Regret-based Federated Causal Discovery

Osman Mian OSMAN.MIAN@CISPA.DE
CISPA Helmholtz Centre for Information Security
Stuhlsatzenhaus 5, 66123 Saarbruecken, Germany

David Kaltenpoth DAVID.KALTENPOTH@CISPA.DE
CISPA Helmholtz Centre for Information Security
Stuhlsatzenhaus 5, 66123 Saarbruecken, Germany

Michael Kamp MICHAEL.KAMP@UK-ESSEN.DE
Institute for AI in medicine IKIM,
Ruhr-University Bochum, and Monash University
Girardetstr. 2, 45131 Essen, Germany

Editor:

Abstract

In critical applications, causal models are the prime choice for their trustworthiness and explainability. If data is inherently distributed and privacy-sensitive, federated learning allows for collaboratively training a joint model. Such approaches for federated causal discovery, however, require sending local causal models, revealing the local data structure. We propose privacy-preserving federated causal discovery by distributed min-max regret optimization. This technique requires clients to only send local regret values, instead of model parameters, ensuring the privacy of sensitive local data. Initial results show that our approach reliably discovers causal networks without ever looking at local data or local causal structures.

Keywords: Federated Learning, Causal Discovery

1. Introduction

We consider a setting where we have multiple distributed datasets over a fixed set of variables X generated from the same causal network such that the datasets can not be pooled together at a single location. Our goal is to discover the overall causal network over these datasets in a privacy preserving manner.

To avoid privacy violations in many critical applications, such as healthcare, we cannot pool data. Discovering causal models is not straightforward when the data is distributed. A naive federated learning approach would be to discover individual causal models for each local dataset, pool those models and parameters associated with the models and compute the likely global causal model governing the process that generated all local datasets. However, federated learning involving parameter sharing may not be completely privacy-preserving and may still allow for reconstruction of local data (Geiping et al., 2020; Lyu and Chen, 2021; Singhal et al., 2021). Applying federated causal discovery in privacy-sensitive applications therefore requires to only transmit insensitive statistics about local data.

While there exist approaches capable of discovering causal networks (Spirtes et al., 2000; Chickering, 2002; Shimizu et al., 2006; Huang et al., 2018; Peters et al., 2014), they are designed to work only on a single dataset. Approaches that do take multiple datasets into account require that data first be pooled together at a single location (Mooij et al., 2016; Zhang et al., 2017; Magliacane et al., 2018), work only for a single target variable (Peters et al., 2016), or place strict assumptions on the underlying causal mechanisms (Shimizu, 2012; Ghassami et al., 2017) that are unlikely to hold in practice.

To solve this problem, we propose a privacy-preserving score-based framework for Regret-based Federated Causal Discovery, called RFcd, which can be instantiated using any consistent score-based causal discovery algorithm. In RFcd multiple clients collaboratively discover a joint causal model governing their local datasets without transmitting any data or local models. The basic idea is as follows. First, each client discovers their best-fitting local causal model. Then, the server repeatedly proposes a causal network to each of the clients, who return only the *regret* relative to the proposed causal model. Regret (Grünwald, 2005) measures how much worse is the best causal model consistent with a proposed network compared to the best-fitting local causal model already discovered. The server then uses these regret values to learn a global causal model by minimizing the worst-case regret returned by all clients. Initial results show that our score-agnostic framework reliably discovers causal networks from distributed data, even when our assumptions may not hold.

2. Related Work

There have been a number of causal discovery approaches in recent years (Spirtes et al., 2000; Chickering, 2002; Shimizu et al., 2006; Peters et al., 2014; Marx and Vreeken, 2017; Huang et al., 2018; Mian et al., 2021) that are designed to work on a single dataset. One way to adapt these approaches to discover causal networks from multiple datasets is to collect all the data centrally. Even if we centrally collect the data, we should still not simply pool it all together, as it is well known that naively pooling the data introduces serious biases in estimation (Lee and Tsui, 1982; Tillman, 2009). Subsequently, methods have been proposed to systematically combine data from multiple sources such that causal networks can be learned (Yang et al., 2018; Squires et al., 2020). One of the popular approaches is to introduce one or more context variables to distinguish the rows of the combined datasets and perform causal discovery on augmented data (Zhang et al., 2017; Magliacane et al., 2018; Ding et al., 2020). Most popular of such approaches is the Joint Causal Inference (Jci) framework (Mooij et al., 2016) defined for constraint-based approaches.

All of the aforementioned approaches, however, require data to be centralized, which is prohibitive in many applications. In such cases, a federated causal discovery approach allows to discover the network without centralizing data. The topic of federated causal discovery is relatively young. Existing approaches allow for causal inference (Xiong et al., 2021) or causal discovery (Shimizu, 2012) given some parametric assumptions. More recent approaches overcome this issue, but either sacrifice a convergence guarantee (Gao et al., 2021), or require sharing additional learning parameters (Ye et al., 2022; Na and Yang, 2010). All of the aforementioned approaches still share local causal models that can reveal sensitive information about data, up to the point of allowing attackers to reconstruct local data from model parameters (Geiping et al., 2020; Lyu and Chen, 2021; Singhal et al., 2021).

3. Preliminaries

3.1 Problem Setting and Notation

We have data X, over m variables, split into multiple different datasets $X^{(1)}, \ldots, X^{(l)}$ of sizes $n^{(1)}, \ldots, n^{(l)}$ drawn i.i.d. from the same global distribution $P(X^{(i)})$ which we assume to correspond to the true causal network G^*. We assume that the data cannot be shared to create a single dataset due to e.g., privacy concerns.

Our goal is to discover the true causal network G^* from local datasets $X^{(i)}$, which we assume to be the same for all for all i. While on each dataset $X^{(i)}$ we can try to find a *local* causal graph $G^{(i)}$, if each of the datasets $X^{(i)}$ contains only few samples, none of these learned graphs $G^{(i)}$ will be a good description of the true generating process. Note that discovering local causal models $G^{(i)}$ will, in the limit with respect to local dataset size, uncover the true network G^*. For finite dataset sizes, however, locally discovered causal models $G^{(i)}$ can vary substantially from G^*.

We therefore develop a method to discover a causal network over the data X which does not require us to pool the data. Before we can do so, we first introduce the necessary assumptions to perform causal discovery in our setting.

3.2 Assumptions for Causal Discovery

We assume that the shared distribution $P(X^{(i)})$ and the true causal network G^* satisfy the following assumptions generally made in the field of causal discovery.

First is the *Causal Markov Condition* which states that every node in G^* is independent of its non-descendants conditional on its parents in P. Second is the *Causal Faithfulness* assumption is that if sets U, V are independent given W in P then W d-separates U and V in G^*. Together, the Causal Markov Condition and Causal Faithfulness condition entail that the conditional independence relations in P correspond precisely to d-separation relations in G^*. Last, we make the *Causal Sufficiency* assumption, telling us that all common parents of all observed variables are themselves observed. That is, there are no latent confounders affecting our data. The sufficiency and faithfulness conditions, however, are not a must. We make these assumptions to simplify explanation as our initial proof-of-concept is presented over score-based approaches all of which make these aforementioned assumptions. In practice, the assumptions of our framework will be as strict as the assumptions of the score-based approach that is used inside it, as we explain later in Sec. 6.

When all of the above assumptions hold, certain algorithms, e.g. Ges (Chickering, 2002) and the Pc algorithm (Spirtes et al., 2000) have been shown to be consistent for a single dataset and discover causal networks up to the Markov equivalence class (Glymour et al., 2019).

3.3 Learning from Regrets

Let $L(X^{(i)}; G)$ be a scoring function, usually of the form $L(X^{(i)}; G) = -\log p(X^{(i)} \mid G) + L(G)$ where $L(G)$ is a function penalizing the complexity of the graph G and $-\log p(X^{(i)} \mid G)$ is the log-likelihood assuming a certain class of generating functions, e.g. linear relationships between each variable and its parents in G.

Due to privacy concerns, we do not pool local datasets $X^{(i)}$ or the local causal models $G^{(i)}$. Instead, each client shares only the *regret* of a proposed network G with respect to the local causal network $G^{(i)}$ on its local dataset X^i, formally defined as

$$R_i(G) \coloneqq L(X^{(i)}; G) - \min_{G^{(i)}} L(X^{(i)}; G^{(i)}) \,,$$

where we drop the dependence on the data $X^{(i)}$ to simplify notation.

It measures how much worse the proposed network G is compared to the *best* network for the data $X^{(i)}$. Our goal then is to minimize the *maximum* regret over each of the datasets $X^{(i)}$, i.e. want to find $\widehat{G}$ as follows

$$\widehat{G} = \operatorname*{argmin}_{G} \max_{i} R_i(G) \ . \tag{1}$$

The obtained network $\widehat{G}$ is the one which is least bad compared to any of the individual networks. It trades off errors relative to one network $G^{(i)}$ for errors relative to another network $G^{(j)}$ and tries to minimize them as much as possible. Intuitively, local networks in finite-sample setting may not be identical to each other. Our aim, therefore, is to systematically unify them by looking for the next best thing: a network $\widehat{G}$ that is 'nearly' as good as each $G^{(i)}$. In doing so, we can give the best worst-case guarantees over all datasets.

4. Regret-based Federated Causal Discovery

In this section we describe the practical implementation of our framework defined in Eq 1 which we refer to as the Regret-based Federated Causal Discovery (RFcd). We begin by dealing with the intractability of finding the minima in Eq. (1) in a score-agnostic manner.

To this end, let $\mathcal{A}$ be any score-based structure discovery algorithm, e.g. GES (Chickering, 2002) or GSP (Solus et al., 2017) and let L be any consistent score used within $\mathcal{A}$ such as the BIC-score (Schwarz, 1978) or MDL-based score (Mian et al., 2021). Then we can replace the terms $\min_{G^{(i)}} L(X^{(i)}; G^{(i)})$ in the regret term as follows

$$\begin{aligned} \widehat{G} &= \operatorname*{argmin}_{G} \max_{i} \left(\widehat{R_i}(G)\right), \text{ where} \\ \widehat{R_i}(G) &\coloneqq L(X^{(i)}; G) - L(X^{(i)}; \widehat{G^{(i)}}) \,, \end{aligned} \tag{2}$$

where $\widehat{G^{(i)}}$ is the graph learned by $\mathcal{A}$ on $X^{(i)}$, and $L(X^{(i)}; G)$ is the score that evaluates how well does G fit the data $X^{(i)}$, The idea is that when $\mathcal{A}$ is a consistent algorithm with respect to L then as $n^{(i)} \to \infty$ we find that $\widehat{G^{(i)}} \to G^* \leftarrow \operatorname{argmin}_{G^{(i)}} L(X^{(i)}; G^{(i)})$ so that for sufficiently large datasets the above replacement is harmless.

Next, given the goal in Eq. (2), how do we estimate $\widehat{G}$? We can, in theory, find the true causal network by exhaustively searching over the space of DAGs and taking the one minimizes Eq. 2. Such an approach, however quickly becomes infeasible as the space of DAGs grows super-exponentially in the number of nodes and the loss landscape associated with L does not exhibit any structural regularities, which makes optimal Bayesian structure discovery NP-Hard (Chickering et al., 2004).

In practice, we can implement the above approach as a beam search which is guaranteed to find the optimum, given correct beam size, but also allows to reduce beam size to trade optimality for runtime. That is, starting from the empty network G_0 we evaluate at every step every one-edge extension G of the b best networks $G_{t,1}, \ldots, G_{t,b}$ from the previous step and keep the b best networks from the current step $G_{t+1,1}, \ldots, G_{t+1,b}$. We repeat this until no further extensions of any of the networks $G_{t,j}$ improve upon the best network already found. Then we set

$$\widehat{G}_{\mathcal{B}} = \operatorname*{argmin}_{G_{t,j}} \max_{i} \left(L(X^{(i)}; G_{t,j}) - L(X^{(i)}; \widehat{G^{(i)}}) \right) ,$$

to be the best performing network discovered so far.

We can use the above formulation to perform federated causal discovery as shown in Algorithm 1. Given a server S and l different clients $C^{(1)}, \ldots, C^{(l)}$ each with their own private datasets $X^{(1)}, \ldots, X^{(l)}$, the server communicates the algorithm $\mathcal{A}$ and the scoring metric L to each of the clients. Each client then runs $\mathcal{A}$ on its own data to learn the local $\widehat{G^{(i)}}$ (line 4). The server then sends an empty network G_0 to each client and receives the regrets $r_0^{(i)}$ w.r.t locally learned networks back from each $C^{(i)}$. Next, the server calculates the worst-case regret r_0 (lines 5-7) and initializes a beam B of size b with the state (G_0, r_0) (line 8). Then the search process begins. At each search-step, all possible single edge extensions of each DAG in the beam are enumerated and their worst-case regret calculated via communication between the server and clients (lines 11-14). The top b extensions with lowest worst-case regret are then retained as the new beam (line 15). An immediate advantage of our search procedure is that it is guaranteed to converge. This is because regret defined using a consistent score L used within $\mathcal{A}$ can never go below 0, and we only take steps that reduce regret. Hence we continue the search until convergence. During the entire learning process, the server neither sees the data, nor the locally learned causal networks for each dataset. The only communication that takes place between server and client is the regret value r_G for a query DAG G.

Setting beam size to $\binom{m^2}{(m^2-m)/2}$ is equivalent to an exhaustive search where we are guaranteed to find the global optimum. This, however, is only suitable for networks with small number of variables and quickly becomes infeasible even for variable sizes $m = 8$. Alternatively, setting $b = 1$ results in a greedy DAG search algorithm which is only guaranteed to discover correct causal network if the underlying structure is a tree. In practice, we find that setting beam-sizes as small as 10 already performs well even though our search is only guaranteed to find local optima in those cases.

5. Experiments

We now provide initial results of our work on RFcd. For our assessment, we are mainly interested in answering the following two questions: (1) Can RFcd discover causal networks over data from multiple environments? (2) How well does RFcd perform on networks where our assumptions may not hold? We describe our experimental setup and then provide empirical results to answer these questions.

We instantiate RFcd framework using Ges as the score-based algorithm $\mathcal{A}$ and two scoring criteria which are proven to be consistent: namely the BIC-score (Schwarz, 1978)

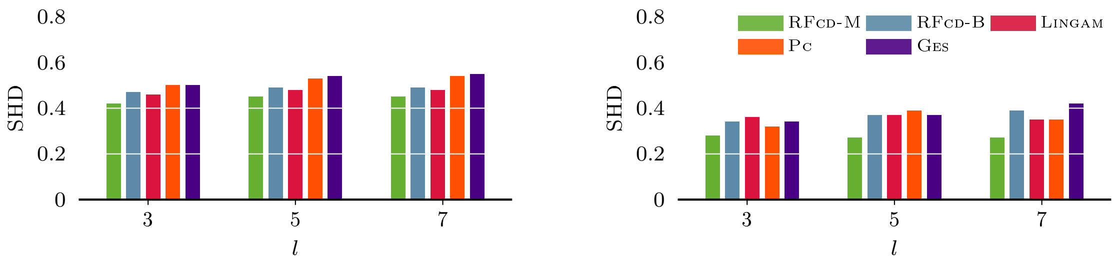
Algorithm 1: Causal Discovery using RFcd framework

```
Input: algorithm $\mathcal{A}$, consistent scoring criteria $L$, beam size $b$
Output: causal network $\mathbf{G}$
$B \leftarrow \emptyset$
$r_0 = 0$, $G_0 \leftarrow \emptyset$
for $i = 1 \dots l$ do
    $c[i]$.learn$(\mathcal{A}, L)$
    $r_0^i \leftarrow c[i]$.regret$(G_0, L)$
    if $r_0^i > r_0$ then
        $r_0 \leftarrow r_0^i$
$B \leftarrow B \oplus (G_0, r_0)$
repeat
    $Q \leftarrow B$.copy() // priority queue to rank DAGs in order of increasing regret
    $\mathbb{G} \leftarrow$ all admissible single edge extensions of DAGs in $B$
    foreach $G \in \mathbb{G}$ do
        $r_G \leftarrow$ max$(c[i]$.regret$(G, L))$ for $i = 1 \dots l$
        $Q \leftarrow Q \oplus (G, r_G)$
    $B \leftarrow$ first $b$ entries in $Q$
until convergence;
$\mathbf{G}^*, r_{G^*} \leftarrow$ first entry in $B$
return $\mathbf{G}^*$
```

and spline-based MDL score proposed by Mian et al. (2021). We refer to these instantiations as RFcd-B and RFcd-M respectively. We compare to Direct-Lingam (Shimizu, 2012) as representative ANM based method for federated causal discovery over multiple groups. As baseline, we use Ges (Chickering, 2002) and Pc (Spirtes et al., 2000) to locally discover causal networks within each environments and take a union over the discovered networks to predict the global causal network, Note that this still requires us to share the local causal networks. Comparing directly to CdNod (Zhang et al., 2017) or the Jci framework (Mooij et al., 2016) is not possible as they require that we first collect all the data in one place and are therefore not applicable in our setting.

Under our assumptions in Sec. 3.2, causal networks are only identifiable up till Markov equivalence class (Glymour et al., 2019). We therefore convert the causal network predicted by RFcd to a Completed Partial Directed Acyclic Graph[1] (CPDAG) before calculating any evaluation statistics. We evaluate the predicted networks in terms of structural similarity using the Structural Hamming Distance (Shd) (Tsamardinos et al., 2006). Let G and H be the ground truth resp. predicted causal DAG, then Shd(G, H) counts the edges where the two causal DAGs differ. To keep results comparable across different settings we normalize Shd between 0 and 1. To avoid susceptibility to practical issues like var-sortability (Reisach

[1] A Completed Partial Directed Acyclic Graph (CPDAG) is a graph that represents the Markov equivalence class of DAGs



(a)

(b)

Figure 1: [Lower is better] SHD for our identifying causal structure with synthetic data generated from networks with 5 variables (left) and 10 variables (right). Results for RFcd-M and RFcd-B are reported for beam size=10

et al., 2021) we standardize all data to zero mean and unit variance before running the experiments. Next we provide our evaluation results.

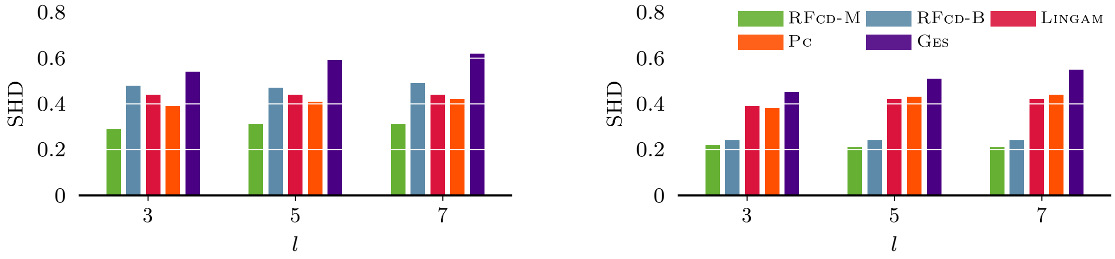
Can RFcd discover causal networks over data from multiple environments? We start with the setting where we generate multiple datasets using the same underlying distribution. We have number of environments $l \in \{3, 5, 7\}$, number of variables $m \in \{5, 10\}$, samples per environment $n = 1000$, and beam-size $b \in \{1, 5, 10\}$ as our experimental setting. We simulate DAGs using the Erdős-Rényi model and generate each variable's data as a non-linear polynomial function of its causal parents, using additive Gaussian noise in half of the cases and additive uniform noise for the other half of the instances. We use the Causal Discovery Toolbox (Kalainathan and Goudet, 2019) to generate this data.

We report the Shd for each approach in Fig. 1. We see that RFcd-M outperforms Lingam and the baseline algorithms for both variable sizes 5 and 10. We find that RFcd-B is better than the baselines but performs worse than RFcd-M possibly due to a lenient penalty on the model parameters as compared to RFcd-M. This lenient penalty allows for inclusion of more, potentially spurious, edges when working with limited data. Furthermore, we see that both RFcd-B and RFcd-M exhibit consistent performance even as the number of environments grows, whereas the performances of other approaches degrades with increasing number of environments.

We find that for greedy DAG search, varying beam-size does not have a drastic effect on the quality of the predicted causal networks. This is because the beam quickly gets dominated by correlated local expansions of a given DAG, which have similar worst-case regrets. To maintain a diverse set of states in the beam, we would need to choose a very large beam size that results in unrealistic run-times even for networks as small as 5 variables. We report the average Shd for varying beam sizes in Table 1.

Table 1: [Lower is better] Average SHD for varying beam-sizes b

b	m	RFcd-M	RFcd-B
1	5	0.44	0.51
5		0.46	0.51
10		0.46	0.50
1	10	0.28	0.35
5		0.27	0.37
10		0.28	0.37



(a)

(b)

Figure 2: [Lower is better] Shd for causal structures with synthetic homogeneous data (left) and heterogeneous data (right) over RFcd-M, RFcd-B, Lingam, Pc, and Ges. Results for RFcd-M and RFcd-B are reported for beam size=10

How well does RFcd perform on networks where our assumptions may not hold? Second, and more interestingly, we investigate a challenging setting where we generate data in each environment using different interventional distributions from a *fixed* underlying causal network. This means that the data for each environment comes from a different (sub-)network, which may be either observational or interventional. We provide no prior knowledge about the type of the dataset nor about the intervention targets to the methods.

We run all methods and report the results in Fig. 2 where we see both RFcd-M and RFcd-B clearly outperform the competitors, even in this challenging setting. This shows that RFcd is robust to situations where data across each environment is heterogeneous. One explanation of this result is that increase in worst-case regret when including an intervened edge in an interventional dataset is smaller than the decrease in the same for observational datasets where the edge is present, thereby preferring the inclusion of the edge, even though it may be absent in some of the datasets.

6. Discussion and Future Work

In this work we have presented a federated causal learning framework that is based on minimizing the worst-case regret over distributed environments. We showed empirically that using the min-max regret framework we can instantiate a federated causal learning approach that only requires each environment to communicate regret values to the server, and discovers meaningful causal structure from data, even when some of our assumptions may not hold.

Despite the initial promising results, RFcd is still a work in progress. Since our proposed beam search is over the space of DAGs, we need the causal sufficiency assumption. In general, however, it is possible to relax both causal faithfulness and causal sufficiency assumptions by using any consistent pair $(\mathcal{A}, L)$ within our regret-based framework that does not require these assumptions. One additional, non-trivial modification that this entails is that beam-search must now be defined over the space of Partial Ancestral Graphs (PAGs) instead of DAGs. How to efficiently enumerate PAGs and their consistent extensions is one of our current lines of investigation. Moreover, greedy DAG search with a fixed beam-size can get stuck in local optima, which we see in Table 1. Given a large enough beam-size we can, in theory, arrive at the true causal network but such an approach quickly becomes infeasible as the space of DAGs grows super-exponentially in the number of nodes. The problem of optimal Bayesian structure learning is known to be NP-Hard after all (Chickering et al., 2004). To improve this, we need to formulate the RFcd framework using a Ges-like procedure such that we search over the space of Markov equivalence class instead of the space of DAGs. This requires us to first give formal identifiability guarantees for our score defined in Eq. (1). We are therefore working on proving consistency of our proposed score as well as instantiating a federated Greedy Equivalence Search (Chickering, 2002) procedure as an upgrade to the beam-search. These improvements will allow us to transform RFcd into a non-parametric score-based causal discovery framework that can be instantiated using *any* consistent score. Finally, we will investigate privacy preserving guarantees of our approach, with the goal of providing guarantees under certain conditions.

7. Conclusion

We propose a federated causal learning framework that is different from existing work in the sense that it does not require any of the environments to share either their data or their local causal networks with the central server. We showed that a simple instantiation of our framework using beam-search is already effective in discovering causal networks in a federated setting, even in cases where our assumptions of homogeneous data may not hold.

As a continuation of this work we would like to investigate formal privacy guarantees that can be given using RFcd framework. We aim to further provide consistency guarantees for the proposed optimization score and use these guarantees to evolve beam search into a more complex search procedure such as the Greedy Equivalence Search(Chickering, 2002; Ramsey et al., 2017). This will allow us to transport our theoretical guarantees into a practical instantiation and to further improve our results.

References

David Maxwell Chickering. Optimal structure identification with greedy search. *JMLR*, 3: 507–554, 2002.

Max Chickering, David Heckerman, and Chris Meek. Large-sample learning of bayesian networks is np-hard. *JMLR*, 5, 2004.

Chenwei Ding, Biwei Huang, Mingming Gong, Kun Zhang, Tongliang Liu, and Dacheng Tao. Score-based causal discovery from heterogeneous data. 2020.

Erdun Gao, Junjia Chen, Li Shen, Tongliang Liu, Mingming Gong, and Howard Bondell. Federated causal discovery. *arXiv preprint arXiv:2112.03555*, 2021.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients - how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 16937–16947. Curran Associates, Inc., 2020.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. *arXiv preprint arXiv:1705.09644*, 2017.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 2019.

Peter Grünwald. Minimum description length tutorial. In Peter Grünwald and I.J. Myung, editors, *Advances in Minimum Description Length*. MIT Press, 2005.

B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour. Generalized score functions for causal discovery. In *KDD*. ACM, 2018.

Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*, 2019.

Sik-Yum Lee and Kwok-Leung Tsui. Covariance structure analysis in several populations. *Psychometrika*, 47, 1982.

Lingjuan Lyu and Chen Chen. A novel attribute reconstruction attack in federated learning. *arXiv preprint arXiv:2108.06910*, 2021.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *NIPS*, volume 31, 2018.

Alexander Marx and Jilles Vreeken. Telling Cause from Effect using MDL-based Local and Global Regression. In *ICDM*, pages 307–316. IEEE, 2017.

Osman Mian, Alexander Marx, and Jilles Vreeken. Discovering fully oriented causal networks. In *AAAI*, 2021.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *JMLR*, 21, 2016.

Yongchan Na and Jihoon Yang. Distributed bayesian network structure learning. In *2010 IEEE International Symposium on Industrial Electronics*, pages 1607–1611. IEEE, 2010.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *JMLR*, 15, 2014.

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

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *J. Data Sci. Anal.*, 2017.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34, 2021.

Gideon Schwarz. Estimating the dimension of a model. *Annals Stat.*, 6(2):461–464, 1978.

Shohei Shimizu. Joint estimation of linear non-gaussian acyclic models. *Neurocomputing*, 81, 2012.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *JMLR*, 7, 2006.

Karan Singhal, Hakim Sidahmed, Zachary Garrett, Shanshan Wu, John Rush, and Sushant Prakash. Federated reconstruction: Partially local federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Liam Solus, Yuhao Wang, Lenka Matejovicova, and Caroline Uhler. Consistency guarantees for permutation-based causal inference algorithms. *arXiv preprint arXiv:1702.03530*, 2017.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT Press, 2000.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048. PMLR, 2020.

Robert E Tillman. Structure learning with independent non-identically distributed data. In *ICML*, pages 1041–1048, 2009.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Ruoxuan Xiong, Allison Koenecke, Michael Powell, Zhu Shen, Joshua T Vogelstein, and Susan Athey. Federated causal inference in heterogeneous observational data. *arXiv preprint arXiv:2107.11732*, 2021.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *ICML*, pages 5541–5550. PMLR, 2018.

Qiaoling Ye, Arash A Amini, and Qing Zhou. Distributed learning of generalized linear causal networks. *arXiv preprint arXiv:2201.09194*, 2022.

Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI*, 2017.