Latent Variable Models for Bayesian Causal Discovery

Jithendaraa Subramanian [1 2] **Yashas Annadani** [3] **Ivaxi Sheth** [1 4] **Stefan Bauer*** [3] **Derek Nowrouzezahrai** [1 2] **Samira Ebrahimi Kahou** [1 4 5]

Abstract

Learning predictors that do not rely on spurious correlations involves building causal representations. However, learning such a representation is very challenging. We, therefore, formulate the problem of learning a causal representation from high dimensional data and study causal recovery with synthetic data. This work introduces a latent variable decoder model, Decoder BCD, for Bayesian causal discovery and performs experiments in mildly supervised and unsupervised settings. We present a series of synthetic experiments to characterize important factors for causal discovery and show that using known intervention targets as labels helps in unsupervised Bayesian inference over structure and parameters of linear Gaussian additive noise latent structural causal models.

1. Introduction

Exploiting structure in the data to infer latent variables and capture causal mechanisms is crucial for causal representation learning (Schölkopf et al., 2021). Such a representation would allow for counterfactual reasoning in a manner similar to that of humans, thereby moving away from models that rely on exploiting spurious correlations for prediction.

Causal mechanisms are usually modelled as Bayesian Networks or Directed Acyclic Graphs (DAG) and given information about the causal variables, one can learn the DAG with structure learning algorithms. Recently, there has been a flurry of works advancing structure learning algorithms (Shimizu et al., 2006; Zheng et al., 2018; He et al., 2019; Pamfil et al., 2020; Lorch et al., 2021; Annadani et al., 2021; Ng et al., 2021; Cundy et al., 2021; Deleu et al., 2022) that learn the structure of a DAG given data samples (of causal variables). Most of these works cast the discrete optimization of learning a DAG into a continuous one that is optimized through gradient descent, thereby sidestepping computational intractability arising from the super-exponential nature of DAG search in the discrete case. However, all approaches learn a causal DAG on the premise that one has full access to the true causal variables which might not be realistic. A more realistic assumption would be that we have partial or no access to true causal variables and that one has to infer the structure *along with* the causal variables.

Here, we use a fully differentiable latent variable model, Decoder BCD, to study the problem of Bayesian structure learning in linear Gaussian additive noise models, from high dimensional data. We perform synthetic experiments to analyze why unsupervised causal discovery in latent variable models is difficult. Section 2 explains preliminaries for the setup and section 3 gives the problem setup. In section 4, we introduce Decoder BCD, a decoder model for Bayesian Causal Discovery in the latent space before discussing experiments and our findings in section 5. We discuss related work in section 6 before concluding in section 7.

2. Preliminaries

Structural Causal Models (SCM): We operate in the framework of SCM (Pearl, 2009) where node Z_i represents a random causal variable with an independent noise variable $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, and parents $Pa_G(Z_i)$ corresponding to a DAG G. We focus primarily on the family of linear Gaussian additive noise models. However, for a DAG to be identifiable from data, one either has to observe a non-Gaussian setting, or, in the case of a Gaussian one, have an equal noise variance assumption (all $\sigma_i = \sigma$) (Peters & Buhlmann, 2014). Since we are in the Gaussian setting, we assume the latter. Thus, we have d causal variables $Z = [Z_1, ...Z_d]$, and a joint distribution entailed by the DAG G such that,

$$P(Z_1, ...Z_d) = \prod_{i=1}^{d} P(Z_i | Pa_G(Z_i)) \tag{1}$$

and the exact values of the d random variables are given by $z_i = f(Pa_G(z_i)) + \epsilon_i$, where $f(.)$ is a linear function. In our case, f is the weighted sum of values taken on

*Work done while visiting Mila [1]Mila - Québec AI Institute [2]McGill University, Montréal [3]KTH, Stockholm [4]ÉTS Montréal [5]CIFAR AI Chair. Correspondence to: Jithendaraa Subramanian <jithen.subra@gmail.com>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2022 by the author(s).

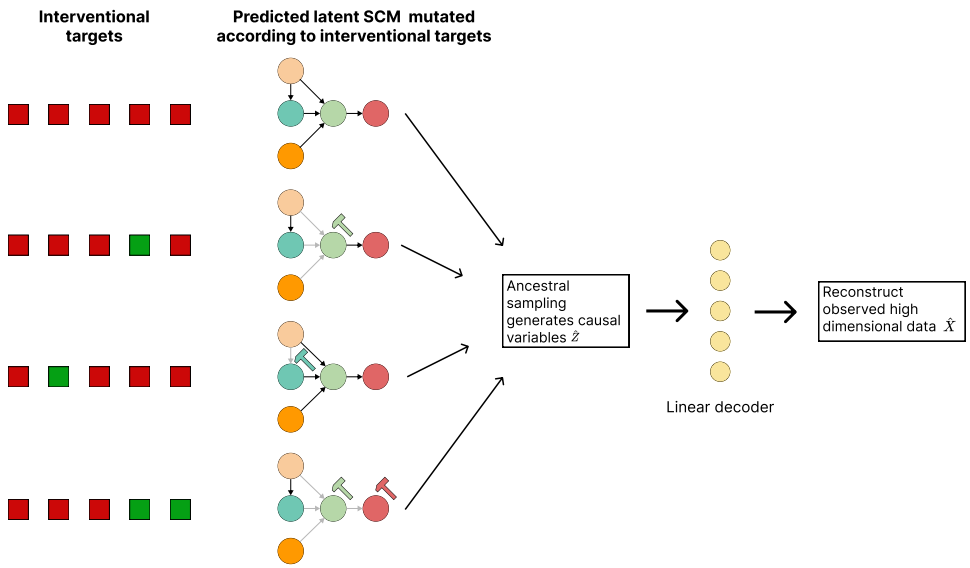

Figure 1. An illustration of the latent variable (decoder) model for Bayesian causal discovery

by random variables $Pa_G(Z_i)$, the weights given by the weighted adjacency matrix W, such that $z = W^T z + \epsilon$.

Bayesian Causal Discovery Nets (BCD Nets): We build Decoder BCD, our latent variable decoder model, upon BCD Nets (Cundy et al., 2021). Given samples of true causal variables, BCD Nets is a Bayesian structure learning method that obtains a posterior distribution over causal structures that best explains the data. Similar to other structure learning works (Lorch et al., 2021), BCD Nets introduces the problem of structure learning as a continuous constrained optimization problem. However, the DAG is parameterized such that one always ends up with a DAG and is therefore a hard constraint (in contrast to DiBS which has a soft DAG constraint). This is achieved by formulating the weighted adjacency matrix as $W = (PLP^T)^T$ where P is a permutation matrix and L is strictly lower triangular. When P is identity, this is equivalent to having a DAG of fixed ordering with each node j having its possible parents only in the node range $[j+1, d]$. P allows one to transition between node orderings by permuting the rows and columns of L. Apart from estimating W (via P and L), BCD Nets also infers Σ, the noise covariance, for the noise variables on each node in the DAG.

Thus, overall, BCD Nets formulates the Bayesian Structure Learning problem as inference of P, L, and Σ with a unique factorization of the posterior as $q_\phi(P, L, \Sigma) = q_\phi(P|L, \Sigma) \cdot q_\phi(L, \Sigma)$. The model is trained on an ELBO loss (eq. 2) with a horseshoe prior on L, Gumbel Sinkhorn prior (Jang et al., 2016) on P, and a Gaussian prior on Σ.

$$\mathbb{E}_{(L,\Sigma)\sim q_\phi}\Big[\mathbb{E}_{P\sim q_\phi(.|L,\Sigma)}\big[\log p(X|P,L,\Sigma) - \log \frac{q_\phi(P|L,\Sigma)}{p(P|L,\Sigma)}\big] - \log \frac{q_\phi(L,\Sigma)}{p(L,\Sigma)}\Big] \tag{2}$$

For finer details, we refer the reader to the original work (Cundy et al., 2021). Our work focuses more on *extending* BCD Nets to the high dimensional setting and studying unsupervised graph recovery in the latent space.

3. The Problem Setup

This work revolves around the Bayesian inference of causal variables Z and the causal structure G. Given n samples of high dimensional data $X \in \mathbb{R}^{n\times D}$, we wish to recover a distribution over graph structures G – the (weighted) adjacency matrix – and the causal variables $Z \in \mathbb{R}^{n\times d}$. Our setup revolves mostly around the recovery in linear isotropic Gaussian additive noise SCM, which is identifiable. Models like DiBS (Lorch et al., 2021), VCN (Annadani et al., 2021), and BCD Nets (Cundy et al., 2021) can recover the Ground Truth (GT) DAG given only observational and interventional data. Given observational data, recovery is possible up to a Markov Equivalence Class (MEC).

We first generate a random ER (Erdös & Rényi, 1959) DAG — with weighted adjacency matrix W_{GT} with sparsity pattern G_{GT}, the adjacency matrix — and consider this the ground truth and set the noise covariance to be $\Sigma_{GT} = \sigma^2_{GT} I$, since we have an isotropic Gaussian assumption. In our experiments, σ_{GT} is usually set to 0.1.

Data generation of the true causal variables: The data generation is done by an ancestral sampling process, compactly given by $z = W^T z + \epsilon$. One could also get interventional data from this setup: **(i)** Choose the node set N being intervened upon, **(ii)** for every node in N, zero out the particular column in W to get the mutated DAG $\tilde{W}$ and **(iii)** perform ancestral sampling using $z = \tilde{W}^T z + \epsilon$. This process is repeated multiple times to get n samples of z, and we will then organize these samples into a (n, d) matrix and term it z_{GT}, or simply, z [1].

Generating the true high dimensional data: We assume the observed low level data is a linear projection of the causal variables given by $X_{GT} = z_{GT} P'$, where $P' \in \mathbb{R}^{d \times D}$ is a projection matrix and we have $D >> d$ in the real world. For most of the upcoming experiments, we will study recovery in the (simpler) limiting case where d equal to D. This work studies recovery in latent space in a case where direct access to true causal variables is not given, in contrast to existing structure learning works.

4. Decoder BCD

BCD Nets performs Bayesian inference over P, L, Σ given samples of true causal variables to best explain the data by training an ELBO loss (see eq. 2). Decoder BCD tries to infer the decoder parameters in addition to inferring P, L, Σ. It is trained over X_{GT} instead of over z_{GT} as in BCD Nets. Thus, we relax the assumption that we have access to samples of true causal variables. We can only access the high dimensional data that has to be explained by (Z, G) and we have to fit a structure, $\hat{G}$, and estimate edge weights, $\hat{W}$, to fit our best guess of the causal variables, $\hat{z}$. Algorithm 1 summarizes the inference mechanism of Decoder BCD. A diagrammatic overview is given in Figure 1.

Algorithm 1 Decoder BCD for causal discovery from high dimensional data

1. Initialize random distributions for P, L, Σ
2. For **train_steps**:

(i) Sample $\hat{P}, \hat{L}, \hat{\Sigma} \sim q_\phi(P, L, \Sigma)$
(ii) $\hat{W} = (\hat{P}\hat{L}\hat{P^T})^T$
(iii) Perform ancestral sampling: $\hat{z} = \hat{W}^T \hat{z} + \epsilon$; $\epsilon \sim \mathcal{N}(0, \hat{\sigma}^2)$ and $\hat{\Sigma} = \hat{\sigma}^2 I$
(iv) Decode $\hat{z}$ to obtain $\hat{X}$
(v) Update parameters of the distribution $P(P, L, \Sigma)$ with loss as $MSE(X, \hat{X})$
(vi) For supervised experiments, add an additional KL loss between true and posterior observational joint: $KL(q(z_1, ...z_d)||p(z_1, ...z_d))$

[1] These variables will also be referred to, at times, as samples of true causal variables

5. Experiments and Findings

For all our experiments, for simplicity, we will stick to just learning the decoder and inferring the edge matrix L since this makes the optimization simpler for our studies. Such an assumption of fixing the permutation P to the GT, and thereby, the node orderings, is not unreasonable (He et al., 2019). Here, the focus is solely on inferring the edges L in latent space whilst learning a decoder. In all experiments, we train the model for 5000 steps across 20 random seeds, with a learning rate of 0.002 on ER-2 DAGs. We consider the case of the higher dimensional data being $D = 10$ dimensions that is generated by data from a $d = 6$ node underlying SCM.

Metrics: In our experiments, we refer to the expected Structural Hamming Distance across 64 samples of the inferred DAG as SHD, $MSE(L, \hat{L})$ is the MSE between predicted $\hat{L}$ and L_{GT}, AUROC (a value of 0.5 denotes a random baseline with null edges), and KL(true || learned) refers to the KL divergence between the posterior observational joint and the GT observational joint distributions.

5.1. Learning edge matrix L with supervision

For the supervised experiments, we add an additional KL loss on the inferred posterior observational joint $q(z_1, ...z_d)$ and the prior observational GT joint distribution, $p(z_1, ...z_d) \sim \mathcal{N}(\mu_z, \Sigma_z)$, where Σ_z is calculated with W_{GT} instead of with $\hat{W}$. The estimation of the prior and posterior observational joint distribution is detailed in A.1.

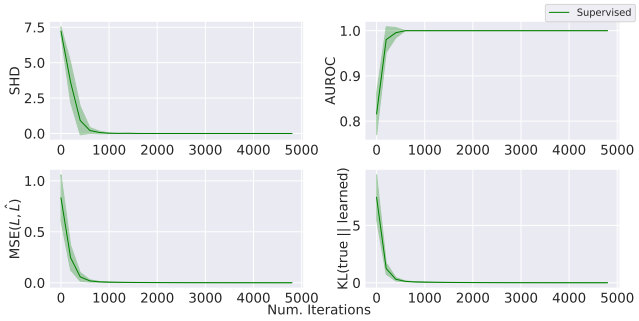

Figure 2. Supervised learning of L on 600 observational data points with $d = 6, D = 10$

Finding 1: From figure 2, we can see the expected SHD approaching 0 as KL(true || learned) and $MSE(L, \hat{L})$ approach 0. Using a KL over the observational joint distribution results in complete graph recovery in the supervised case. This is expected since we provide a mild signal for the model to uncover the true causal variables. Instead of providing prior over the samples of true causal variables, we use the true observational joint distribution as a signal.

5.2. Unsupervised learning of a single edge weight

For this setting, instead of inferring the whole lower triangular edge matrix L, we infer only the last edge at position $(d, d-1)$. The other elements of the matrix are fixed to the GT and we observe graph recovery in this case. This subsection is split into two parts to study recovery with (i) observational data and (ii) a mix of observational and interventional data, to analyze the effects of interventional data.

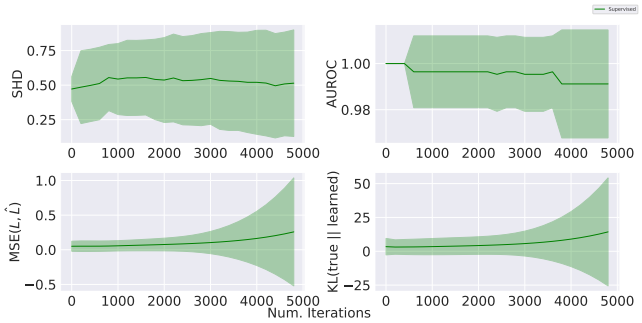

Figure 3. Unsupervised learning of a single edge weight with 1800 observational data points

We use 1800 observational data points for case (i) and 1800 data points (50-50 split of observational and interventional data) for case (ii). The interventional data generation process for single node and multi node interventions is detailed in A.2. The interventional values are fixed to 100.

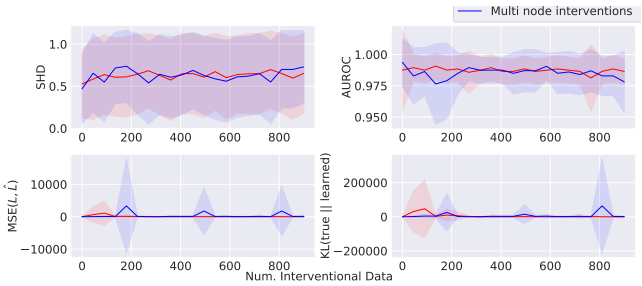

Figure 4. Unsupervised learning of a single edge weight with 900 observational and 900 interventional data points

Finding 2: Figures 3 and 4 reveal that using observational and/or interventional data with single node or multi node interventions with fixed intervention values is not sufficient to learn to orient a single edge in the unsupervised case.

5.3. Unsupervised learning of edge weight L

In this experiment and in the next, we explore the problem of learning the entire lower triangular edge matrix L in an unsupervised setting. First, we consider the learning problem with various amounts of observational data to analyze its effect on edge recovery in the latent space, which is shown in figure 5.

Figure 5. Unsupervised learning of the edge weight matrix with various amounts of observational data

Figure 5 shows that all metrics diverge with training but there is no trend with respect to the amount of observational data that the model is given.

Figure 6. Unsupervised learning of edge weights with random single node and multi node interventions

We now consider learning L using interventional targets as labels to learn the structure in the latent space with a mix of 300 observational data points and 3300 interventional data points. For this experiment, we retrain Decoder BCD multiple times from scratch — each time with the same 300 observational data points but with more interventional data than the previous run. This helps us understand the usefulness of interventional data for graph recovery. This result is illustrated in figure 6. For the interventional data points, we chose to use a fixed intervention value of 100.0. The reason for this particular value was that we had to choose a value that is far from 0 – the mean of all the nodes in the causal graph. Note that in a linear gaussian additive noise SCM, if one has 0 mean of the error variables ϵ, then all nodes in the graph have 0 mean.

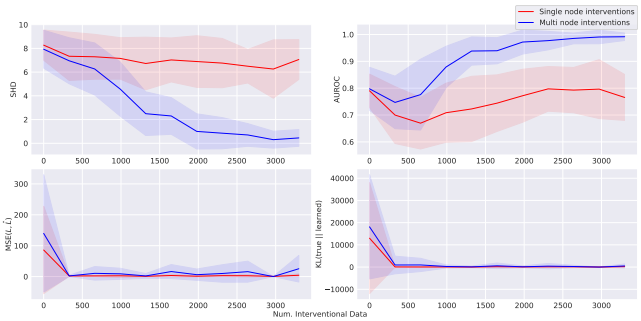

Figure 7. Unsupervised learning of edge weights with uniform single node and multi node interventions

Finally, other than learning from randomly (single or multi) intervened nodes, we also performed an experiment to observe the effect of randomly chosen interventional values, rather than a fixed interventional value. Thus, instead of an intervention value of 100, we randomly sampled interventional values in Uniform$(-10.0, 10.0)$ for each data point and repeated our previous experiment for the same amount of observational (300) and interventional (3300) data points. We summarize our findings for this experiment in figure 7.

Finding 4: Figures 6 and 7 show that both single node and multi node interventions help in recovering the edge weights, measured across all the 4 metrics. However, multi node interventions with uniformly sampled intervention values results in the better inference of the structure and parameters of the latent SCM.

6. Related Works

To address the challenges of causal discovery, a variety of methods have been proposed. Some of these methods are based on structure learning using observational data and some take in to account interventional data (Lorch et al., 2021; Scherrer et al., 2021; Ke et al., 2019; Brouillard et al., 2020).

There has been an increasing focus on Bayesian structure learning (Yu et al., 2019; Annadani et al., 2021; Lorch et al., 2021; Cundy et al., 2021; Deleu et al., 2022) to quantify epistemic uncertainty that is crucial for reinforcement learning and active learning settings. Charpentier et al. (2022) follow almost exactly the same approach as Cundy et al. (2021) except that they operate on nonlinear Gaussian SCM instead of a linear one. There also exist many maximum likelihood based methods, one such example is (He et al., 2019). It is one of the few works that learn a structure in the latent space but they do not operate in a causality-based or SCM framework.

Markov Chain Monte Carlo (MCMC) is a popular technique for sampling from complex high dimensional probability distributions, such as the posterior distribution of DAGs. (Madigan et al., 1995) uses Metropolis-Hastings (Metropolis et al., 1953) to predict the posterior distribution through Markov space to perform single edge addition or deletion. (Eaton & Murphy, 2007) propose a hybrid MCMC algorithm that uses an exact score based algorithm. (Kuipers et al., 2022) and (Viinikka et al., 2020) use more efficient MCMC samplers. (Deleu et al., 2022) uses a novel class of probabilistic models, GFlowNets (Bengio et al., 2021), which model distribution over discrete entities like DAGs to approximate the posterior in place of MCMC algorithms. Section A.3 discusses related work in more detail. Finally, our work involved randomly selecting nodes to intervene on while also randomly selecting the values for the interventions. (Tigas et al., 2022) uses a mutual information objective to learn where (which nodes) and how (with what values) to perform interventions in an active learning scenario to recover the edges more efficiently. However, unlike ours, their SCM is not in the latent space.

7. Conclusion

In this work, we introduced our latent variable model, Decoder BCD, and studied the causal representation learning problem. We explored the cases where edge recovery fails – learning to orient a single edge or learning with only observational data. To address this, we propose using interventional targets as labels to allow recovery of edges and edge weights in an SCM. Our experiments show that this is a promising direction for the unsupervised Bayesian causal discovery in latent space. However, our hypothesis of the observed data having a latent linear SCM and linear projection of the latent causal variables to higher dimension is a limitation (refer A.5) when it comes to mechanisms in the real world. Future work should explore nonlinear projections of the causal variables as well as nonlinear and non-Gaussian SCMs. Finally, we discuss some key challenges for future study in Appendix A.4.

Acknowledgements

The authors are grateful to Nan Rosemary Ke, Anirudh Goyal, Tristan Deleu and Sébastien Lachapelle for fruitful discussions and feedback. The authors are also thanful to Compute Canada and CIFAR for the compute and resources that made this work possible.

References

Annadani, Y., Rothfuss, J., Lacoste, A., Scherrer, N., Goyal, A., Bengio, Y., and Bauer, S. Variational causal networks: Approximate bayesian inference over causal structures, 2021.

Bengio, E., Jain, M., Korablyov, M., Precup, D., and Bengio, Y. Flow network based generative models for non-iterative diverse candidate generation, 2021. URL https://arxiv.org/abs/2106.04399.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Charpentier, B., Kibler, S., and Günnemann, S. Differentiable dag sampling, 2022. URL https://arxiv.org/abs/2203.08509.

Cundy, C., Grover, A., and Ermon, S. Bcd nets: Scalable variational approaches for bayesian causal discovery. In *Advances in Neural Information Processing Systems 34 pre-proceedings (NeurIPS)*, 2021.

Deleu, T., Góis, A., Emezue, C., Rankawat, M., Lacoste-Julien, S., Bauer, S., and Bengio, Y. Bayesian structure learning with generative flow networks, 2022. URL https://arxiv.org/abs/2202.13903.

Eaton, D. and Murphy, K. Exact bayesian structure learning from uncertain interventions. In *Artificial intelligence and statistics*, pp. 107–114. PMLR, 2007.

Erdös, P. and Rényi, A. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

Geiger, D. and Heckerman, D. Learning gaussian networks. In *Uncertainty Proceedings 1994*, pp. 235–243. Elsevier, 1994.

Ghoshal, A. and Honorio, J. Learning linear structural equation models in polynomial time and sample complexity, 2017.

He, J., Gong, Y., Marino, J., Mori, G., and Lehrmann, A. Variational autoencoders with jointly optimized latent dependency structure. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJgsCjCqt7.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax, 2016. URL https://arxiv.org/abs/1611.01144.

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Schölkopf, B., Mozer, M. C., Pal, C., and Bengio, Y. Learning neural causal models from unknown interventions, 2019. URL https://arxiv.org/abs/1910.01075.

Kuipers, J., Suter, P., and Moffa, G. Efficient sampling and structure learning of bayesian networks. *Journal of Computational and Graphical Statistics*, pp. 1–12, 2022.

Loh, P.-L. and Bühlmann, P. High-dimensional learning of linear causal networks via inverse covariance estimation, 2013.

Lorch, L., Rothfuss, J., Schölkopf, B., and Krause, A. Dibs: Differentiable bayesian structure learning, 2021.

Madigan, D., York, J., and Allard, D. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pp. 215–232, 1995.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi: 10.1063/1.1699114. URL http://link.aip.org/link/?JCP/21/1087/1.

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

Ng, I., Ghassami, A., and Zhang, K. On the role of sparsity and dag constraints for learning linear dags, 2021.

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Beaumont, P., Georgatzis, K., and Aragam, B. Dynotears: Structure learning from time-series data, 2020.

Pearl, J. *Causality*. Cambridge university press, 2009.

Peters, J. and Buhlmann, P. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014. ISSN 00063444. URL http://www.jstor.org/stable/43305605.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Scherrer, N., Bilaniuk, O., Annadani, Y., Goyal, A., Schwab, P., Schölkopf, B., Mozer, M. C., Bengio, Y., Bauer, S., and Ke, N. R. Learning neural causal models with active interventions, 2021. URL https://arxiv.org/abs/2109.02429.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Towards causal representation learning, 2021. URL https://arxiv.org/abs/2102.11107.

Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), jun 2020. doi:

10.1214/19-aos1857. URL https://doi.org/10.1214%2F19-aos1857.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Tigas, P., Annadani, Y., Jesson, A., Schölkopf, B., Gal, Y., and Bauer, S. Interventions, where and how? experimental design for causal models at scale, 2022. URL https://arxiv.org/abs/2203.02016.

Viinikka, J., Hyttinen, A., Pensar, J., and Koivisto, M. Towards scalable bayesian learning of causal dags. *Advances in Neural Information Processing Systems*, 33: 6584–6594, 2020.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. Causalvae: Disentangled representation learning via neural structural causal models, 2020. URL https://arxiv.org/abs/2004.08697.

Yu, Y., Chen, J., Gao, T., and Yu, M. Dag-gnn: Dag structure learning with graph neural networks, 2019.

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning, 2018.

A. Appendix

A.1. KL Loss for the mildly supervised experiments

Given $\hat{W} = (P\hat{L}P^T)^T$, from inferred $\hat{L}$, one can obtain the mean and covariance of the observational joint distribution $q(z_1, ...z_d)$ as follows:

$$z = \hat{W}^T z + \epsilon; \quad \epsilon \sim \mathcal{N}(0, \sigma) \tag{3}$$

$$z = (I - \hat{W})^{-T}\epsilon \tag{4}$$

$$q(z_1, ...z_d) \sim \mathcal{N}(\hat{\mu_z}, \hat{\Sigma_z}) \tag{5}$$

$$\hat{\mu_z} = 0 \; and \; \hat{\Sigma_z} = (I - \hat{W})^{-T}\Sigma(I - \hat{W})^{-1} \tag{6}$$

To estimate the prior GT observational joint distribution, one would use W in place of $\hat{W}$ in equation 6.

A.2. Generating interventional data for experiments that use single node and multi node interventions

Suppose we have to generate i interventional data points. We split the data generation process into $s = 20$ sets, each set generating i/s interventional data points. *For single node interventions*, we randomly choose a node and sample i/s data points. The process is repeated s times randomly to generate the i data points. *For multi node interventions*, we randomly choose a number between [2, d] to decide on the number of nodes to intervene on (call this x). We then choose x nodes without replacement and perform the interventions on these nodes and sample i/s data points. The process is repeated s times randomly to generate the i data points.

A.3. More Related Work

Since discrete optimization is hard and often involves enumeration of possible structures, the super-exponential nature of structure learning has resulted in the community resorting to relaxing the discrete optimization problem into a continuous one (Lorch et al., 2021; Cundy et al., 2021; Annadani et al., 2021; Scherrer et al., 2021; Ke et al., 2019; Zheng et al., 2018) and learning the parameters using gradient descent. (Loh & Bühlmann, 2013) propose a scalable, scoring-based DAG learning approach to recover high dimensional, sparse causal graphs in a non-Gaussian setting where only some but not all exogenous noise variables are expected to be non-Gaussian. (Ghoshal & Honorio, 2017) learns a linear structural equation model in polynomial time. (Ke et al., 2019) learns the causal structure from unknown interventions but operates on the Bernoulli distribution while (Scherrer et al., 2021) is in an active learning framework and the system determines the intervention that will be most useful in gaining knowledge about the graph structure. (Yang et al., 2020) proposes a variational autoencoder parameterised by exogenous variables to learn causal semantics of the data. Another family of works introduce assumptions to functional and parametric form of the data-generation structure. They exploit symmetries to learn the causal structure (Peters et al., 2017; Mooij et al., 2016).

Approaches to the problem are mostly employ score-based or constraint-based optimization. Most modern methods use some sort of a scoring function to rank estimated structure and use it to rank structures and optimizing for the score is expected to return the ground truth DAG. Popular scoring functions include Bayesian Information Criterion (BIC) and Bayesian Gaussian Equivalent marginal likelihood score (Geiger & Heckerman, 1994). These methods typically use a regularization over the structure to induce sparsity and/or acyclicity. Some methods impose hard constraints as well that ensure the search is done only over the space of DAGs. (Shah & Peters, 2020) is a constraint based approach that tests for conditional independence.

A.4. Key challenges for future study

One of the most important scientific questions of causal representation learning is regarding the **relationship between high dimensional, observed variables and the low dimensional, causal variables**: In this work, we perform synthetic data generation of z and project it to higher dimensions by using a random projection matrix P'. We begin on the premise that

real-world, high dimensional data can be explained by a few causal variables and the inferring these variables and their structure is the problem of causal inference that the brain solves for performing intelligent tasks. Thus, there must exist an operation that maps the low dimensional causal variables to the high dimensional, observed variables (eg. images, videos). For our problem setting, we assume this is true and try to generate high dimensional samples that are "causally consistent" by performing a linear projection $X = zP'$. However, we do not know how this process of projection to higher dimensions might happen in reality.

What's the right loss function for unsupervised causal discovery? In all experiments, we found that the MSE over high dimensional data, X, goes down but this does not necessarily mean that graph recovery in the latent space gets better. Therefore, we need to look for alternative losses with a property such that reduction in loss over X guarantees a better recovery in the latent space (i.e., better graph structure recovery or better estimates of edge weight matrix L). Ideally, such a loss should result in a reduction in the KL divergence between the inferred posterior observational joint distribution and the GT observational joint distribution. We propose that this a better metric to measure since in the supervised experiments, getting a low enough value of this metric results in the SHD dropping steeply to 0.

A.5. Limitations

A limitation of this work is that we do not know if it is practical to assume a linear projection – it is just a formulation that we explore. Additionally, if it *is* a linear operation, are there any properties that the projection matrix P' must hold to maintain this "causal consistency" in higher dimensions? If P' needs to hold some properties for causal inference to be performed from high dimensions, *what* exactly are these properties? It is easy to see that a random projection matrix (which transforms a d-dimensional vector to D-dimensional vector) can be random enough to completely destroy the encoded information due to the causal generation process that occurred in the lower dimensions, and thus the high dimensional data could no longer be "causally consistent" for us to perform inference. And finally, one needs to focus on the question of whether the projection operation could be nonlinear.