Differentiable Causal Discovery Under Unmeasured Confounding

Rohit Bhattacharya
Johns Hopkins University
rbhattacharya@jhu.edu

Tushar Nagarajan
University of Texas at Austin
tushar@cs.utexas.edu

Daniel Malinsky
Columbia University
d.malinsky@columbia.edu

Ilya Shpitser
Johns Hopkins University
ilyas@cs.jhu.edu

Abstract

The data drawn from biological, economic, and social systems are often confounded due to the presence of unmeasured variables. Prior work in causal discovery has focused on discrete search procedures for selecting acyclic directed mixed graphs (ADMGs), specifically ancestral ADMGs, that encode ordinary conditional independence constraints among the observed variables of the system. However, confounded systems also exhibit more general equality restrictions that cannot be represented via these graphs, placing a limit on the kinds of structures that can be learned using ancestral ADMGs. In this work, we derive differentiable algebraic constraints that fully characterize the space of ancestral ADMGs, as well as more general classes of ADMGs, arid ADMGs and bow-free ADMGs, that capture all equality restrictions on the observed variables. We use these constraints to cast causal discovery as a continuous optimization problem and design differentiable procedures to find the best fitting ADMG when the data comes from a confounded linear system of equations with correlated errors. We demonstrate the efficacy of our method through simulations and application to a protein expression dataset. Code implementing our methods is open-source and publicly available at `https://gitlab.com/rbhatta8/dcd` and will be incorporated into the `Ananke` package.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 INTRODUCTION

Biological, economic, and social systems are often affected by unmeasured (latent) variables. In such scenarios, statistical and causal models of a directed acyclic graph (DAG) over the observed variables do not faithfully capture the underlying causal process. The most popular graphical structures used to summarize constraints on the observed data distribution are a special class of acyclic directed mixed graphs (ADMGs) with directed and bidirected edges, known as ancestral ADMGs (Richardson and Spirtes, 2002).

Ancestral ADMGs capture all ordinary conditional independence constraints on the observed margin, but they do not capture more general non-parametric equality restrictions, commonly referred to as Verma constraints (Verma and Pearl, 1990; Tian and Pearl, 2002; Robins, 1986). While ADMGs without the ancestral restriction are capable of capturing all such equality constraints (Evans, 2018a), the associated parametric models are not guaranteed to form smooth curved exponential families with globally identifiable parameters – an important pre-condition for score-based model selection. A smooth parameterization for arbitrary ADMGs is known only when all observed variables are either binary or discrete (Evans and Richardson, 2014). For the common scenario when the data comes from a linear Gaussian system of structural equations, the statistical model of an ADMG is almost-everywhere identified if the ADMG is bow-free (Brito and Pearl, 2002), and is globally identified and forms a smooth curved exponential family if and only if the ADMG is arid (Drton et al., 2011; Shpitser et al., 2018). From a causal perspective, arid and bow-free ADMGs, like ancestral ADMGs, have the desirable property of preserving ancestral relationships in the underlying latent variable DAG, while also capturing all non-parametric equality restrictions on the observed margin (Shpitser et al., 2018).

We introduce a structure learning procedure for selecting arid, bow-free, or ancestral ADMGs from observational data. Our learning approach is based on reformulating the usual discrete combinatorial search problem into a more tractable constrained continuous optimization program. Such a reformulation was first proposed by Zheng et al. (2018) for the special case when the search space is restricted to DAGs. Subsequent extensions such as Yu et al. (2019), Zhang et al. (2019), and Zheng et al. (2020) also restrict the search space in a similar fashion. In this work, we derive differentiable algebraic constraints on the adjacency matrices of the directed and bidirected portions of an ADMG that fully characterize the space of arid ADMGs. We also derive similar algebraic constraints that characterize the space of ancestral and bow-free ADMGs that are quite useful in practice and connect our work to prior methods. Having derived these differentiable constraints, we select the best fitting graph in the class by optimizing a penalized likelihood-based score. While the constraints we derive in this paper are non-parametric, we focus our causal discovery methods on distributions that arise from linear Gaussian systems of equations.

Causal discovery methods for learning ancestral ADMGs from data are well developed (Spirtes et al., 2000; Colombo et al., 2012; Ogarrio et al., 2016), but procedures for more general ADMGs are understudied. Hyttinen et al. (2014) propose a constraint-based satisfiability solver approach for mixed graphs with cycles. However, their proposal relies on an independence oracle that does not address how to perform valid statistical tests for arbitrarily complex equality restrictions and their procedure may lead to models where the corresponding statistical parameters are not identified (so goodness-of-fit cannot be evaluated). A score-based approach to discovery for linear Gaussian bow-free ADMGs was proposed in Nowzohour et al. (2017). Their method relies on heuristics that may lead to local optima and is not guaranteed to be consistent. Similar issues are faced by the method in Wang and Drton (2020), which makes a linear non-Gaussian assumption. Currently, there does not exist any consistent fully score-based procedure for learning general ADMGs (besides exhaustive enumeration which is intractable); there are greedy algorithms (Bernstein et al., 2020) and hybrid greedy algorithms (Ogarrio et al., 2016) for ancestral ADMGs, but these are computationally intensive due to the large discrete search space and extending these to arid or bow-free ADMGs would be non-trivial. The procedure we propose has the benefit of being easy to adapt to either ancestral, arid, or bow-free ADMGs while avoiding the need to solve a complicated discrete search problem, instead exploiting state-of-the-art advances in continuous optimization.

Our structure learning procedure for arid and ancestral graphs is consistent in the following sense: asymptotically, convergence to the global optimum implies that the corresponding ADMG is either the true model or one that belongs to the same equivalence class. That is, if the optimization procedure succeeds in finding the global optimum, the resulting graph is either the true underlying structure or one that implies the same set of equality constraints on the observed data. While the L_0-regularized objective we propose is non-convex and so our optimization scheme may result in local optima, we show via experiments and application to protein expression data that our proposal works quite well in practice. We believe the algebraic constraints on their own are also valuable for further research at the intersection of non-convex optimization techniques for L_0-regularization and causal discovery.

We begin with a motivating example and background on the structure learning problem for partially-observed systems in Sections 2 and 3. In Section 4 we derive differentiable algebraic constraints that characterize arid, bow-free, and ancestral ADMGs. In Section 5 we use these to formulate the first (to our knowledge) tractable method for learning arid ADMGs from observational data, by extending the continuous optimization scheme of causal discovery. Simply by modifying the constraint in the optimization program, the same procedure may also be leveraged to learn bow-free or ancestral graphs. Finally we evaluate the performance of our algorithms in simulation experiments and on protein expression data in Section 6.

2 MOTIVATING EXAMPLE

To motivate our work, we present an example of how our method may be used to reconstruct complex interactions in a network of genes, which is related to the data application we present in Section 6.

Consider a scenario in which an analyst has access to gene expression data on four genes: A, B, C, and D. Assume that the analyst is confident (due to prior analysis or background knowledge) about the structure corresponding to non-dashed edges shown in Fig. 1(a), i.e., that A regulates C and B regulates D but A and B are independent. This leaves an important ambiguity regarding regulatory explanations of co-expression of genes C and D.

An observed correlation between C and D may be explained in different ways that provide very different mechanistic interpretations. If the hypothesis class is restricted to DAGs, the only explanations available to the analyst are that C is a cause of D or vice-versa as

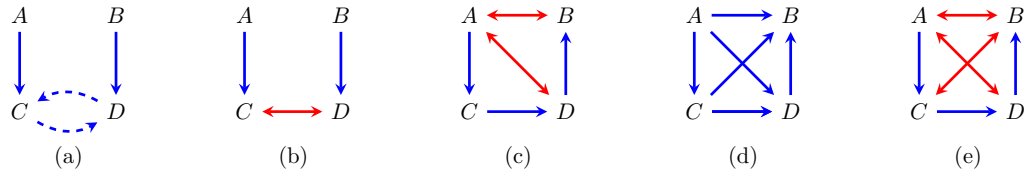


Figure 1: (a) A DAG if $C \to D$ or $D \to C$ exists but not both. (b) An ADMG that posits an unmeasured confounder between C and D. (c) An (arid) ADMG encoding a Verma constraint between C and B. (d) The ancestral version of (c). (e) A non-arid bow-free ADMG that is a super model of (c).

shown in Fig. 1(a). If the analyst proceeds with either of these explanations and performs a gene-knockout experiment where C (or D) is removed but sees no change in D (respectively C), then the causal DAG fails to be a faithful representation of the true underlying mechanism. The correlation may instead be explained by an ADMG as in Fig. 1(b) where $C \leftrightarrow D$ indicates that C and D are dependent due to the presence of at least one unmeasured confounding gene that regulates both of them. That is, if we had data on these unmeasured genes U the corresponding DAG would have contained a structure $C \leftarrow U \to D$. However, given observations only on A, B, C, D, Fig. 1(b) provides a faithful representation of this underlying mechanism on the observed variables. It correctly encodes that intervention on C or D has no downstream effects on the other.

Importantly, each of these different explanations are not just different from a mechanistic point of view but also imply different independence restrictions on the observed data. The two DAGs in Fig. 1(a) imply that $A \perp\!\!\!\perp D \mid C$ or $B \perp\!\!\!\perp C \mid D$ respectively, whereas Fig. 1(b) implies $A \perp\!\!\!\perp D$ and $B \perp\!\!\!\perp C$. Hence, a causal discovery procedure that seeks the best fitting structure from the hypothesis class of ADMGs, will be able to distinguish between these different explanations and choose the correct one.

Some mechanisms, such as the one shown in Fig. 1(c), are not distinguishable using ordinary conditional independence statements alone. In this graph, the only pair of genes with no edge between them is B and C. The absence of this edge implies that C does not directly regulate the expression of B and only does so through D. This missing edge does not correspond to any ordinary conditional independence (there are no independence constraints implied by the model at all), but does encode a Verma constraint, namely that $B \perp\!\!\!\perp C \mid D$ in a re-weighed distribution derived from the joint, $p(A, B, C, D)/p(C|A)$.

The following ADMG classes will be important in this work. An ADMG $\mathcal{G} = (V, E)$ is said to be *ancestral* if for any pair of vertices $V_i, V_j \in V$, a directed path $V_i \to \cdots \to V_j$ and bidirected edge $V_i \leftrightarrow V_j$ do not both appear in $\mathcal{G}$. An ADMG $\mathcal{G}$ is said to be *arid* if it does not contain any *c-trees*. A c-tree is a subgraph of $\mathcal{G}$ whose directed edges form an arborescence (the directed graph analogue of a tree) and bidirected edges form a single bidirected connected component within the subgraph. It is easy to confirm that the ADMG in Fig. 1(b) is ancestral while the one in Fig. 1(c) is arid but not ancestral. An ADMG is called *bow-free* if for any pair of vertices, $V_i \to V_j$ and $V_i \leftrightarrow V_j$ do not both appear in $\mathcal{G}$. A graph that is bow-free but neither arid nor ancestral is displayed in Fig. 1(e). The relation between these graph classes is the following:

$$\text{Ancestral} \subset \text{Arid} \subset \text{Bow-free}$$

Ancestral graphs can "hide" certain important information because they encode only ordinary conditional independence constraints. An ancestral graph that encodes the same ordinary independence constraints as the arid graph in Fig. 1(c) is shown in Fig. 1(d). It is a complete graph since there are no conditional independence constraints in Fig. 1(c). That is, the absence of any $C \to B$ edge in Fig. 1(c) is "masked" to preserve the ancestrality property. We can potentially learn a more informative structure if we do not limit our search space to the class of ancestral graphs.

3 GRAPHICAL INTERPRETATION OF LINEAR SEMs

In this section, we review linear SEMs and their graphical representations. We use capital letters (e.g. V) to denote sets of variables and nodes on a graph interchangeably and capital letters with an index (e.g. V_i) to refer to a specific variable or node in V. We also make use of the following standard matrix notation: A_{ij} refers to the element in the i^{th} row and j^{th} column of a matrix A, indexing $A_{-i,-j}$ refers to the sub matrix obtained by excluding the i^{th} row and j^{th} column of A, and $A_{:,i}$ refers to the i^{th} column of A.

3.1 Linear SEMs and DAGs

Consider a linear SEM on d variables parameterized by a weight matrix $\theta \in \mathbb{R}^{d\times d}$. For each variable $V_i \in V$, we have a structural equation $V_i \leftarrow \sum_{V_j \in V} \theta_{ji} V_j + \epsilon_i$, where the noise terms ϵ_i are mutually independent. That is, $\epsilon_i \perp\!\!\!\perp \epsilon_j$ for all $i \neq j$. Let $\mathcal{G}(\theta)$ and $D(\theta) \in \{0,1\}^{d\times d}$ be the induced directed graph and corresponding binary adjacency matrix obtained as follows: $V_i \to V_j$ exists in $\mathcal{G}(\theta)$ and $D(\theta)_{ij} = 1$ if and only if $\theta_{ij} \neq 0$. The induced graph $\mathcal{G}$ has no directed cycles if and only if θ can be made upper-triangular via a permutation of vertex labelings (McKay et al., 2004). Such an SEM is said to be *recursive* or *acyclic* and the corresponding probability distribution $p(V)$ is said to be Markov with respect to the DAG $\mathcal{G}(\theta)$. This means that conditional independence statements in $p(V)$ can be read off from $\mathcal{G}$ via the well-known d-separation criterion (Pearl, 2009).

3.2 Systems with Unmeasured Confounding

A set of observed variables is called *causally insufficient* if there exist unobserved variables, commonly referred to as latent confounders, that cause two or more observed variables in the system. In the linear SEM setting, unmeasured variables manifest as correlated errors (Pearl, 2009). Such an SEM on d variables can be parameterized by two real-valued matrices $\delta, \beta \in \mathbb{R}^{d\times d}$ as follows. For each $V_i \in V$, we have a structural equation $V_i \leftarrow \sum_{V_j \in V} \delta_{ji} V_j + \epsilon_i$, and the dependence between the noise terms $\epsilon = (\epsilon_1, ..., \epsilon_d)$ is summarized via their covariance matrix $\beta = \mathbb{E}[\epsilon\epsilon^T]$. In the case when each noise term ϵ_i is normally distributed the induced distribution $p(V)$ is jointly normal with mean zero and covariance matrix $\Sigma = (I-\delta)^{-T}\beta(I-\delta)^{-1}$. The induced graph $\mathcal{G}$ is a mixed graph consisting of directed ($\to$) and bidirected ($\leftrightarrow$) edges and can be represented via two adjacency matrices D and B. $V_i \to V_j$ exists in $\mathcal{G}$ and $D_{ij} = 1$ if and only if $\delta_{ij} \neq 0$. $V_i \leftrightarrow V_j$ exists in $\mathcal{G}$ and $B_{ij} = B_{ji} = 1$ if and only if $\beta_{ij} \neq 0$. That is, the adjacency matrix B corresponding to bidirected edges in $\mathcal{G}$ is symmetric as the covariance matrix β itself is symmetric (and positive definite).

We consider three classes of mixed graphs to represent causally insufficient linear SEMs: ancestral, arid, and bow-free ADMGs. All of these have no directed cycles and lack specific substructures as defined in the previous section. A distribution $p(V)$ induced by a linear Gaussian SEM is said to be Markov with respect to an ADMG $\mathcal{G}$ if absence of an edge between V_i and V_j implies $\delta_{ij} = \delta_{ji} = \beta_{ij} = \beta_{ji} = 0$ which in turn implies equality restrictions on the support of all possible covariance matrices $\Sigma(\mathcal{G})$ by forcing certain polynomial functions of entries in the covariance matrix to evaluate to 0 (Yao and Evans, 2019). To facilitate causal discovery, we assume a generalized version of faithfulness, similar to the one in Ghassami et al. (2020), stating that if a distribution $p(V)$ is induced by a linear Gaussian SEM where $\delta_{ij} = \delta_{ji} = \beta_{ij} = \beta_{ji} = 0$ then there is no edge present between V_i and V_j in $\mathcal{G}$. In other words, we define $p(V)$ to be Markov and faithful with respect to $\mathcal{G}$ if absence of edges in $\mathcal{G}$ occurs if and only if the corresponding entries in δ and β are 0.

As a concrete example, let Σ denote the covariance matrix of standardized normal random variables A, B, C, D drawn from a linear SEM that is Markov with respect to the ADMG in Fig. 1(c), and let δ and β denote the corresponding normalized coefficient matrices. By standard rules of path analysis (Wright, 1921, 1934), the Verma constraint due to the missing edge in Fig. 1(c) corresponds to the equality constraint:

$$\Sigma_{BC} - \delta_{CD}\delta_{DB} - \delta_{AC}\beta_{AB} - \delta_{AC}\beta_{AD}\delta_{DB} = 0.$$

Since entries in the covariance matrix are rational functions of δ and β, the above constraint can be re-expressed solely in terms of entries in Σ. Our faithfulness assumption is used to ensure that such polynomial functions of the covariance matrix do not "accidentally" evaluate to zero, and only do so due to a missing edge in the underlying ADMG.

As mentioned earlier, ancestral ADMGs cannot encode such generalized equality restrictions but arid and bow-free ADMGs can. For any ADMG $\mathcal{G}$, an arid ADMG that shares all non-parametric equality constraints with $\mathcal{G}$ may be constructed by an operation called maximal arid projection (Shpitser et al., 2018). We also consider bow-free ADMGs because the algebraic constraint characterizing the bow-free property is simpler than the one characterizing the arid property. Though the lack of global identifiability in bow-free ADMG models (only almost everywhere identifiable) can pose problems for model convergence, we confirm in our experiments that enforcing only the weaker bow-free property is often sufficient for accurate causal discovery in practice.

4 DIFFERENTIABLE ALGEBRAIC CONSTRAINTS

We now introduce differentiable algebraic constraints that precisely characterize when the parameters of a linear SEM induce a graph that belongs to any one of the ADMG classes described in the previous section. Our results are summarized in Table 1 in terms of the binary adjacency matrices but as we explain below, the results extend in a straightforward manner to real-valued matrices that parameterize a linear SEM. In Table 1, $A \circ B$ denotes the Hadamard (elementwise)

Algorithm 1 GREENERY (D, B)

1: greenery $\leftarrow 0$ and $I \leftarrow d \times d$ identity matrix
2: **for** i in $(1, \ldots, d)$ **do**
3: $\quad D_f, B_f \leftarrow D, B$
4: $\quad$ **for** j in $(1, \ldots, d-1)$ **do**
5: $\quad\quad t \leftarrow$ row sums of $e^{B_f} \circ D_f$ ▷ $1 \times d$ vector
6: $\quad\quad f \leftarrow \tanh(t + I_i)$ ▷ $1 \times d$ vector
7: $\quad\quad F \leftarrow [f^T; \ldots; f^T]^T$ ▷ $d \times d$ matrix
8: $\quad\quad D_f \leftarrow D_f \circ F$ and $B_f \leftarrow B_f \circ F \circ F^T$
9: $\quad C \leftarrow e^{D_f} \circ e^{B_f}$
10: $\quad$ greenery $+=$ sum$(C_{:,i})$ ▷ sum of i^{th} column
11: **return** greenery $- d$

ADMG	Algebraic Constraint
Ancestral	$\text{trace}(e^D) - d + \text{sum}(e^D \circ B) = 0$
Arid	$\text{trace}(e^D) - d + \text{GREENERY}(D, B) = 0$
Bow-free	$\text{trace}(e^D) - d + \text{sum}(D \circ B) = 0$

Table 1: Differentiable algebraic constraints that characterize the space of binary adjacency matrices that fall within each ADMG class. The GREENERY algorithm to penalize c-trees is described in Algorithm 1.

matrix product between A and B and e^A denotes the exponential of a square matrix A defined as the infinite Taylor series, $e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k$. We formalize the properties of our constraints in the following theorem.

Theorem 1. *The constraints shown in Table 1 are satisfied if and only if the adjacency matrices satisfy the relevant property of ancestrality, aridity, and bow-freeness respectively.*

We defer formal proofs to the Appendix but briefly provide intuition for our results. For a binary square matrix A, corresponding to a directed/bidirected adjacency matrix, the entry A^k_{ij} counts the number of directed/bidirected walks of length k from V_i to V_j; see for example Butler (2008). For $k = 0$, D^k is the identity matrix by definition and for $k \geq 1$, each diagonal entry of the matrix D^k appearing in the infinite series e^D thus corresponds to the number of directed walks of length k from a vertex back to itself, i.e., the number of directed cycles of length k. The quantity $\text{trace}(e^D) - d$ is therefore a weighted count of the number of directed cycles in the induced graph and is zero precisely when no such cycles exist. Hence, this term appears in all algebraic constraints presented in Table 1 as requiring $\text{trace}(e^D) - d = 0$ enforces acylicity.

Similar reasoning can be used to show that requiring $\text{sum}(e^D \circ B) = 0$ enforces ancestrality. An entry i, j of the matrix $D^k \circ B$ appearing in the infinite series counts the number of violations of ancestrality due to a directed path from V_i to V_j of length k and a bidirected edge $V_i \leftrightarrow V_j$. The sum of all such terms is then precisely zero when the induced graph is ancestral. The bow-free constraint $\text{sum}(D \circ B) = 0$ is simply a special case of the ancestral constraint where directed paths of length ≥ 2 need not be considered.

C-trees are known to be linked to the identification of causal parameters, specifically, the effect of each variable's parents on the variable itself (Shpitser and Pearl, 2006; Huang and Valtorta, 2006). The outer loop of Algorithm 1 iterates over each vertex V_i to determine if there is a V_i-rooted c-tree. The inner loop performs the following recursive simplification at most $d - 1$ times. At each step, the sum of the j^{th} row of the matrix $e^{B_f} \circ D_f$ is zero if and only if there are no bidirected paths from V_j to any of its direct children. If this criterion – called primal fixability – is met, the effect of V_j on its children is identified and the post-intervention distribution can be summarized by a new graph with all incoming edges into V_j removed (Bhattacharya et al., 2020). Lines 6-8 are the algebraic operations that correspond to deletion of incoming directed and bidirected edges into primal fixable vertices, except V_i itself as it is the root node of interest. The hyperbolic tangent function is used to ensure that recursive applications of the operation do not result in large values. At the end of the recursion, the co-existence of directed and bidirected paths to V_i imply the existence of a c-tree. Hence, the quantity $\text{sum}(C_{:,i})$ is non-negative and is zero if and only if there is no V_i-rooted c-tree. Concrete examples of applying Algorithm 1, and its connections to primal fixing are provided in Appendix A.

It is easy to see that the above results and intuitions can be applied to arbitrary non-negative real-valued matrices D and B. Theorem 1 then extends in a straightforward manner to parameters of a linear SEM by noting that for any real-valued matrix A, the matrix $A \circ A$ is real-valued and non-negative.

Corollary 1.1. *The result in Theorem 1 and the constraints in Table 1 can be applied to linear SEMs by plugging in $D \equiv \delta \circ \delta$ and $B \equiv \beta' \circ \beta'$, where $\beta'_{ij} = \beta_{ij}$ for $i \neq j$ and 0 otherwise.*

Finally, while the matrix exponential makes theoretical arguments simple, the resulting constraints are not numerically stable as pointed out in Yu et al. (2019). The following corollary provides a more stable alternative that we use in our implementations.

Corollary 1.2. *The results in Theorem 1 and Corollary 1.1 hold if every occurrence of a matrix exponential e^A is replaced with the matrix power $(I + cA)^d$ for any $c > 0$, where I is the identity matrix.*

5 DIFFERENTIABLE SCORE BASED CAUSAL DISCOVERY

Let θ be the parameters of a linear SEM. We use θ here to refer to a generic parameter vector that can be reshaped into the appropriate parameter matrices δ, and β as discussed in Section 3. Let $\mathcal{G}(\theta)$ be the corresponding induced graph. Given a dataset $X \in \mathbb{R}^{n\times d}$ drawn from the linear SEM and a hypothesis class $\mathbb{G}$ that corresponds to one of ancestral, arid, or bow-free ADMGs, the combinatorial problem of finding an optimal set of parameters $\theta^* \in \Theta$ that minimizes some score $f(X;\theta)$ such that $\mathcal{G}(\theta) \in \mathbb{G}$ can be rephrased as a more tractable continous program.

$$\begin{aligned} &\min_{\theta\in\Theta} \ f(X;\theta) \\ &\text{s.t.} \ \ \mathcal{G}(\theta) \in \mathbb{G} \end{aligned} \quad \Longleftrightarrow \quad \begin{aligned} &\min_{\theta\in\Theta} \ f(X;\theta) \\ &\text{s.t.} \ \ h(\theta) = 0. \end{aligned} \tag{1}$$

The results in the previous section in Theorem 1, its Corollaries and Table 1 tell us how to pick the appropriate function $h(\theta)$ for each hypothesis class $\mathbb{G}$. We now discuss choices of score function $f(X;\theta)$ and procedures to minimize it for different hypothesis classes.

5.1 Choice of Score Function

Given a dataset $X \in \mathbb{R}^{n\times d}$, the Bayesian Information Criterion (BIC) is given by $-2\ln(\mathcal{L}(X;\theta)) + \ln(n)\sum_{i=1}^{\dim(\theta)} \mathbb{I}(\theta_i \neq 0)$, where $\mathcal{L}(\cdot)$ is the likelihood function and $\dim(\theta)$ is the dimensionality of θ. The BIC is consistent for model selection in curved exponential families (Schwarz, 1978; Haughton, 1988), i.e., as $n \to \infty$ the BIC attains its minimum at the true model (or one that is observationally equivalent to it). This results in the following desirable theoretical property when the BIC is used as our objective function.

Theorem 2. *Let $p(V;\theta^*)$ be a distribution in the curved exponential family that is Markov and faithful with respect to an arid ADMG $\mathcal{G}^*$. Finding the global optimum of the continuous program in display (1) with $f \equiv BIC$ yields an ADMG $\mathcal{G}(\theta)$ that implies the same equality restrictions as $\mathcal{G}^*$.*

However, the presence of the indicator function makes the BIC non-differentiable and optimization of L_0 objectives like the BIC is known to be NP-hard (Natarajan, 1995). While L_1 regularization is a popular alternative, it often leads to inconsistent model selection and overshrinkage of coefficients (Fan and Li, 2001). Several procedures have been devised in order to provide approximations of the BIC score; see Huang et al. (2018) for an overview. In this work, we consider the approximate BIC (ABIC) obtained via replacement of the indicator function with the hyperbolic tangent function as outlined in Su et al. (2016) and Nabi and Su (2017). That is, we seek to optimize $-2\ln(\mathcal{L}(X;\theta)) + \lambda\sum_{i=1}^{\dim(\theta)} \tanh(c|\theta_i|)$, where $c > 0$ is a constant that controls the sharpness of the approximation of the indicator function and λ controls the strength of regularization. As highlighted in Su et al. (2016), the ABIC is relatively insensitive to the choice of c. The main hyperparameter is the regularization strength λ. In our experiments we set $c = \ln(n)$ and report results for different choices of λ. In the next section we discuss our strategy to optimize the ABIC subject to the constraint that θ induces a valid ADMG within a hypothesis class $\mathbb{G}$.

Algorithm 2 REGULARIZED RICF

Inputs: $(X, \text{tol}, \text{max iterations}, h, \rho, \alpha, \lambda)$
Initialize estimates δ^t and β^t and set $c = \ln(n)$
Define $\text{LS}(\theta)$ as $\frac{1}{2n}\sum_{i=1}^{d} ||X_{:,i} - X\delta_{:,i} - Z^{(i)}\beta_{:,i}||_2^2$
for t in $(1, \ldots, \text{max iterations})$ **do**
  $\forall i \in (1,\ldots,d)$ compute $\epsilon_i \leftarrow X_{:,i} - \delta^t_{:,i}X$
  $\forall i \in (1,\ldots,d)$ compute $Z^{(i)} \in \mathbb{R}^{n\times d}$ as
    $Z^{(i)}_{:,i} = 0$ and $Z^{(i)}_{:,-i} \leftarrow \epsilon_{-i}(\beta^t_{-i,-i})^{-T}$
  $\delta^{t+1}, \beta^{t+1} \leftarrow \text{argmin}_{\theta\in\Theta} \big\{\text{LS}(\theta) + \frac{\rho}{2}|h(\theta)|^2$
    $+ \alpha h(\theta) + \lambda\sum_{i=1}^{\dim(\theta)} \tanh(c|\theta_i|)\big\}$
  $\forall i \in (1,\ldots,d)$ compute $\epsilon_i \leftarrow X_{:,i} - \delta^{t+1}_{:,i}X$
  $\forall i \in (1,\ldots,d)$ set $\beta^{t+1}_{ii} \leftarrow \text{var}(\epsilon_i)$
  if $||\delta^{t+1} - \delta^t + \beta^{t+1} - \beta^t|| < \text{tol}$ **then** break
return δ^t, β^t

Algorithm 3 DIFFERENTIABLE DISCOVERY

Inputs: $(X, \text{tol}, \text{max iterations}, s, h, \lambda, r \in (0,1))$
Initialize $\theta^t, \alpha^t, m^t \leftarrow 1$
while $t <$ max iterations and $h(\theta^t) >$ tol **do**
  $\theta^{t+1} \leftarrow \theta^*$ from REGULARIZED RICF with
    inputs $(X, 10^{-4}, m^t, h, \rho, \alpha^t, \lambda)$
    where ρ is such that $h(\theta^*) < rh(\theta^t)$
  $\alpha^{t+1} \leftarrow \alpha^t + \rho h(\theta^{t+1})$ and $m^{t+1} \leftarrow m^t + s$
return $\mathcal{G}(\theta^t)$

5.2 Solving the Continuous Program

We formulate the optimization objective as minimizing the ABIC subject to one of the algebraic equality constraints in Table 1. We use the augmented Lagrangian formulation (Bertsekas, 1997) to convert the problem into an unconstrained optimization problem with a quadratic penalty term, which can be solved using a dual ascent approach. Specifically, in each iteration we first solve the primal equation:

$$\min_{\theta\in\Theta} \ \text{ABIC}_\lambda(X;\theta) + \frac{\rho}{2}|h(\theta)|^2 + \alpha h(\theta),$$

where ρ is the penalty weight and α is the Lagrange multiplier. Then we solve the dual equation $\alpha \leftarrow$

$\alpha + \rho h(\theta^*)$. Intuitively, optimizing the primal objective with a large value of ρ would force $h(\theta)$ to be very close to zero thus satisfying the equality constraint.

However, unlike DAG models, maximum likelihood estimation of parameters under the restrictions of an ADMG does not correspond to a simple least squares regression that can be solved in one step. Drton et al. (2009) proposed an iterative procedure known as Residual Iterative Conditional Fitting (RICF) that produces a sequence of maximum likelihood estimates for δ and β under the constraints implied by a fixed ADMG $\mathcal{G}$. Each RICF step is guaranteed to produce better estimates than the previous step and the overall procedure is guaranteed to converge to a local optimum or saddle point when $\mathcal{G}(\theta)$ is arid/ancestral, i.e., globally identified (Drton et al., 2011).

In Algorithm 2 we describe a modification of RICF that directly inherits the aforementioned properties with respect to the regularized maximum likelihood objective, and can be used to solve the primal equation of our procedure. Briefly, for Gaussian ADMG models, maximization of the likelihood corresponds to minimization of a least squares regression problem where each variable i is regressed on its direct parents $V_j \to V_i$ and pseudo-variables Z formed from the residual noise terms and bidirected coefficients of its siblings $V_j \leftrightarrow V_i$. At each RICF step, we compute Z with respect to the current parameter estimates, and then solve the primal equation in line 7 of the algorithm. We repeat this until convergence or a prespecified maximum number of iterations. As RICF is not expected to converge during initial iterations of the augmented Lagrangian procedure when the penalty applied to $h(\theta)$ is quite small (resulting in non-arid graphs), we start with a small number of maximum RICF iterations and at each dual step increment this number. The penalty ρ applied to $h(\theta)$ is increased according to a fixed schedule where ρ is multiplied by a factor of 10 (up to a maximum value of 10^{16}) each time the inequality in line 4 of the algorithm is not satisfied. Our simulations show this works quite well in practice with convergence of the algorithm obtained typically within 10-15 steps of the augmented Lagrangian procedure.

We summarize our structure learning algorithm in Algorithm 3. Though optimization of the objective in display (1) is non-convex, standard properties of dual ascent procedures as well as the RICF algorithm guarantee that at each step in the process we recover parameter estimates that do not increase the objective we are trying to minimize. Further, per Theorem 2, if optimization of the ABIC objective for a given level of λ provides a good enough approximation of the BIC, the global minimizer (if found by our optimization procedure) yields a graph that implies the same equality restrictions as the true graph.

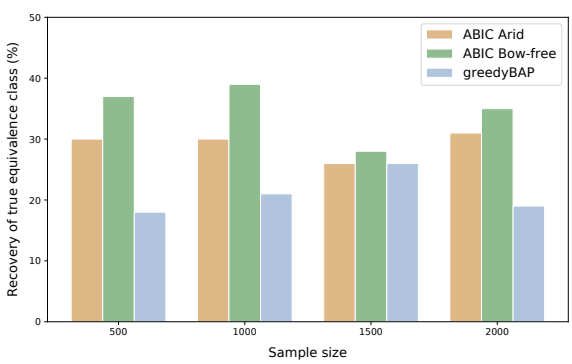
5.3 Reporting Equivalent Structures

Our procedure only reports a single ADMG but there may exist multiple ADMGs that imply the same equality restrictions on the observed data. In the linear Gaussian setting, exact recovery of the skeleton of the ADMG (i.e., adjacencies without any orientations) is possible, but complete determination of all edge orientations is not. Reporting the uncertainty in edge orientations is important for downstream causal inference tasks. When limiting our hypothesis class $\mathbb{G}$ to ancestral ADMGs, the non-parametric equivalence class can be represented via a Partial Ancestral Graph (PAG). After obtaining a single ADMG using our procedure, we can easily reconstruct its equivalence class using rules in Zhang (2008) to create the summary PAG. For arid and bow-free ADMGs, a full theory of equivalence that captures Verma constraints is still an open problem. Thus, while we are able to recover the exact skeleton, we coarsen reporting of edge orientations by converting the estimated ADMG into an ancestral ADMG and reporting the PAG. Connections in this PAG may be pruned using sound rules from Nowzohour et al. (2017) and Zhang et al. (2020) though we do not pursue this approach here. Deriving a summary structure that captures the class of all ADMGs that are equivalent up to equality restrictions is an important problem but outside the scope of this work.

6 EXPERIMENTS

For a given ADMG, we generate data as follows. For each $V_i \to V_j$ we uniformly sample δ_{ij} from $\pm[0.5, 2.0]$, for $V_i \leftrightarrow V_j$, we sample $\beta_{ij} = \beta_{ji}$ from $\pm[0.4, 0.7]$, and for each β_{ii} we sample from $\pm[0.7, 1.2]$ and add $\text{sum}(|\beta_{i,-i}|)$ to ensure positive definiteness of β.

Since randomly generated ADMGs are unlikely to exhibit Verma constraints, we first consider recovery of the ADMG shown in Fig. 1(c) and two other ADMGs $A \to B \to C \to D, B \leftrightarrow D$ and a Markov equivalent ADMG obtained by replacing $A \to B$ with $A \leftrightarrow B$ which have Verma constraints established in the prior literature. Exact recovery of Fig. 1(c) is possible while the latter ADMGs can be recovered up to ambiguity in the adjacency between A and B as $A \to B$ or $A \leftrightarrow B$. We compare our arid and bow-free algorithms to the greedyBAP method proposed in (Nowzohour et al., 2017) (the only other method available for recovering such constraints). Since greedyBAP is designed to perform random restarts, we allow all methods 5 uniformly random restarts and pick the final best fitting ADMG. As mentioned earlier, our main hyperparam-



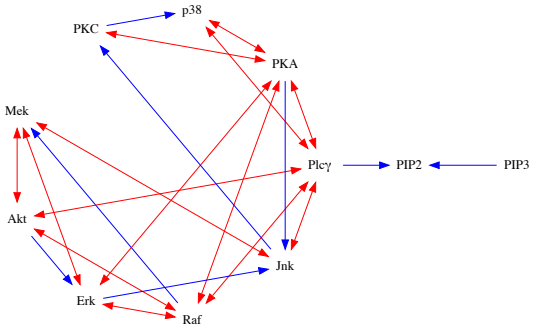


Figure 2: Left: Rate of recovery of the true equivalence class of an ADMG with a Verma constraint as a function of sample size. Right: Application of the ABIC bow-free method to the Sachs et al. (2005) dataset.

	Skeleton		Arrowhead		Tail	
Method	tpr ↑	fdr ↓	tpr ↑	fdr↓	tpr ↑	fdr ↓
gBAP (Nowzohour et al., 2017)	0.80	0.30	0.41	0.58	0.11	0.65
ABIC (bow-free)	**0.89**	**0.17**	**0.72**	**0.29**	**0.30**	**0.45**

	Skeleton		Arrowhead		Tail	
Method	tpr ↑	fdr ↓	tpr ↑	fdr↓	tpr ↑	fdr ↓
FCI (Spirtes et al., 2000)	0.51	0.12	0.41	0.53	0.10	0.73
gSPo (Bernstein et al., 2020)	**0.88**	0.27	0.46	0.59	0.32	0.81
ABIC (ancestral)	0.85	**0.11**	**0.72**	**0.23**	**0.66**	**0.47**

Table 2: Comparison of our method to greedyBAP (left) and FCI/greedySPo (right) for recovering 10 variable bow-free and ancestral ADMGs, respectively. We report true positive rate (tpr) and false discovery rate (fdr) — the fraction of predicted edges that are present in the target structure or the fraction that are absent from the target structure respectively — for skeleton, arrowhead and tail recovery. (↑/↓ indicates higher/lower is better.)

eter is the regularization strength λ, which we set to 0.05 for all experiments. Choice of other hyperparameters and additional experiments with varying λ are provided in Appendix D, E. We generate 100 datasets for each sample size of $[500, 1000, 1500, 2000]$ from a uniform sample of the 3 aforementioned ADMGs. The results are summarized via barplots in Fig. 2.

The ABIC arid and bow-free procedures both outperform the greedyBAP procedure in recovering the true equivalence class. The highest recovery rate is shown by the bow-free procedure with 39% at $n = 1000$. Though this seems low, these results are quite promising in light of geometric arguments in Evans (2018b) that show reliable recovery of Verma constraints may require very large sample sizes. In examining the modes of failure of each algorithm, our ABIC procedures often fail to recover the true ADMG by returning a super model of the true equivalence class while the greedyBAP procedure often returns an incorrect independence model; see Fig. C in Appendix E. The former kind of mistake does not yield bias in downstream inference tasks while the latter does. Our bow-free procedure yields more accurate results than the arid one most likely due to posing an easier optimization problem. In the 400 runs used to generate plots in Fig. 2, the bow-free procedure failed to converge only 3 times and the arid one never failed to converge, which is consistent with established theoretical results on almost-everywhere and global identifiability of these models.

For larger randomly generated arid ADMGs, to save computation time, we only compare our bow-free procedure with greedyBAP, and for ancestral ADMGs, we compare our ancestral procedure with FCI (Spirtes et al., 2000) and greedySPo (Bernstein et al., 2020). We also obtained results for GFCI (Ogarrio et al., 2016) and M3HC (Tsirlis et al., 2018). These were slightly worse than the results for FCI and greedySPo so we only report the latter results. Runs of the M3HC algorithm typically ended with convergence warnings.[1] Random arid/ancestral ADMGs on 10 and 15 variables were generated by first producing a random bow-free ADMG with directed and bidirected edge probabilities of 0.4 and 0.3 respectively, and then applying the maximal arid/ancestral projection. We report true positive and false discovery rates for exact skeleton recovery of the true ADMG as well as recovery of tails and arrowheads in the true PAG for 100 datasets of 1000 samples each. For FCI, we used a significance level of 0.15 which gave the most competitive results. Our method performs favorably in recovery of both arid and ancestral ADMGs. Results for 10 variables, which roughly

[1] Code from https://github.com/mensxmachina/M3HC.

matches the dimensionality of our data application, are summarized in Table 2. Results for 15 variables showing the same trends are in Appendix E.

Finally we apply our ABIC bow-free method to a cleaned version of the protein expression dataset in Sachs et al. (2005) from Ramsey and Andrews (2018). The result is shown in the right panel of Fig. 2. The precision and recall of our procedure with respect to the true adjacencies provided in Ramsey and Andrews (2018) are 0.77 and 0.61 respectively. We do not provide evaluation of orientations as there is no consensus regarding many of them. However, we briefly highlight the importance of a Verma restriction in producing a model that is consistent with an intervention experiment performed by Sachs et al. (2005). The authors found that manipulation of Erk produced no downstream effect on PKA though they are correlated. The ADMG in Fig. 2 has an edge Erk $\leftrightarrow$ PKA that is consistent with this finding. Moreover, this edge cannot be oriented in either direction without producing different independence models than the one implied by Fig. 2. This is due to a Verma restriction between Akt and PKC; we provide more details in Appendix B. We confirm that orienting the edge as Erk $\leftarrow$ PKA or Erk $\rightarrow$ PKA leads to an increase in the BIC score, indicating that the Verma restriction capturing the ground truth is preferred over these other explanations.

7 CONCLUSION

We have extended the continuous optimization scheme of causal discovery to include models that capture all equality constraints on the observed margin of hidden variable linear SEMs with Gaussian errors. The differentiable algebraic constraints we provided are non-parametric and may thus enable future development of non-parametric causal discovery methods. Our method may also help explore questions regarding distributional equivalence and Markov equivalence with respect to all equality restrictions in ADMG models. The authors in Shpitser et al. (2014) made progress on equivalence theory for 4-variable ADMGs by enumerating all possible 4-variable ADMGs and evaluating the BIC score for each one, grouping graphs with equal scores to form an "empirical equivalence class." A similar approach could be pursued for larger graphs using our proposed causal discovery procedure. If relevant patterns in larger empirical equivalence classes become apparent, this may result in progress towards a characterization for nested Markov equivalence.

Acknowledgements

The authors would like to thank Razieh Nabi for her insightful comments regarding approximations of the Bayesian Information Criterion. This project is sponsored in part by the NSF CAREER grant 1942239. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Bernstein, D., Saeed, B., Squires, C., and Uhler, C. (2020). Ordering-based causal structure learning in the presence of latent variables. In *International Conference on Artificial Intelligence and Statistics*, pages 4098–4108. PMLR.

Bertsekas, D. P. (1997). Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334.

Bhattacharya, R., Nabi, R., and Shpitser, I. (2020). Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*.

Brito, C. and Pearl, J. (2002). A new identification condition for recursive models with correlated errors. *Structural Equation Modeling*, 9(4):459–474.

Butler, S. K. (2008). *Eigenvalues and structures of graphs*. PhD thesis, UC San Diego.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, pages 294–321.

Drton, M., Eichler, M., and Richardson, T. S. (2009). Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10:2329–2348.

Drton, M., Foygel, R., and Sullivant, S. (2011). Global identifiability of linear structural equation models. *Annals of Statistics*, 39(2):865–886.

Evans, R. J. (2018a). Margins of discrete Bayesian networks. *Annals of Statistics*, 46(6A):2623–2656.

Evans, R. J. (2018b). Model selection and local geometry. *arXiv preprint arXiv:1801.08364*.

Evans, R. J. and Richardson, T. S. (2014). Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, pages 1452–1482.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Ghassami, A., Yang, A., Kiyavash, N., and Zhang, K. (2020). Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In *Proceedings of the 37th International Conference on Machine Learning*.

Haughton, D. M. (1988). On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16(1):342–355.

Huang, J., Jiao, Y., Liu, Y., and Lu, X. (2018). A constructive approach to L_0 penalized regression. *The Journal of Machine Learning Research*, 19(1):403–439.

Huang, Y. and Valtorta, M. (2006). Pearl's calculus of intervention is complete. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 13–16.

Hyttinen, A., Eberhardt, F., and Järvisalo, M. (2014). Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 340–349.

McKay, B. D., Oggier, F. E., Royle, G. F., Sloane, N. J. A., Wanless, I. M., and Wilf, H. S. (2004). Acyclic digraphs and eigenvalues of (0, 1)-matrices. *Journal of Integer Sequences*, 7(2):3.

Nabi, R. and Su, X. (2017). coxphMIC: An R package for sparse estimation of Cox proportional hazards models via approximated information criteria. *R Journal*, 9(1):229–238.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234.

Nowzohour, C., Maathuis, M. H., Evans, R. J., Bühlmann, P., et al. (2017). Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electronic Journal of Statistics*, 11(2):5342–5374.

Ogarrio, J. M., Spirtes, P. L., and Ramsey, J. D. (2016). A hybrid causal search algorithm for latent variable models. In *Proceedings of the 8th International Conference on Probabilistic Graphical Models*, pages 368–379.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Ramsey, J. and Andrews, B. (2018). FASK with interventional knowledge recovers edges from the Sachs model. *arXiv preprint arXiv:1805.03108*.

Richardson, T. S. and Spirtes, P. (2002). Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

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

Shpitser, I., Evans, R. J., and Richardson, T. S. (2018). Acyclic linear SEMs obey the nested Markov property. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence*.

Shpitser, I., Evans, R. J., Richardson, T. S., and Robins, J. M. (2014). Introduction to nested markov models. *Behaviormetrika*, 41(1):3–39.

Shpitser, I. and Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence*.

Spirtes, P. L., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Su, X., Wijayasinghe, C. S., Fan, J., and Zhang, Y. (2016). Sparse estimation of Cox proportional hazards models via approximated information criteria. *Biometrics*, 72(3):751–759.

Tian, J. and Pearl, J. (2002). On the testable implications of causal models with hidden variables. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 519–527.

Tsirlis, K., Lagani, V., Triantafillou, S., and Tsamardinos, I. (2018). On scoring maximal ancestral graphs with the max–min hill climbing algorithm. *International Journal of Approximate Reasoning*, 102:74–85.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence*.

Wang, Y. S. and Drton, M. (2020). Causal discovery with unobserved confounding and non-Gaussian data. *arXiv preprint arXiv:2007.11131*.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20:557–580.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, 5(3):161–215.

Yao, B. and Evans, R. J. (2019). Constraints in Gaussian graphical models. *arXiv preprint arXiv:1911.12754*.

Yu, Y., Chen, J., Gao, T., and Yu, M. (2019). DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7154–7163.

Zhang, C., Chen, B., and Pearl, J. (2020). A simultaneous discover-identify approach to causal inference

in linear models. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, pages 10318–10325.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896.

Zhang, M., Jiang, S., Cui, Z., Garnett, R., and Chen, Y. (2019). D-VAE: A variational autoencoder for directed acyclic graphs. In *Advances in Neural Information Processing Systems*, pages 1588–1600.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. P. (2020). Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425.