresponding SEM as $X = XW + Z$, where $W \in \mathbb{R}^{d \times d}$ is a weighted graph such that $w_{ij} \neq 0 \Leftrightarrow a_{ij} = 1$, and $Z$ are random noise variables. If there was no noise in the system, then we can define the corresponding constraint equation on $W$ as $X - XW = 0$. From this constraint equation we have one term of the scoring function- a least squares loss metric calculated with the Frobenius norm:

$$\ell(W; X) = \frac{1}{2n} \|X - XW\|_F^2$$

Applying such a loss to scoring DAGs has been demonstrated to recover the Markov Equivalence Class with high probability on finite-samples and in high-dimensions, with demonstrated consistency for Gaussian[22], [27] and non-Gaussian SEMs[25].

Because the goal is to learn a sparse DAG, the authors add a $\ell_1$-regularization term to the weighted adjacency matrix: $\|W\|_1$. These two terms then provide the regularized score function:

$$F(W) = \frac{1}{2n} \|X - XW\|_F^2 + \lambda \|W\|_1$$

And similar to other score-based search approaches, the goal is to minimize the scoring function $F$ across the space of DAGs: $\min_{W \in \mathbb{R}^{d \times d}} F(W)$, subject to $G(W) \in \mathbb{D}$. Note that although the

<table>
<thead>
<tr>
<th># of Nodes</th>
<th># of DAGs</th>
<th># of Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>$2^2$</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>$2^6$</td>
</tr>
<tr>
<td>4</td>
<td>543</td>
<td>$2^{12}$</td>
</tr>
<tr>
<td>5</td>
<td>29281</td>
<td>$2^{20}$</td>
</tr>
<tr>
<td>6</td>
<td>3781503</td>
<td>$2^{30}$</td>
</tr>
<tr>
<td>7</td>
<td>113879265</td>
<td>$2^{42}$</td>
</tr>
<tr>
<td>8</td>
<td>783702329343</td>
<td>$2^{56}$</td>
</tr>
<tr>
<td>9</td>
<td>1213442454842881</td>
<td>$2^{72}$</td>
</tr>
<tr>
<td>10</td>
<td>4175098976430598143</td>
<td>$2^{90}$</td>
</tr>
</tbody>
</table>
scoring function $F$ is continuous, the search task is defined over the discrete space of DAGs. This is the formulation adopted by other score-based search approaches that vary in their scoring functions: BDe(u)[23], BGe[24], BIC[6], and MDL[5]. However the combinatorial nature of the discrete search space is NP-hard to solve[7]. The authors’ innovation is the definition of an additional ‘DAGness’ metric: a smooth characterization of DAGs capturing how severe the violations of acyclicity are as the weighted adjacency matrix moves away from a DAG representation.

Thus the problem transforms from a discrete search to a constrained optimization task: $\min_{W \in \mathbb{R}^{d \times d}} F(W)$, subject to $h(W) = 0$, where $h : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ is the new measure of ‘DAGness’ with the following attributes:

i) $h(W) = 0 \Leftrightarrow G(W) \in \mathbb{D}$ (ie, $h = 0$ iff $W$ is acyclic)

ii) Distances within $h$ correspond to degrees of acyclicility, or ‘DAGness’)

iii) $h$ is smooth

iv) $h$ is differentiable

The derivation of the ‘DAGness’ metric arrives from the observation that given a binary adjacency matrix $B$ of dimension $d$ with zero diagonal elements (ie, no nodes cycling onto themselves), then the value of the $i$th element of $B^k$ is equal to the count of all paths of length $k$ that loop back to node $i$. As a simple demonstration one can consider a 2-node binary adjacency matrix and count all of the two-length cycles on the first node. In such an example, $(B^2)_{11} = b^2_{11} + b_{12} * b_{21} = 0 + b_{12} * b_{21}$, which equals one if and only if $B$ includes the only possible two-length cycle on the first node (ie, $b_{12} = 1$ and $b_{21} = 1$).

Hence, from the trace of a binary adjacency matrix exponentiated to $k$ one can count all of the length-$k$ cycles: $(TrB^k)$. So an infinite sum over all lengths $k$ provides a measure of acyclicility, or ‘DAGness’, in binary adjacency matrix $B$. This is represented by the matrix exponential:

$$Tr e^B = Tr(\sum_{k=0}^{\infty} \frac{B^k}{k!}) = \sum_{k=0}^{\infty} \frac{1}{k!} Tr B^k$$

. However, note that $Tr(B^0) = Tr(I) = d$ regardless of the value of $B$. Therefore, the condition of acyclicility for adjacency matrix $B$ of dimension $d$ would be:

$$Tr e^B - d = 0$$

Note that the weighted adjacency matrix $W$ can have negative values, which is problematic for this metric as negative paths would cancel positive ones. But because we don’t need an exact count of cycles so much as a metric that corresponds to DAGness when constrained to zero, and the actual DAG is defined only by the presence of zeros in the weighted adjacency matrix: then we can instead apply the same approach after squaring each element of $W$ (ie, the weighted adjacency matrix’s Hadamard product with itself). Therefore, the resulting ‘DAGness’ metric is defined as:

$h(W) = Tr e^{W \odot W} - d$

Furthermore, the result is both smooth and differentiable ($\nabla h(W) = (e^{W \odot W})^T \odot 2W$).

Rather than employing search methods over the discrete space of DAGs, the authors can now apply an augmented Lagrangian optimization to the scoring function $F$ under the non-convex constraint of $h(W) = 0$. In practice this entails alternating between optimizing the scoring function and updating the ‘DAGness’ constraint, and finally thresholding $W$ to enforce acyclicility.

The authors generated synthesized data to test their approach, generating random graphs using an Erdos-Renyi or scale-free models and sampling edges to construct a $W$ weighted adjacency matrix. From there, they would then use the underlying $W$ to generate data: $X = W^TX + z$, where $z$ was noise sampled from several different distributions: Gaussian, Exponential and Gumbel. The authors demonstrated that their optimization approach does in fact estimate DAGs and offers improved performance over several search based algorithms: FGS[29], PC[8], LiNGAM[11].

Subsequent researchers followed this unsupervised learning approach to causal discovery, with several papers developing different smooth metrics for a "DAGness" constraint. In DAG-GNN[42], Yu uses the new optimization framework to replace the SEM-based scoring function in Zheng’s paper with a more flexible graph neural network approach. Additionally, Yu proposes a polynomial alternative...
to the 'DAGness' constraint that better fits the automatic differentiation framework of deep learning. Lee[39] addressed the computational cost of Zheng’s NO-TEARS approach, suggesting an alternative acyclicity constraint calculated from the spectral radius of the adjacency matrix, which enforces acyclic properties while scaling in $O(d^2)$ complexity as opposed to the $O(d^3)$ of NO-TEARS. They also adjusted the scoring function to address non-linearity, in this case by incorporating polynomial regression.

DAGness priors: $h(W) = 0$

$h(W) = \text{trace} [\exp(W \odot W)] - d; \ (Zheng[35])$

$h(W) = \text{trace} ((I + \alpha W \odot W)^d) - d; \ (Yu[42])$

$h(W) = p(W \odot W); \ (Lee[39])$

$\odot := \text{HadamardProduct}$

$p() := \text{SpectralRadius}(\max_i |\gamma_i|)$

Acyclicity priors: eg, avoiding cycles of length 2, so preventing $(i, j) = (j, i) = 1$ in the adjacency matrix

Sparsity priors: eg, L1 norm on adjacency matrix

Altogether, applying an unsupervised learning approach to causal discovery entails the selection of a scoring function, a constraint metric, a regularization penalty for sparseness, as well as their respective hyperparameters. Each of these selections represents assumptions of the underlying causal structure, and the weighted combination of these priors then presents a large hyperparameter space. Knowing the correct priors to associate with a given joint-distribution becomes a difficult search task in and of itself. By comparison, supervised learning techniques ease this explicit selection of priors, relying instead on training to learn a mapping between data and underlying causal dynamics.

### 4 Supervised Causal Discovery

A fundamental challenge for supervised learning methods are their reliance on ground truth causal structures at training time, data which is not readily available at a sufficient scale. The neural causal discovery method proposed by Ke[53], Causal Structure Induction with a Supervised Approach (CSIvA), addresses the lack of ground truth causal structures by synthesizing data from increasingly complex causal distributions. They use this synthesized dataset to train a transformer-based neural network architecture which maps observational data to causal graphs, thus transforming the search-task to an end-to-end, supervised optimization task driven by gradient descent.

This approach incorporates interventional data when estimating causal structures and the neural networks allow for non-linear dependencies. The authors first demonstrate that this causal discovery model can generalize from the training distribution of synthesized data to a test set of differing synthesized data. They then test the model on naturalistic data from several widely referenced Bayesian networks and demonstrate that their approach generalizes to this more natural environment as well.

Two preceding supervised causal discovery approaches included Lopez-Paz’s RCC[26] and Li’s DAG-EQ[45]. RCC relied on a kernel mean embedding classification, while DAG-EQ operated directly on covariance matrices. However DAG-EQ only modeled linear dependency, which is impractical for real-world data, and neither incorporated interventional data, which is important for sufficient identifiability of the causal structure.[15] (see figure 5)
The supervised learning approach of CSIvA\cite{53} takes as input observational and interventional data and estimates the underlying causal structure graph. (See figure 6.) The ground truth and estimated graphs take the form of an adjacency matrix. Deriving the training set entails a random sampling of graphs which vary over the number of nodes, number of edges, and each node’s conditional probability distribution. The resulting training set thus consists of data sampled from the joint probability distribution of each selected graph, with the underlying graph structure serving as the ground truth. Intervenational data is generated in this process by fixing the value of any given node.

4.1 Encoder-Decoder Transformers

The authors approach the challenge of learning the causal structure in a fashion similar to how sequence-to-sequence transformers are used to predict the next word in NLP applications. However in this case, the model is estimating the causal structure by predicting the next value of the adjacency matrix. Furthermore, the authors incorporate the approach suggested by Kossen\cite{48} for distributing self-attention between datapoints, so a single input to the transformer consists of an array of synthesized data which have a known ground truth causal structure. In broad terms, layers of the encoder alternate between distributing attention across nodes within a data sample (ie, the columns of the input data array), and across data samples within a node (ie, the rows of the input data array). The former is designed to encode relationships between multiple nodes, while the latter captures information about the distribution of data within a given node. The encoder results in a summary vector for each node, generated using the common weighted-average technique employed by transformers: in this instance the node embeddings are used as ‘queries’ and the embeddings across the data columns are used in the normal ‘key-value’ attention mechanism.

The decoder then uses the summary vector to autoregressively estimate the causal structure; in other words, sequentially estimate the binary values of the adjacency matrix where each estimation is conditioned on the previously estimated adjacency matrix elements. See figure 7 for details.

4.2 Data Synthesis

A critical element of this approach to supervised learning of causal structures is the synthesis of observational and interventional data. The generating causal structures are sampled from a distribution that varies across: number of nodes (5 to 80), number of edges (Erdos-Renyi metric of 1 or 2), and the conditional probability distributions at each node. Given:

\[ N \times S \] matrix \( X = S \) samples of a causal Bayesian network with \( N \) nodes;
Figure 7: This diagram presents the structure of the encoder transformer. One can see the alternating layers of attention, initially distributing attention across nodes and then across samples

\[ A = \text{weighted adjacency matrix; } \]
\[ Z = \text{matrix of random distribution of values } N(0,0.1); \]

Then for continuous nodes, three types of conditional probability distributions were generated:

- **Linear Data:** \( X_{n,:} = A_{n,:}X + Z_{n,:}, \) with biases initialized with \( U[-0.5,0.5] \), weights truncated to normal distribution with sd 1.5, nodes with interventions sampled from \( U[-1,1] \)
- **ANM Data** (nonlinear with additive noise models): \( X_{n,:} = F_{n,:}X + 0.4xZ_{n,:}, \) where \( F \) is a fully connected neural network with weights randomly initialized from \( N(0,1) \). Nodes and interventions sampled from \( N(2,1) \).
- **NN Data** (nonlinear and non-additive noise models using neural networks): \( X_{n,:} = F_{n,:}(X,Z_{n,:}) \) with \( F \) a fully connected network with same initialization and nodes/interventions sampling as above.

For discrete nodes, two types were generated:

- **MLP Data:** Two fully connected layers of hidden dimensionality 32 and randomly initialized network. Nodes with interventions/values randomly and independently sampled from \( U\{1,\ldots,K\} \), where \( K \) represents number of categories of the discrete variable
- **Dirichlet Data** (conditional-probability table generators) Conditional probability table filled in by sampling categorical distribution from Dirichlet prior with symmetric parameters \( \alpha \). Values of \( F\alpha \) less than 1 provide lower entropy distributions, while \( \alpha \) greater than one provide less information about causal relationships among variables. Nodes/interventions sampled the same as above.

### 4.3 Results

Performance of CSIvA is tested against four other causal discovery methods in three increasingly challenging settings:

- In-distribution experiments (training and test distributions are identical)
- Generalization to out-of-distribution test data (OOD)
- Generalization to OOD test data sampled from a naturalistic causal Bayesian network (CBN)
The four benchmark causal discovery methods that CSIvA is compared to consist of: DAG-GNN [42]; non-linear ICP (Invariant Causal Prediction [32]); DCDI (Differentiable Causal Discovery from Interventional Data) [44]; and ENCO (Efficient Neural Causal Discovery [49]).

Each of these methods are unsupervised and all except ICP rely on neural network functional representations. DAG-GNN uses graph neural networks, but does not incorporate interventional data; DCDI uses normalizing flows to leverage interventional data; ENCO uses neural networks to separately model nodes and edges. Non-linear ICP relies on causality’s inherent robustness to environment changes to estimate underlying causal structures and was excluded from comparisons of graphs with more than 20 nodes due to scaling issues.

The In-distribution experiments provided the easiest proof-of-concept, where CSIvA demonstrated a capacity to learn a mapping between data and the corresponding adjacency matrix, and outperformed the comparison methods across each synthetic data type. Examples of estimated DAGs tested on an in-distribution experiment are shown in figure 8. Furthermore, the relative improvement in CSIvA’s performance increases with the number of nodes.

The out-of-distribution experiments tested CSIvA’s capacity to generalize from the training data to synthetic datasets of varying complexity (ER-1, ER-2, and ER-3 distributions), measuring a small increase loss in Hamming when testing on new distributions. In general terms, results largely corroborate the intuitive conjecture that CSIvA performs better on simpler graphs, and that distance between complexity of the training set and the testing set is inversely related to performance. However, the results show a couple of exceptions: for example, when generalizing to ER-1 and ER-3 test data, the model trained on ER-2 data actually matched or outperformed the models that were trained on ER-1 and ER-2 data, respectively (see figure 9). This is a potential focus for further supervised causal learning research: elucidating the relationship between training data and robust generalization.
Figure 9: Generalizing across graph density for a MLP trained dataset. Training on ER-2 density outperforms other the other trained models for all three density test datasets. A future avenue of study may be towards the characteristics of training sets that are best suited for generalizing.

Finally, the authors demonstrated CSIVa’s capacity to outperform the comparison models when generalizing to a naturalistic CBN distribution. Specifically, they test CSIVa’s capacity to generalize on a benchmark set of natural causal Bayesian networks from the BnLearn repository: Sachs, Asia, and Child[3], [4], [10]. Two versions of the CSIVa model were trained on synthetic data with: one on MLP generated data and another on Dirichlet data, with hyperparameters for the generated data chosen a priori.

5 Discussion

5.1 Summary

The motivation of this review posited that recent advancements in machine learning—particularly the powerful functional representations of neural networks—did not readily apply to advancements in the social science domain. We noted that this may be partially attributed to the need in social science research to step beyond inference and towards causal reasoning. Because the social sciences often lack knowledge of causal priors, we emphasized that the ability to winnow the set of potential causal DAGs from a complex observational dataset would prove a valuable first step in advancing causal research within a social science context. As such, we focused this review on how deep learning techniques were advancing causal discovery methods, and to what extent this progress would be applicable in the social science domain.

We described the transition of causal discovery methodology from a search over the discrete space of potential DAGs to a continuous optimization task that better scales to larger numbers of nodes. The unsupervised learning approaches relied on numerous assumptions on the generating process underlying the observed data distribution, which was manifest in the large hyperparameter space inherent with selecting and weighting components of the scoring function, ‘DAGness’ constraints, and sparsity regularization[35], [39], [42].

Within the context of social science research, navigating this hyperparameter space to accurately reflect complex, messy datasets poses replacing one difficult search problem with another. By comparison, a supervised approach to causal discovery (CSIVA) leveraged the powerful functional representations of neural networks to facilitate a much more adaptive mapping between the available data and potential causal structures. The flexibility of neural causal networks to adapt to the available observational and interventional data and the demonstration of its capacity to generalize across increasingly complex synthetic and naturalistic datasets brings its functionality closer to the use case of social science research. However, there are several limitations with the current methodology that should be considered.

5.2 Limitations

Firstly, although CSIVa has made progress by demonstrating a capacity to generalize to naturalistic distributions from the BnLearn repository, it should be recognized that these demonstrations are still substantial simplifications of messy data we would likely see within real social science research. Even the easier in-distribution tests occasionally misspecify the underlying causal structure—an issue with
potentially notable consequences if social science researchers were relying on this approach without due caution.

Secondly, this approach currently does not address latent confounders, a challenge that is almost assuredly present in real-life applications.

Lastly, it’s worth emphasizing that causal discovery is fundamentally a task of inference. In other words, rather than extracting the complete and accurate causal structure from an observed dataset, these causal discovery methods seek to winnow the realm of potential causal structures to promising candidates. Just as in other machine learning methods, this poses the possibility of inadvertently reflecting shortcomings and biases within the training data into the resulting estimated causal structures—especially if future progress entails synthesizing datasets that are modeled after observed data distributions, as suggested in section 5.4. Neural causal discovery’s advantage of avoiding the explicit priors and large hyperparameter space of supervised causal discovery methods introduces a ‘black box’ challenge, where researchers may be overly trusting of estimated causal structures while lacking the rationale that’s hidden within the neural layers of a transformer. All of this is to say that neural causal discovery methods pose the same risks of biases and unchecked assumptions as other machine learning methods applied within the social science context, and accordingly require a similar level of scrutiny.

5.3 Data Synthesis for Developing Causal Discovery Techniques

It’s important to recognize that the success of the supervised neural causal discovery approach is very dependent on its method of synthesizing training and validation datasets with ground truth causal structures. Accordingly, a fundamental requirement is the capacity to generalize from a synthetic training set to real-world data. This question of training on synthetic data for real-world causal discovery applications has been the subject of some debate in causal literature. Reisach cautions against the reliance on synthetic datasets, demonstrating how synthesized datasets generated from additive noise models exhibit marginal variance that increases along the causal order, artificially elevating their performance in a way that doesn’t generalize to real-world applications.

Another paper, Gentzel, offers a similar concern about the reliance on synthetic training data, and suggest that establishing a benchmark empirical dataset for evaluating causal techniques would be an important step forward. The author goes on to suggest that large-scale software systems may provide an environment for measuring causal data with corresponding interventions.

For example, they suggest a Postgres database where query runtime would be an output variable of interest, and running the same query with different settings of key parameters could be considered interventions. To demonstrate their point, the authors apply two causal discovery techniques (PS and GES) to this dataset, and then compare the resulting models’ performances in predicting the change of distribution resulting from an intervention. They find substantial differences in relative performance between the causal discovery methods when estimating interventional distributions in synthetic data vs empirical data. (See figure 10.)

Notably, Gentzel are suggesting a shift in evaluation from the Hamming distances calculated on a ground truth causal structure, to accuracy in estimating the result of interventions in empirical datasets. While this is important context for considering the reliability of supervised causal discovery, the approach outlined by Gentzel suggests important issues worthy of further evaluation.

Firstly, this approach of shifting the validation method of causal discovery techniques to estimating interventional distributions in real datasets still leaves supervised learning methods reliant on synthetic data for training. Secondly, unlike checking the Hamming distance between ground truth and causal structures, this approach of validation could vary significantly depending on the intervention being estimated. So this technique would likely benefit from incorporating an increasingly diverse set of estimated interventions, which of course makes the task of gathering such a natural validation dataset more challenging. Lastly, for this validation method to prove more persuasive within a social science context, it would likely require finding natural validation datasets related to the field.

5.4 Further Work

The recent advancement of neural causal learning and its dependence on synthetic data for ground truth causal structures suggest two avenues for continuing progress: development of a causal discovery benchmark library, and an exploration of alternative, downstream validation methods that do not rely on a ground truth causal structure.
Figure 10: Results of testing the mean Total Variance Distance (TVD)\cite{37} between the estimated and ground truth interventional datasets for three different causal discovery search-based methods: Greedy Equivalence Search (GES), max-min hill-climbing (MMHC), and PC. This metric is compared across three different datasets, two synthetic and one real. The two synthetic datasets were generated by running PC (a) and GES (b), respectively on the real, empirical dataset. The resulting estimated causal structures were each treated as ground truth for synthetic charts a and b, and used to synthesize data. Mean TVD was calculated on the respective synthetic datasets and compared to the real empirical one, and showed substantial variation in the relative performance of each model across each dataset type. This is shown as a demonstration of how causal discovery methods can be sensitive to the generation methods of the synthetic datasets they operate on.

\[
TV_{\hat{P},P,T=t}(O) = \frac{1}{2} \sum_{o \in \Omega(O)} |P(O = o|do(T = t)) - \hat{P}(O = o|do(T = t))|, \text{ for outcome } o, \text{ intervention } t, \text{ and } \Omega(O) \text{ is the domain of outcome variable } O
\]

5.4.1 Causal Discovery Benchmark Library

The development of a naturalistic causal discovery benchmark library that can be used for training and validating new learning-based techniques would substantially facilitate progress towards developing applicable neural causal learning methods. This would not only push discovery models into a more realistic data environment, but it would also allow researchers to more readily compare the relative performance of causal discovery techniques. Ideally, such a library would include the following features:

- The generative structures of the benchmark library would be modeled after real data, both in terms of the complexity of the underlying causal structure and the joint distributions of the variables. Ideally this could include distributions modeled after real data collected in the social science domain to develop and demonstrate applicability within the field.

- Provided a given generative structure, there ideally would be a capacity for researchers to control elements of the training and testing environment in defined ways:
  - Presence/prevalence of known/unknown interventions
  - Size of the causal structure
  - Sparsity of connections
  - Complexity of relationships between variables

- Ideally such a library could include larger generative structures modeled with unobserved, latent variables to better align with a real-world environment

This is similar to the bnlearn\cite{16} r-package which offers various benchmark bayesian networks, several of which were used in papers referenced in this review. However, the approach suggested here would further develop the idea, allowing for more complex models with non-linear relationships that are more closely modeled after real-world data. Accordingly, recent methods of data imputation such as DECAF\cite{46} could be leveraged to synthesize much more naturalistic datasets with known causal structures.
5.4.2 Alternative Validation Methods

Many of the techniques reviewed here relied on the Hamming distance from a ground truth causal structure to validate model performance. Exploring alternative validation methods that do not require ground truth causal structures could include searching for real-world, interventional datasets, ideally from within the field of social science research. This would allow researchers to test causal discovery methods on the downstream task of predicting the shift in distribution resulting from a previously unseen intervention. Finding such data may be a prohibitively difficult task within the social science domain, but if a neural causal discovery method was able to demonstrate the ability to reliably predict unseen interventional distributions from real social science datasets, this would go a long ways towards demonstrating the applicability of supervised causal discovery methods in real-use causal learning tasks.

References


