Differentiable and Transportable Structure Learning

Anonymous

16 May 2022 (modified: 26 May 2022) NeurIPS 2022 Conference Blind Submission Readers: Conference, Paper5595 Senior Area Chairs, Paper5595 Area Chairs, Paper5595 Reviewers, Paper5595 Authors Show Revisions (/revisions?id=ORtwEmymLq)

Keywords: causality

TL;DR: Transportability is an important property of structure learning, yet it is not satisfied with gradient-based structure learning; we reintroduce it.

Abstract: We are interested in unsupervised structure learning with a particular focus on directed acyclic graphical (DAG) models. Compute required to infer these structures is typicall exponential in the amount of variables, as inference requires a sweep of a combinatorially large space of potential structures. That is, until recent advances allowed to search this space differentiable metric, drastically reducing search time. While this technique--- named NOTEARS ---is widely considered a seminal work in DAG-discovery, it concedes an important prope differentiability: transportability. In our paper we introduce D-Struct which recovers transportability in the found structures through a novel architecture and loss function, while remain differentiable. As D-Struct remains differentiable, one can easily adopt our method in differentiable architectures as was previously done with NOTEARS. In our experiments we empiric Struct with respect to edge accuracy and the structural Hamming distance.

Supplementary Material: pdf (/attachment?id=ORtwEmymLq&name=supplementary_material)

Revealed to Jeroen Berrevoets, Nabeel Seedat, Fergus Imrie, Mihaela van der Schaar, Paper5595 Senior Area Chairs

15 May 2022 (modified: 26 May 2022) NeurIPS 2022 Conference Submission

Authors: *Jeroen Berrevoets (/profile?id=~Jeroen_Berrevoets1), Nabeel Seedat (/profile?id=~Nabeel_Seedat1), Fergus Imrie (/profile?id=~Fergus_Imrie1), Mihaela van der (/profile?id=~Mihaela_van_der_Schaar2)*

Add **Official Commer**

Reply Type: all Author: everybody Visible To: all readers Hidden From: nobody

[−] **Official Review of Paper5595 by Reviewer dyeR**

NeurIPS 2022 Conference Paper5595 Reviewer dyeR

16 Jul 2022 NeurIPS 2022 Conference Paper5595 Official Review Readers: Program Chairs, Paper5595 Senior Area Chairs, Paper5595 Area Chairs, Paper5595 Reviewers Submitted, Paper5595 Ethics Reviewers, Paper5595 Authors

Summary:
The paper proposes to consider the consistency among structures during the DAG learning process from different datasets. Authors propose a D-struct algorithm to enforce such consistency, by introducing a new regularization loss in the objective. Results on multiple sources and multiples of the same datasets show that the proposed method has better accuracy.

Strengths And Weaknesses:
Strengths:

1. The proposed method separates structure learning from parameter learning in the differential DAG learning algorithms, which is interesting
2. Even within a single dataset, separate learning data folds or batches seems a neglective point and authors showed the importance.
3. Multiple data sources are also considered, which is novel from existing works.

Weaknesses:

1. The presentation and related concepts are not clearly presented or misrepresented
2. It would be helpful if the authors can provide comparison with a multi-task DAG learning algorithm baseline (Multi-task Learning of Order-Consistent Causal Graphs), which makes less assumptions about this work.

Questions:
1. It is indeed true that transportability of the dag structure is limited as the weighted adjacency matrix is both parameter and structure. Would existing DAG learning works using binary masks be able to handle such problems?
2. Definition 1 is not precise. (1) it is not clear whether x_d is in the path or not. (2) is it for one node X_k or all nodes X_k?
3. Line 97: does the back-forth relationship require the faithfulness assumption?
4. Line 126: would a typical score-based DAG learning algorithm also guarantee transportability?
5. Line 161: "the single solution" how about Markov equivalent graphs in G?
6. Section 3.2: some details are not clear here. Do you sort by the values of covariates, ascending or descending?
7. The performance gain is great. However, authors should study whether the performance gain is from imposing the transportability or a clever division of data so it improves the accuracy of DAG learning by being more generalizable.
8. Computational efficiency: it is also not clear how the parallel learning scheme is used. Are they computed at the same time with multiple GPU cores?

Limitations:
The limitations were discussed adequately.

Ethics Flag: No
Soundness: 2 fair
Presentation: 3 good
Contribution: 2 fair
Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.
Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Code Of Conduct: Yes

Add **Offi**

[−] Official Review of Paper5595 by Reviewer F1dW

NeurIPS 2022 Conference Paper5595 Reviewer F1dW

11 Jul 2022 NeurIPS 2022 Conference Paper5595 Official Review Readers: Program Chairs, Paper5595 Senior Area Chairs, Paper5595 Area Chairs, Paper5595 Reviewers Submitted, Paper5595 Ethics Reviewers, Paper5595 Authors

Summary:
This work studies the problem of inferring a DAG from a probability distribution, points out problems with existing algorithms, and proposes a new technique "D-Struct" to overcome the problems. The work is closely based on recent work on differentiable DAG learning (ref [30]). This technique is briefly reviewed and claimed to have shortcomings, in particular the results are not uniquely determined by the DAG structure underlying the observed distribution. It is argued (in section 2.2) that this is because the technique uses a relaxation of the DAG constraint which has spurious solutions. The proposed remedy is a sort of bootstrap in which the observations are subsampled, inference is done for each subsample, and the results combined. Experiments with distributions based on random graphs are done to show that the new method is both faster and identifies a unique DAG.

Strengths And Weaknesses:
Strengths: The problem is clearly explained and the motivation is good; the new proposal is simple.

Weaknesses: I found the arguments sketchy and the description of the experiments to be insufficient.

Questions:
I was not totally convinced by the comments in 2.2 explaining the "non transportability" of differentiable structure learning, and would like to see more argument and/or examples.

I would like to see more detail about the experiments: are the Erdos-Renyi graphs the underlying DAGs ? Is the dimension d in the tables the number of nodes ? The authors observed non transportability for NOTEARS, is it consistent with the claims in section 2.2 ?

I would like to see more detail about the procedure in section 3.2 to divide a single dataset into subsets.

Limitations:
Technical work not discussing specific applications.

Ethics Flag: No
Soundness: 3 good
Presentation: 2 fair
Contribution: 3 good
Rating: 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.
Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Code Of Conduct: Yes

Add Offi

[−] Official Review of Paper5595 by Reviewer as89

NeurIPS 2022 Conference Paper5595 Reviewer as89

08 Jul 2022 (modified: 15 Jul 2022) NeurIPS 2022 Conference Paper5595 Official Review Readers: Program Chairs, Paper5595 Senior Area Chairs, Paper5595 Area Chairs, Paper5595 Reviewers Submitted, Paper5595 Ethics Reviewers, Paper5595 Authors

Summary:
The authors consider unsupervised structure learning over DAG models. They proposed, D-struct, a transportable and differentiable algorithm. The proposed D-struct solves the transportability problem in the prior NOTEARS method. In experiments, they show their approach attain better results over the Erdos-Renyi graph and is efficient during inference.

Strengths And Weaknesses:
Strengths:

- The paper is smooth to read through. The authors presented a useful idea on top of the NOTREAR algorithm.
- Well documented appendix and code.

Weakness

- I hope the authors would consider more baselines and highlight the contribution of this work not just by comparing to one single baseline. Currently, this area has many other new works, including:

DAG-GNN: DAG Structure Learning with Graph Neural Networks. ICML2020

Learning Sparse Nonparametric DAGs. AIStats2020

Gradient-based neural dag learning. ICLR2020.

- How to conduct the conditional independence test on the graph may need more explanations. How the proposed work is different from those historical CIT-based works may need explanations.

Questions:
1. In terms of Definition 2, how do you compare if the learned DAGs are equivalent or not? Since DAGs may have different orientations, did the comparison algorithm account for that?
2. I don't quite understand the motivation and result in Figures 3 and 4. Would you please explain that?
3. From my current understanding, the main novelty is presented in Equation 4. Could you elaborate more on this equation, and how it achieves transportability? Do we have any other candidate function to be chosen? Why is this the best one?

Limitations:
The authors mentioned that the proposed method D-struct cannot be applied to reveal causal structures.

Ethics Flag: No
Ethics Review Area: I don't know
Soundness: 2 fair
Presentation: 3 good
Contribution: 2 fair
Rating: 5: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.

Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

Code Of Conduct: Yes

Add Offi

[−] **Official Review of Paper5595 by Reviewer 8Wft**

NeurIPS 2022 Conference Paper5595 Reviewer 8Wft

21 Jun 2022 NeurIPS 2022 Conference Paper5595 Official Review Readers: Program Chairs, Paper5595 Senior Area Chairs, Paper5595 Area Chairs, Paper5595 Reviewers Submitted, Paper5595 Ethics Reviewers, Paper5595 Authors

Summary:

The paper proposes to extend NOTEARS to encourage the algorithm to find the same graph for different datasets. The proposed method, D-struct, consists of K NOTEARS models that are run in parallel over K different datasets. The loss function is composed of the original NOTEARS objective plus an extra term that penalises each model if its predicted graph differs from the average of the K graphs. D-struct is shown to outperform the original NOTEARS on synthetic data in terms of computational efficiency and solution quality.

Strengths And Weaknesses:

Strengths

- The paper is well written and easy to follow.
- The proposed method seems to produce good results, and is simple and easy to implement on top of the already popular NOTEARS algorithm.

Weaknesses

- I am not convinced the discussion around transportability is theoretically sound (see questions below).
- The penalisation term is a heuristic with a somewhat weak motivation. I think some similarly simple baselines (e.g. running NOTEARS on K bootstrap samples and aggregating the results) are missing.
- The experimental procedure is not entirely clear (see questions below).

Minor issues

- In Algorithm 1, it seems each of the K sets of parameters is trained on the same batch of data. Shouldn't the batch be sampled from D_k instead?
- Line 42: In "While we why this is the case in section 2.2", there is probably a verb missing after "we".
- Line 57: DSF is only defined later on line 144. Probably best to spell out the acronym here instead.
- Line 69: Unless NOTEARS generalizes all differentiable structure learning methods, which I believe is not the case, one should not say "without loss of generality" here.

Questions:

Transportability

1. I am confused about the overall reasoning around transportability. First of all, I believe conditional independence tests (CITs) methods are not guaranteed to be transportable in the sense of the definition used in the paper. Without further assumptions on the distribution P or class of DAGs, structure learning algorithms only retrieve an I-equivalence class (i.e. a set of graphs encoding the same independence relationships). Unless I am missing something (and I might very well be) this renders the development and motivation of the method incorrect.
2. On that note, how do the different graphs retrieved by NOTEARS (and D-struct) compare to each other? Do they belong to the same I-equivalence class?

Experimental Procedure

1. How are the confidence intervals in each table computed? I assume it is across the 5 random ER graphs, but is probably best to make this clear in the text.
2. How sensitive is D-struct to the data splits? Some of the results for D-struct are unexpected, e.g. the results for varying number of instances (n) are unusual; NOTEARS improves with more data, as one would expect, but that is not always true for D-struct. Could it not be due to the randomness introduced by the data splits? How did the authors control for that?
3. In Table 1 and Figure 6, it is probably best to mention the reference values for n, s, and d. That is, the values at which each parameter is kept fixed when varying the others. Judging by the results, it is n=1000, s=2d and d=5, but probably best to mention it. I could not find a similar pattern in the tables in the appendix though. Maybe there is a typo or different reference values were used there.
4. How does the number of parameters in D-struct compares to that of NOTEARS?
5. What would happen if one were to run NOTEARS with K bootstrap samples from the data, and average the K graphs a posteriori? Would that be much worse than D-struct? I think that would be an important baseline.

Computational Efficiency

1. Figure 5 seems incomplete. The speedups are computed with respect to what? Running time or number of steps to convergence?
2. The experimental procedure is not clear. Was K the same in all cases? How does the number of parameters in D-struct compares to that of NOTEARS?
3. The use of the word parallel here is a bit confusing. Is each of the K datasets used in D-struct ran in parallel? If so, shouldn't the speedup be proportional to K?
4. Also, I would expect the speedup to improve with n, which is not always the case. Why could that be? How many random seeds were used to compute these results?

Limitations:

The authors did a good job that stating clearly that their method does not retrieve causal relationships. I do not see any major limitation that should be specifically addressed.

Ethics Flag: No

Soundness: 1 poor

Presentation: 3 good

Contribution: 2 fair

Rating: 3: Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility and incompletely addressed ethical considerations.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes

Add Offi