Inference of Causal Effects when Control Variables are Unknown

Ludvig Hult[1] **Dave Zachariah**[1]

[1]Systems and Control, Uppsala Universitet, Uppsala, Sweden

Abstract

Conventional methods in causal effect inference typically rely on specifying a valid set of control variables. When this set is unknown or misspecified, inferences will be erroneous. We propose a method for inferring average causal effects when all potential confounders are observed, but the control variables are unknown. When the data-generating process belongs to the class of acyclical linear structural causal models, we prove that the method yields asymptotically valid confidence intervals. Our results build upon a smooth characterization of linear directed acyclic graphs. We verify the capability of the method to produce valid confidence intervals for average causal effects using synthetic data, even when the appropriate specification of control variables is unknown.

1 INTRODUCTION

When applied researchers aim to assess the causal effect of some policy or exposure, they must often infer it from observational data. This requires controlling for variations in the outcome of interest that arise from confounding factors. After selecting a set of control variables, inferences are often drawn using regression models. But selecting a valid control variable set is in general hard and the use of invalid sets produces misleading inferences, see. e.g., Carlson and Wu [2012], Bernerth and Aguinis [2016]. It is therefore of practical interest to infer causal effects without relying on the researcher to specify the control variables among all observed variables.

In this paper, we will develop such an inferential method under the assumption that there is no unobserved confounding. The method infers average causal effects using asymptotic confidence intervals and obviates the need for specifying control variables.

Consider a random outcome variable y observed after an intervention on another scalar x. We denote the unknown conditional distribution of outcomes under such an intervention as

$$y \sim \tilde{p}(y|x)$$

We consider the scalars x and y to be of zero mean, i.e. $\widetilde{\mathbb{E}}[x] = \widetilde{\mathbb{E}}[y] = 0$, where the tilde denotes that the expectation is taken with respect to the interventional distribution $\tilde{p}$. The conditional mean function $\widetilde{\mathbb{E}}[y|x]$ describes the effect of the intervention and can be summarized by the distribution parameter

$$\boxed{\gamma \coloneqq \frac{\widetilde{\mathrm{Cov}}[x,y]}{\widetilde{\mathrm{Var}}[x]} \;\equiv\; \arg\min_{\bar{\gamma}} \; \widetilde{\mathbb{E}}\left[\left(\widetilde{\mathbb{E}}[y|x] - \bar{\gamma}x\right)^2\right]} \quad (1)$$

Thus γx is an optimal linear approximation of the conditional mean function. When the conditional mean function is linear, the parameter is the average causal effect of the intervention, i.e., $\gamma \equiv \frac{\partial}{\partial x}\widetilde{\mathbb{E}}[y|x]$ [Angrist and Pischke, 2009, Pearl, 2009].

The task is to infer γ using data from a different, *observational* distribution

$$(x_i, y_i, z_i) \sim p(x, y, z), \quad i = 1, \ldots, n \quad (2)$$

where z is a vector of additional random variables. A standard procedure to infer γ is to use the partial regression coefficient

$$\beta \coloneqq \frac{\mathrm{Cov}[\bar{x}, \bar{y}]}{\mathrm{Var}[\bar{x}]}, \quad (3)$$

where $\bar{x}$ and $\bar{y}$ are adjusted according to

$$\begin{aligned} \bar{x} &\coloneqq x - \mathrm{Cov}[x, \bar{z}]\,\mathrm{Cov}[\bar{z}]^{-1}\bar{z} \\ \bar{y} &\coloneqq y - \mathrm{Cov}[y, \bar{z}]\,\mathrm{Cov}[\bar{z}]^{-1}\bar{z}, \end{aligned} \quad (4)$$

where $\bar{z} \subseteq z$ is a set of *control variables* using the terminology in much of regression analysis. If this set were *valid*, the noncausal association between x and y can be blocked. Then

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

$\beta = \gamma$ when the data-generating process is well-described by a linear model [Angrist and Pischke, 2009, Pearl, 2009]. See [Peters et al., 2017, ch. 6.6] for a general definition of valid control variables using structural causal models (SCM). Throughout the paper, we will assume that at least one valid subset of z exists but that it is *unknown*. If a specified $\bar{z}$ contains invalid controls, the resulting inferences become erroneous as the following example illustrates.

Example: Invalid control variables Consider a data-generating process with a causal structure as illustrated in Figure 1a. Only $z_2 \subset z$ constitutes a valid control variable, by blocking the noncausal association between x and y. Neither $\varnothing$ nor z_1 are valid. If the causal structure is unknown or misspecified so that we use $\bar{z} = [z_1, z_2]^\top$ instead of z_2, then inferring β in equation (3) will yield erroneous conclusions about the average causal effect, as shown in Figure 1b. We also illustrate an alternative methodology developed in this paper which, by contrast, does not require a correctly specified causal structure.

Contribution and related work The contribution of this paper is the development of a confidence interval for the average causal effect that obviates the need to specify valid control variables, and we derive its statistical properties.

To decide the valid control variables among z, typically requires the causal structure of the data-generating process. The problem of learning such structures from data, aka. causal discovery, has been studied over a few decades [Spirtes et al., 1993, Pearl, 2009, Peters et al., 2017]. A central challenge of the field is to optimize model fitness over the discrete nature of graphs representing the causal structure. Zheng et al. [2018] proposed a smooth characterization of directed acyclic graphs (DAG) which enables conventional optimization methods to be used. See [Yu et al., 2019, Ke et al., 2020, Brouillard et al., 2020, Zheng et al., 2020, Kyono and Zhang, 2020] for applications and extentions of this methodology.

Our method presented herein utilizes that characterization of DAGs and builds upon the framework of M-estimation. See e.g. the presentation in [Wooldridge, 2010, ch. 12] or Vaart [1998] for an introduction. When imposing DAG-constraints, we find the need to extend the basic M-estimation framework. While the theory of constrained M-estimation has been approached before [Geyer, 1994, Shapiro, 2000, Andrews, 1999, Wang, 1996], we show that the assumptions needed do not hold due to the geometry of the DAG constraints. Moreover, alternative characterizations of DAGs, presented in Wei et al. [2020], would not remedy this problem.

Therefore we take a different approach, inspired by Stoica and Ng [1998], to derive the large-sample properties of the proposed confidence interval and prove its asymptotic validity. Our theoretical results are corroborated by numerical experiments, which demonstrate the ability of the method to correctly infer average causal effects in linear SCMs without specifying valid control variables.

Lastly we emphasize that while our method builds upon insights from the causal discovery literature, its task is to infer the average causal effect and not a causal graph.

2 PROBLEM FORMULATION

We begin by specifying the class models for the data generating process that we will consider and then proceed to define the target quantity that we seek to infer from data.

2.1 MODEL CLASS FOR THE DATA-GENERATING PROCESS

To simplify the notation, we introduce the d-dimensional data vector $v^\top = (x,\, y,\, z^\top)$. Suppose the data-generating process $p(v)$ in (2) belongs to the class of linear SCM. That is, we can express the data vector as

$$v = W^\top v + e, \tag{5}$$

where is e is zero-mean random variable with a diagonal covariance matrix Σ. It is for simplicity assumed to be known here, although as we point out in Section 3 this assumption can be relaxed to a certain degree. We let $W \in \mathbb{R}^{d\times d}$ have zeros on its diagonal. It can be interpreted as a weighted directed graph, by letting $W_{i,j}$ be the weight on the edge from node i to node j. The matrix $W^\top$ is sometimes referred to as the *adjacency matrix* [Shimizu et al., 2011] or the *autoregressive matrix* [Loh and Bühlmann, 2014].

The matrix W is unknown but has certain restrictions. For SCMs it is common to impose a DAG structure on the graph specified by W, since such structure significantly clarifies and simplifies any causal analysis of the model. We will call W a 'DAG-matrix' if the directed graph of the matrix is acyclical. When W is a DAG-matrix, we can interpret the entry $W_{i,j}$ as the expected increase in v_i for every unit increase in v_j, holding all other variables constant.

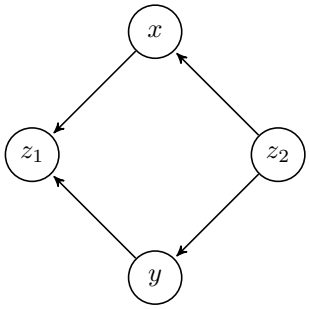
Zheng et al. [2018] introduced the function $h(W) := \operatorname{tr}\exp(W \circ W) - d$, using the trace of the matrix exponential and the element-wise product $\circ$, and showed that

$$W \text{ is DAG-matrix} \Leftrightarrow h(W) = 0$$

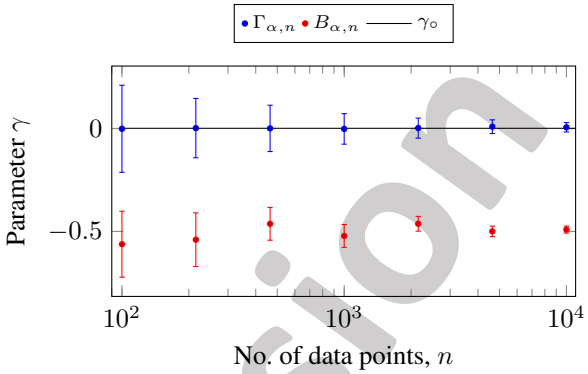
To enable a tractable analysis below, we will also consider the set of all ϵ-almost DAG-matrices, defined as

$$\mathcal{W}_\epsilon = \{W \,|h(W) \leq \epsilon \text{ and } \operatorname{diag}(W) = 0\} \tag{6}$$

Note that when $\epsilon = 0$, the set $\mathcal{W}_0$ is exactly the set of DAG-matrices. When $\epsilon > 0$, cycles are permitted but the magnitude of their effects are bounded. Below we will provide bounds on ϵ that enable a meaningful analysis of $\mathcal{W}_\epsilon$.



(a) Underlying causal structure



(b) 95%-confidence intervals that aim to cover $\gamma_\circ$

Figure 1: Using observational data (2) generated by a linear SCM based on (a), we aim to infer an unknown causal parameter $\gamma_\circ$ (further details in Section 4.2). The causal structure is here unknown and using $z = [z_1, z_2]$ as the control variables, the standard approach based on the ordinary least-squares (OLS) method yields confidence interval $B_{\alpha,n}$ in (b). Since z is invalid due to the collider bias induced by z_1, the inferences are erroneous. Below we develop an inference method that yields calibrated confidence intervals $\Gamma_{\alpha,n}$ when the causal structure in (a), and therefore a set of valid control variables, is unknown.

Given the data-generating process in (5), we can define an *interventional* distribution $\widetilde{p}(v)$ with respect to the first variable x [Pearl, 2009]: Introduce Z, a matrix with ones on the diagonal, except the first element, which is zero, i.e.

$$Z \in \mathbb{R}^{d\times d}, \quad Z_{i,j} = \begin{cases} 1 & \text{if } i = j > 1 \\ 0 & \text{else} \end{cases} \tag{7}$$

Next, introduce a new random vector $\widetilde{e}$, with the same statistical properties as e in (5) for all components, but for its first component, and let $\widetilde{\Sigma}$ denote its diagonal covariance matrix. The interventional distribution $\widetilde{p}(v)$ is then specified by the model

$$v = ZW^\top v + \widetilde{e}, \tag{8}$$

assuming that $(I - ZW^\top)$ is full rank.

2.2 TARGET QUANTITY

For an interventional distribution given by (8), we observe the following result.

Lemma 1. *The average causal effect of x on y in a linear* SCM *with interventional distribution $\widetilde{p}(v)$ is*

$$\gamma(W) = \frac{\widetilde{Cov}[x,y]}{\widetilde{Var}[x]} \equiv \left[(I - ZW^\top)^{-1}\right]_{2,1} \tag{9}$$

*where W is a (possibly non-*DAG*) adjacency matrix.*

The syntax $[.]_{2,1}$ refers to the second row and first column of a matrix. The proof is a direct computation and given in the supplementary material.

We are interested in computing the average causal effect

$$\boxed{\gamma_\circ = \gamma(W_\circ),} \tag{10a}$$

where $W_\circ$ is an ϵ-almost DAG adjacency matrix that optimally fits the observational data using the following criterion,

$$W_\circ := \underset{W\in\mathcal{W}_\epsilon}{\arg\min}\ \mathbb{E}\Big[\|\Sigma^{-1/2}(I - W^\top)v\|^2\Big] \tag{10b}$$

Loh and Bühlmann [2014, corollary 8] show that if the observational distribution $p(z)$ follows (5) and $\epsilon = 0$, then (10b) correctly identifies the unknown matrix. Moreover, Loh and Bühlmann [2014, theorem 9] proves that identifiability is obtained even under limited misspecification of the entries in $\mathrm{Cov}[e] = \Sigma$. Thus the target quantity $\gamma_\circ$ is defined as the average causal effect of the optimally fitted linear SCM and requires no further distributional assumptions.

Our task is to construct a confidence interval $\Gamma_{\alpha,n}$, that is using n data points, and has a coverage probability $1 - \alpha$ for the quantity $\gamma_\circ$.

3 RESULTS

We present the results in this paper in two parts. First, we present the confidence interval for $\gamma_\circ$ with an asymptotically valid coverage probability (Theorem 4). This uses a general result of equality-constrained M-estimation, which we subsequently present (Theorem 5, Corollary 6).

3.1 DERIVATION OF CONFIDENCE INTERVAL

Using the empirical average operator $\mathbb{E}_n$, we define the empirical analog of (10b) as

$$W_n \coloneqq \underset{W \in \mathcal{W}_\epsilon}{\arg\min}\ \mathbb{E}_n\left[\|\Sigma^{-1/2}\left(I - W^\top\right) v\|^2\right] \tag{11}$$

Using W_n and (9) yields a point estimate of $\gamma_\circ$:

$$\gamma_n \coloneqq \gamma(W_n) \tag{12}$$

For notational simplicity, we reparameterize W, which contains zeros along the diagonal, by $\mathrm{vec}(W) = L\theta$, where L is a $d^2 \times d(d-1)$ matrix constructed using a $d^2 \times d^2$ identity matrix removing columns $d(k-1)+k$ for $k = 1, 2, \ldots, d$. Using this parametrization, we formulate the loss function

$$\ell_\theta(v) \coloneqq (L\theta - \mathrm{vec}(I))^\top \left[\Sigma^{-1} \otimes [vv^\top]\right] (L\theta - \mathrm{vec}(I)) \tag{13}$$

using the Kronecker product $\otimes$, and we write

$$\theta_\circ = \underset{h(\mathrm{mat}(L\theta)) \le \epsilon}{\arg\min}\ \mathbb{E}[\ell_\theta(v)] \tag{14}$$

$$\theta_n = \underset{h(\mathrm{mat}(L\theta)) \le \epsilon}{\arg\min}\ \mathbb{E}_n[\ell_\theta(v)] \tag{15}$$

equivalently to (10b) and (11).

While setting $\epsilon = 0$ yields exact DAG-matrices, it also renders the problem ill-suited for inference. The set $\mathcal{W}_0$ is nonconvex, has an empty interior, and constraint qualification does not hold (see Lemma 3 in the supplementary material). Therefore, convex optimization methods, barrier methods, and any method based on first-order optimality will be invalid. Asymptotic analysis of M-estimation typically requires convexity of the tangent cone at the optimum, and that the optimal point is stationary even under the unconstrained formulation [Geyer, 1994, Shapiro, 2000], but neither of these assumptions are fulfilled at most points in the set $\mathcal{W}_0$. To provide a tractable analysis, we consider $\epsilon > 0$ below and expect almost-identification when ϵ is small. We start with a technical lemma.

Lemma 2. *The minimizer $\theta_\circ$ in* (14) *is bounded. If it is also unique, then there is a value of $\epsilon_\star$ such that the minimum is obtained at the boundary $h(\mathrm{mat}(L\theta_\circ)) = \epsilon$ for all $\epsilon < \epsilon_\star$.*

Proof. First, assume that the mimimizer of (14) is not bounded. In that case, there is a sequence of feasible points t_n such that $\|t_n\| \to \infty$, and $\mathbb{E}[\ell_{t_n}(v)]$ is decreasing. This is not possible, since $\ell_t(v)$ is a positive definite quadratic in t. We have established the boundedness $\|\theta_\circ\| < B$, for some B.

Let $\mathsf{Q} = \Sigma^{-1} \otimes \mathbb{E}[vv^\top]$, i.e. a Kronecker product of two positive definite matrices and it follows that Q is positive definite. Then the objective function of (14) is a positive definite quadratic with a global minimum given by the stationary point $\theta_\star =: (\mathsf{Q}^{1/2}L)^\dagger \mathsf{Q}^{1/2}\,\mathrm{vec}(I)$ where † denotes the Moore-Penrose inverse. When $\epsilon = \infty$, then $\theta_\star$ is a feasible point to the minimization problem in (14).

Define $\epsilon_\star = h(\mathrm{mat}(L\theta_\star))$ and consider (14) for any $\epsilon \in (0,\ \epsilon_\star)$. Observe that $\{\theta \mid \|\theta\| \le B \text{ and } h(\mathrm{mat}(L\theta)) \le \epsilon\}$ is compact, the objective function has no stationary points on the feasible set, and $\|\theta_\circ\| < B$. Conclude that $h(\mathrm{mat}(L\theta_\circ)) = \epsilon$. □

Lemma 3. *Assume the solution to* (14) *is unique, and that $\epsilon < \epsilon_\star$ as in Lemma 2. Then the asymptotic distribution of θ_n can be described by*

$$\sqrt{n}\mathcal{J}_n^{-1/2}(\theta_n - \theta_\circ) \xrightarrow{d} \mathcal{N}(0, I) \tag{16}$$

The estimated covariance of the estimator is defined as $\mathcal{J}_n = K_n^{-1}\Pi_n J_n \Pi_n K_n^{-1}$, where $K_n = L^\top\left[\Sigma^{-1} \otimes \mathbb{E}_n\left[vv^\top\right]\right]L$, Π_n is a projection matrix with respect to the orthogonal complement of $\nabla_\theta h(\mathrm{mat}(L\theta_n))$ and $J_n = L^\top \tilde{J}_n L$.

We may compute $\Pi_n = I - (qq^\top)/(q^\top q)$ and $q = L^\top\,\mathrm{vec}(2W_n \circ (\exp[W_n \circ W_n])^\top)$. Furthermore, the matrix $\tilde{J}_n$ has the expression

$$(\tilde{J}_n)_{d(j-1)+i,d(l-1)+k} = \sum_{q,r,o,p=1}^{d} \Big\{\big(\mathbb{E}_n\left[v_i v_q v_o v_k\right] - \mathbb{E}_n\left[v_i v_q\right]\mathbb{E}_n\left[v_o v_k\right]\big)\Sigma^{-1}_{j,r}\Sigma^{-1}_{p,l}(W-I)_{q,r}(W-I)_{o,p}\Big\} \tag{17}$$

Proof. By consistency of M-estimation, (15) will be a consistent estimator for (14). Adding the redundant $\|\theta\| \le B$-constraint in Lemma 2 makes the feasible set compact and thus fulfills the technical conditions [Wooldridge, 2010, Theorem 12.2].

By Lemma 2, we know that the minimum will be obtained at the boundary, in the limit $n \to \infty$. We can therefore impose equality constraints in the minimization:

$$\theta_n = \underset{h(\mathrm{mat}(L\theta)) = \epsilon}{\arg\min}\ \mathbb{E}_n[\ell_\theta(v)] \tag{18}$$

Now apply Corollary 6 derived below. It states the formula for confidence intervals under equality-constrained M-estimation using plug-in estimators of data covariance and cross-moments. The derivation of the expressions for $\tilde{J}_n$, K_n and Π_n from (13) are direct computations presented in the supplementary material as Lemma 5. Technical conditions are presented in Lemma 6. □

We can now state our main result for inferring the average causal effect $\gamma_\circ$.

Theorem 4. *The confidence interval*

$$\Gamma_{\alpha,n} = \left\{ \gamma \in \mathbb{R} \middle| \frac{1}{n} \frac{(\gamma - \gamma_n)^2}{\nabla\gamma(\theta_n)^\top \mathcal{J}_n \nabla\gamma(\theta_n))} \leq \chi^2_{1,\alpha} \right\} \tag{19}$$

has asymptotic coverage probability

$$\lim_{n\to\infty} \mathbb{P}(\gamma_\circ \in \Gamma_{\alpha,n}) = 1 - \alpha, \tag{20}$$

where $\chi^2_{1,\alpha}$ denotes the $(1-\alpha)$ quantile of the chi-squared distribution with 1 degree of freedom.

Proof. Define $\gamma(\theta)$ as the value of $\gamma(\mathrm{mat}(L\theta))$ in (9).

The gradient $\nabla\gamma(\theta_n)$ may be computed on closed form by differentiating (9), obtaining

$$[\nabla_\theta \gamma(\theta)]_k = -\left([MZ \otimes I]\, L\right)_{d+1,k} \tag{21}$$

where $M = (I - ZW)^{-1}$. The computation is mostly keeping track of indices, and presented in supplementary materials as Lemma 7. Using the delta method with equation (21) together with Lemma 3, we establish asymptotic normality. Form the Wald statistic for γ_n, and we may finally define a confidence interval $\Gamma_{\alpha,n}$. □

3.2 M-ESTIMATION ASYMPTOTICS UNDER EQUALITY CONSTRAINTS

Next we derive a general result for the asymptotics of of equality-constrained M-estimation. The key observation is borrowed from Stoica and Ng [1998]: that we can project onto the (generalized) score onto the active constraints. We apply this insight to the more general M-estimation framework and derive complete asymptotic distribution of equality-constrained M-estimators.

In this section 3.2 the function ℓ is not necessarily the same function as defined in (13) but we use the same symbol to ease the mapping between the general result and its application.

Theorem 5. *Assume that technical conditions for consistency of M-estimation holds [Wooldridge, 2010, Theorem 12.2]), as well as*

- *The loss function $\ell_\theta(v)$ is two times continously diffrentiable in v.*
- *$\Theta := \{\theta \in \mathbb{R}^p \mid g(\theta) = 0\}$ for some vector-valued constraint function g such that Θ is bounded.*
- *The Jacobian matrix $\nabla g(\theta_n)$ has full rank for all n.*
- *$\mathbb{E}_n\left[\nabla^2 \ell_\theta(v)\right]$ is invertible for all θ.*
- *$\theta_\circ$ is the unique minimizer of $\mathbb{E}[\ell_\theta(v)]$*

Introduce the definitions $J_\circ := \mathrm{Cov}[\nabla\ell_{\theta_\circ}(v)]$, $K_\circ := \mathbb{E}[\nabla^2\ell_{\theta_\circ}(v)]$ and $\Pi_\circ$ is an orthogonal projector in the complement of the range of the jacobian $\nabla g(\theta_\circ)$. Then we can establish the convergence

$$\sqrt{n}(\theta_n - \theta_\circ) \xrightarrow{d} \mathcal{N}(0, K_\circ^{-1}\Pi_\circ J_\circ \Pi_\circ K_\circ^{-1}).$$

Proof. Uniform weak law of large numbers holds, and Θ must be compact since bounded and closed, so we have that $\theta_\circ$ is consistently estimated by θ_n

Let Q_n be a matrix whose orthonormal columns spans the range of $\nabla g(\theta_n)$ (as in e.g. QR factorization). Construct an orthogonal matrix $[Q_n\, U_n]$. Now, Q_n is a ON basis for the normal of the feasible set Θ, and U_n is a ON basis for the tangent cone of Θ as θ_n.

Begin by a mean-value expansion of $\mathbb{E}_n\left[\nabla\ell_{\theta_n}(v)\right]$.

$$\mathbb{E}_n[\nabla\ell_{\theta_n}(v)] = \mathbb{E}_n[\nabla\ell_{\theta_\circ}(v)] + \mathbb{E}_n[\nabla^2\ell_{\tilde\theta}(v)](\theta_n - \theta_\circ) \tag{22}$$

We have that $I = [Q_n\, U_n] \begin{bmatrix} Q_n^\top \\ U_n^\top \end{bmatrix}$

$$[Q_n\, U_n] \begin{bmatrix} Q_n^\top \\ U_n^\top \end{bmatrix} \mathbb{E}_n[\nabla\ell_{\theta_n}(v)] \tag{23}$$

$$= [Q_n\, U_n] \begin{bmatrix} Q_n^\top \\ U_n^\top \end{bmatrix} \mathbb{E}_n[\nabla\ell_{\theta_\circ}(v)] + \mathbb{E}_n[\nabla^2\ell_{\tilde\theta}(v)](\theta_n - \theta_\circ) \tag{24}$$

By definition $U_n^\top \nabla g(\theta_n) = 0$, and from first order optimality conditions $\nabla\ell_{\theta_n}$ is in the range of $\nabla g(\theta_n)$, so $U_n^\top \nabla\ell_{\theta_n} = 0$.

Rearranging, and using the assumption of invertibility of $\mathbb{E}_n[\nabla^2\ell_{\tilde\theta}(v)]$, we get

$$(\theta_n - \theta_\circ) = \tag{25}$$

$$\mathbb{E}_n\left[\nabla^2\ell_{\tilde\theta}(v)\right]^{-1} [Q_n\, U_n] \begin{bmatrix} Q_n^\top \left(\mathbb{E}_n\left[\nabla\ell_{\theta_n}(v) - \nabla\ell_{\theta_\circ}(v)\right]\right) \\ -U_n^\top \mathbb{E}_n\left[\nabla\ell_{\theta_\circ}(v)\right] \end{bmatrix} \tag{26}$$

Next, we will analyze a certain subexpression separately. Introduce $\Pi_\circ = U_\circ U_\circ^\top$ and $\Pi_n = U_n U_n^\top$.

$$\sqrt{n}\Pi_n \mathbb{E}_n\left[\nabla\ell_{\theta_\circ}(v)\right] = \tag{27}$$

$$\Pi_n \sqrt{n}\left(\mathbb{E}_n\left[\nabla\ell_{\theta_\circ}(v)\right] - \mathbb{E}\left[\nabla\ell_{\theta_\circ}(v)\right]\right) + \Pi_n\sqrt{n}\mathbb{E}\left[\nabla\ell_{\theta_\circ}(v)\right] \tag{28}$$

The first term converges to $\mathcal{N}(0, \Pi_\circ J_\circ \Pi_\circ)$ in distribution. The second term converges to zero in probability, so

$$\sqrt{n}\Pi_n \mathbb{E}_n\left[\nabla\ell_{\theta_\circ}(v)\right] \xrightarrow{d} \mathcal{N}(0, \Pi_\circ J_\circ \Pi_\circ) \tag{29}$$

Finally, we can take the limit of equation (25).

$$\sqrt{n}(\theta_n - \theta_\circ) = \sqrt{n}\underbrace{\left[\mathbb{E}_n\left[\nabla^2 \ell_{\tilde{\theta}}(v)\right]\right]^{-1}}_{\xrightarrow{p} K^{-1}} \underbrace{Q_n Q_n^\top}_{\xrightarrow{p} Q_\circ Q_\circ^\top} \underbrace{\left(\mathbb{E}_n\left(\nabla \ell_{\theta_n}(v)\right] - \mathbb{E}_n\left[\nabla \ell_{\theta_\circ}(v)\right]\right)}_{\xrightarrow{p} 0} - \underbrace{\left[\mathbb{E}_n\left[\nabla^2 \ell_{\tilde{\theta}}(v)\right]\right]^{-1}}_{\xrightarrow{p} K_\circ^{-1}} \underbrace{\sqrt{n}\Pi_n \mathbb{E}_n\left[\nabla \ell_{\theta_\circ}(v)\right]}_{\xrightarrow{d} \mathcal{N}(0, \Pi_\circ J_\circ \Pi_\circ)} \quad (30)$$

For all terms converging in probability we have been using the uniform weak law of large numbers, so we rely on compactness of Θ, and the suitable smoothness of the functions depending on v. We need, for example, the continuity of matrix inversion, QR factorization and orthogonal complements. W e use Slutskys theorem to multiply the terms.

Finally we see $\sqrt{n}(\theta_n - \theta_\circ) \xrightarrow{d} \mathcal{N}(0, K_\circ^{-1}\Pi_\circ J_\circ \Pi_\circ K_\circ^{-1})$ □

Corollary 6. *The asymptotic distribution of Theorem 5 can be reformulated by standardizing it, and plugging in estimates (e.g. K_n) in the place of the population optimal expressions (e.g. $K_\circ$).*

$$\sqrt{n}\mathcal{J}_n^{-1/2}(\theta_n - \theta_\circ) \xrightarrow{d} \mathcal{N}(0, I).$$

with the introduction of

$$\mathcal{J}_n \coloneqq K_n^{-1}\Pi_n J_n \Pi_n K_n^{-1}$$

$$K_n \coloneqq \mathbb{E}_n\left[\nabla^2 \ell_{\theta_n}(v)\right]$$

$$J_n \coloneqq \mathbb{E}_n[\nabla \ell_{\theta_n}(v) \nabla \ell_{\theta_n}(v)^\top] - \mathbb{E}_n[\nabla \ell_{\theta_n}(v)]\mathbb{E}_n[\nabla \ell_{\theta_n}(v)]^\top$$

Proof. This follows from the consistency of plug-in-estimators [Wooldridge, 2010, Theorem 12.2]. □

4 NUMERICAL ILLUSTRATIONS

In the following experiments, data was generated using a linear SCM (5) with a matrix W that is either fixed or random. For random DAG-matrices, we follow Yu et al. [2019, section 4.1]: Let d be the number of nodes in a SCM. Let k be the expected number of edges in a randomly generated DAG. Let M be a random strictly subtriangular matrix where entries are drawn $\text{Bernoulli}(2k/(d-1))$. Let P be a random permutation matrix. Let C be uniformly drawn from the interval $[0.5, 2]$, and set $W = P^\top(C \circ M)P$.

The random vector e in (5) has elements with unit variance and are drawn independently as either Normal(0,1), Exp(1) or Gumbel(0,$6/\pi^2$)). Data was also centered before any other processing.

Throughout all runs, the nominal miscoverage level was set to $\alpha = 5\%$ and $\epsilon = 10^{-7}$.

Remark. In the supplementary material, we study deviations from the linear data model, in which case the average causal effect (10a) of the optimal linear model is still defined.

Remark. In all cases when the data generator is a linear SCM with Gaussian noise, we apply Isserlis' theorem to equation (17), $\mathbb{E}_n[v_i v_q v_o v_k] - \mathbb{E}_n[v_i v_q]\mathbb{E}_n[v_o v_k] = \mathbb{E}_n[v_i v_o]\mathbb{E}_n[v_q v_k] + \mathbb{E}_n[v_i v_k]\mathbb{E}_n[v_q v_o]$. This reduction is especially helpful in high dimensions, when d is large.

4.1 NUMERICAL SEARCH METHOD

In the examples below, we construct the confidence interval (19) by numerically solving problem (15). Here we use the augmented Lagrangian method [Nocedal and Wright, 2006], but other search methods are possible as well.

We define the augmented Lagrangian and the equality converted constraint as

$$\mathcal{L}(\theta, s, \alpha, \rho) = \mathbb{E}_n[\ell_\theta(v)] + \alpha c(\theta, s) + \frac{\rho}{2} c(\theta, s)^2 \quad (31)$$

$$c(\theta, s) = h(\text{mat}(L\theta)) + s^2 - \epsilon$$

The method alternates between the minimization over primal variables (θ,s) and maximization over dual variables (α), starting from a few initialization points, as explicated in Algorithm 1.

Algorithm 1: Augmented Lagrangian Method

Input: θ^0,s^0,ρ^0, α^0, g, μ,$\mathcal{L}$,η,ρ_{max},c
Output: θ_n

```
1  k = 0
3  while c(θ^k, s^k) > η and ρ < ρ_max do
5      θ^{k+1}, s^{k+1} = arg min_{θ,s} L(θ, s, α^k, ρ^k)
6      α^{k+1} = α^k + ρ^k c(θ^{k+1}, s^{k+1})
7      if c(θ^{k+1}, s^{k+1}) > g c(θ^k, s^k) then
8          ρ^{k+1} = μ ρ^k
9      else
10         ρ^{k+1} = ρ^k
11     k = k + 1
12 return θ_n = θ^{k+1}
```

The minimization problem on line 5 is solved via the L-BFGS-B-implementation in the python library `scipy.optimize`, which in turn utilizes the 3.0 version of the FORTRAN library of Zhu et al. [1997]. Since this is a local minimizer, we use the previous optimal primal variables θ^k, s^k as the starting point.

The parameters have default values set to $\theta^0 = 0$, $s^0 = 10$, $\rho^0 = 1$, $\alpha^0 = 0$, $g = 1/4$, $\mu = 2$, $\eta = 10^{-12}$, $\rho_{max} = 10^{20}$. Note that η must be significantly smaller than ϵ, which in turn should be smaller than $\epsilon_\star$. Thus it is advisable to verify that the choice of η is sufficiently small in a given problem. The threshold ρ_{max} is introduced for numerical stability.

The augmented Lagrangian method is guaranteed to find a local minimizer θ_n, under a certain set of assumptions [Nocedal and Wright, 2006, Theorem 17.6]. One of these is constraint qualification at the minimizer, in this case demanding $\nabla c(\theta_*, s_*) \neq 0$ at the optimal primal variables θ_*, s_*. For $\epsilon = 0$ this do not hold, but it does so for $\epsilon > 0$, see Lemma 3 in the supplementary material for a proof. Finding the minimum for $\epsilon \to 0$ will thus require $\rho \to \infty$, and we have introduced the stop condition ρ_{max} on line 3 for practical reasons.

To compute $\gamma_\circ$ we replace $\mathbb{E}_n[..]$ in (31) with $\mathbb{E}[..]$, which has a closed-form expression.

4.2 BASELINE COMPARISON

We first compare the proposed confidence interval $\Gamma_{n,\alpha}$ in (19) with a standard OLS-based confidence interval $B_{n,\alpha}$ for (3) that is computed using HC0 standard errors [Wooldridge, 2010]. To use OLS we must specify a set of control variables, which we take to be z. When this set is valid, we expect $\Gamma_{n,\alpha}$ and $B_{n,\alpha}$ to be similar. When the set is invalid, we expect them to diverge.

We use the linear Gaussian data model with the matrix in (5) set to be either

$$W' = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \text{ or } W'' = \begin{bmatrix} 0 & 0.4 & 0 \\ 0 & 0 & 0 \\ 0.7 & 0.2 & 0 \end{bmatrix}$$

The graph of W' is illustrated in Figure 1a, while Figure 1b demonstrates the ability of $\Gamma_{n,\alpha}$ to correctly infer $\gamma_\circ$ without specifying a set of control variables. By contrast, $B_{n,\alpha}$ is clearly biased from incorrectly controlling for the collider z_1.

Corresponding results for W'' are shown in Figure 2. As expected, the resulting intervals $\Gamma_{n,\alpha}$ and $B_{n,\alpha}$ are virtually identical since z constitutes a valid set of control variables.

4.3 CALIBRATION AND NORMALITY

To assess the calibration of $\Gamma_{\alpha,n}$, we set n to be 10^2 or 10^4 and generate repeated datasets from a linear Gaussian data model with matrix

$$W = \begin{bmatrix} 0 & -2 & 1.6 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1.2 & 0 & -0.5 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

corresponding to a graph illustrated in Figure 3.

The coverage probability $\mathbb{P}(\gamma_\circ \in \Gamma_{\alpha,n})$ was estimated to be 94.6% and 94.9% for $n = 10^2$ and 10^4, respectively, using

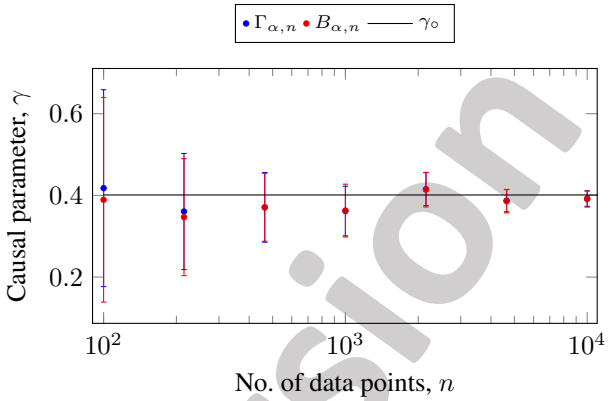


Figure 2: $(1-\alpha)$-confidence intervals for $\gamma_\circ$ computed under a linear Gaussian SCM with matrix W'', for which z is valid control variable.

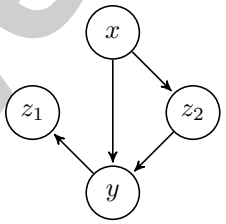


Figure 3: Causal structure of W in experiment for Calibration and Normality check, where $z = [z_1,\ z_2]$ is not a valid set of control variables.

1000 Monte Carlo simulations. This is close to $1-\alpha = 95\%$ and corroborates Theorem 4. Figure 4 supports the result further by showing a Normality plot for the point estimate γ_n over all Monte Carlo simulations.

4.4 COMPARISON WITH A CAUSAL DISCOVERY METHOD

We compare our method with an alternative method of inferring the average causal effect by learning a linear SCM adjacency matrix W using DirectLiNGAM [Shimizu et al., 2011, Hyvärinen and Smith, 2013]. Then we can compute bootstrap confidence intervals, although they lack theoretical coverage guarantees. We used the official python implementation, version 1.5.1 from PyPI `https://pypi.org/project/lingam/1.5.1/`.

We generate a random adjacency matrix W for a graph on $d = 10$ nodes and $k = 1$, but with the random seed set to the lowest nonnegative integer that yielded a nonzero γ to make the comparison interesting. We use $n = 10^4$ observations.

For LiNGAM, we computed the confidence interval (CI) using 100 bootstrap samples. For a comparable evaluation of its coverage, we considered the target quantity $\gamma_\circ$ to be the effect obtained when using LiNGAM with a large numbere

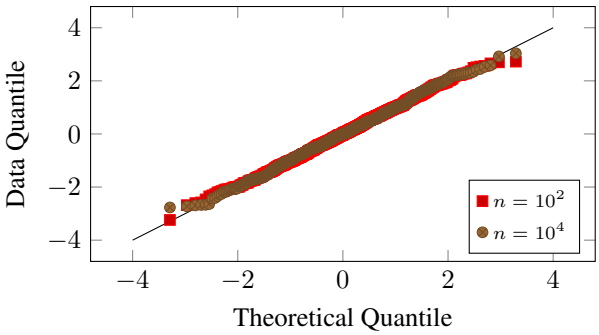


Figure 4: Normal probability plot for realizations of γ_n. Approximate normality is achieved even under moderate sample sizes.

Table 1: Comparison of empirical coverage rate (CR) and the average width of the Confidence Interval (CI) for LiNGAM Bootstrap CI and the CI $\Gamma_{\alpha,n}$ proposed in this article. The nominal CR was set to exceed $1 - \alpha = 95\%$

Noise	Method	CR	Avg CI width	Avg γ_n
Normal	LiNGAM	100%	2.01	0.64
	our	99%	0.15	1.79
Exp	LiNGAM	92%	0.08	1.77
	our	100%	0.54	1.79
Gumbel	LiNGAM	85%	0.07	1.77
	our	100%	0.46	1.79

of data points ($n' = 10^6$). 100 Monte Carlo runs were used and the results are presented in Table 1.

The results show that when data is Gaussian, our proposed method yields both well-calibrated and tighter CIs, than LiNGAM method which has a very wide CI. This expected as LiNGAM was designed for non-Gaussian data. Indeed, for the non-Gaussian examples, LiNGAM produces tighter CIs but they all undercover. By constrast, our method produces more conservative CIs that do not undercover and yield consistent inferences.

4.5 SENSITIVITY WITH RESPECT TO DAG TOLERANCE

Let $\gamma_\circ(\epsilon)$ denote the average causal effect (10a) when setting a specific value ϵ in (10b). When data-generating process is given by a linear SCM(5), we have that the approximation gap $|\gamma - \gamma_\circ(0)| = 0$, where γ is given by (9). The gap should decrease with ϵ such that ideally $\lim_{\epsilon\to 0} |\gamma - \gamma_\circ(\epsilon)| = 0$ and, moreover. An analytical study is, however, beyond the scope of the tools considered herein and we therefore resort to a numerical sensitivity study.

First, we generate random DAG-matrices W. For every W, we form the numerically approximation $\hat{\gamma}_\circ(\epsilon)$ by replacing

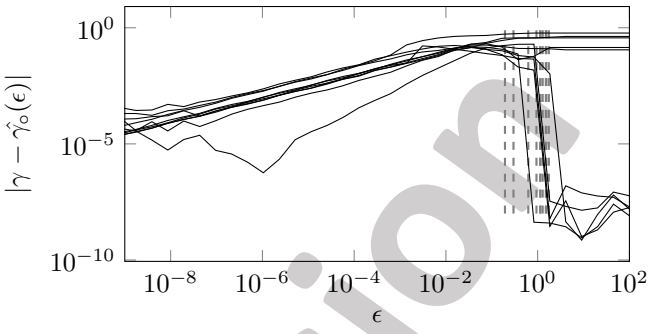


Figure 5: The error between γ (9) for a randomly generated matrix W and the numerically evaluated $\hat{\gamma}_\circ(\epsilon)$ from (10a) and (10b), over a range of ϵ. Each solid line corresponds to the error for a randomly drawn matrix, with a corresponding value of $\epsilon_\star$ shown as a vertical grey dashed line. For $\epsilon \lesssim 10^{-7}$ the numerical precision of our numerical solver limits the precision of the results.

$\mathbb{E}_n$ with the closed for expression for $\mathbb{E}$ in (31). In Figure 5, we illustrate the approximation gap $|\gamma - \hat{\gamma}_\circ(\epsilon)|$. As expected the gap decreases sharply with ϵ, until we reach finite precision effects arising mainly from the L-BFGS-B implementation.

For some of the random matrices, we notice that when $\epsilon > \epsilon_\star$ we obtain unreliable approximations. A more detailed discussion is provided in Section 1.2.1 in the supplementary material.

In the work of Ng et al. [2020], it is shown that the convergence guarantees for augmented Lagrangian method do not hold and that its precision is finite as it terminates when the quadratic penalty ρ approaches infinity — in agreement both with our theoretical and experimental results.

5 CONCLUSION

We have developed a method that is capable of inferring average causal effects without the need to specify valid control variables, when the data-generating process can be described by a linear SCM. The methodology is based on characterizing DAG-structures, which involve combinatorial constraints, using a continuously differentiable constraint. By considering a class of almost-DAG matrices, we derive an asymptotically valid confidence interval building on a theory of equality-constrained M-estimation. The theoretical results were further corroborated in numerical studies with synthetic data.

Further research includes developing numerical search methods that are better tailored to approximate the constrained M-estimator upon which the confidence interval is based. Another research direction is the study of the properties of (10b) when $\epsilon \in (0, \epsilon_\star)$.

Author Contributions

Ludvig Hult made the numerical simulations, the theoretical derivations and typeset the technical parts as well as produced all figueres and diagram. All code is due to Ludvig Hult.

Dave Zachariah concieved the idea, guided the work and supported the article authoring.

Acknowledgements

This work was partly supported by the Swedish Research Council under contract 2018-05040 and the *Wallenberg AI, Autonomous Systems and Software Program (WASP)* funded by Knut and Alice Wallenberg Foundation.

References

Donald W. K. Andrews. Estimation When a Parameter is on a Boundary. *Econometrica*, 67(6):1341–1383, November 1999. doi: 10.1111/1468-0262.00082.

Joshua David Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press, Princeton, 2009. ISBN 978-0-691-12034-8 978-0-691-12035-5.

Jeremy B. Bernerth and Herman Aguinis. A critical review and best-practice recommendations for control variable usage. *Personnel Psychology*, 69(1):229–283, Feb 2016. doi: 10.1111/peps.12103.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In Larochelle et al. [2020].

Kevin D. Carlson and Jinpei Wu. The illusion of statistical control: Control variable practice in management research. *Organizational Research Methods*, 15 (3):413–435, Jul 2012. doi: 10.1177/1094428111428817.

Charles J. Geyer. On the Asymptotics of Constrained M-Estimation. *The Annals of Statistics*, 22(4):1993–2010, December 1994. doi: 10.1214/aos/1176325768.

Aapo Hyvärinen and Stephen M. Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(1), 2013. ISSN 1532-4435.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. In Larochelle et al. [2020].

Trent Kyono and Yao Zhang. Castle: Regularization via auxiliary causal graph discovery. In Larochelle et al. [2020].

H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020.

Po-Ling Loh and Peter Bühlmann. High-Dimensional Learning of Linear Causal Networks via Inverse Covariance Estimation. *Journal of Machine Learning Research*, 15(88):3065–3105, 2014. URL `http://jmlr.org/papers/v15/loh14a.html`.

Ignavier Ng, Sébastien Lachapelle, Nan Rosemary Ke, and Simon Lacoste-Julien. On the convergence of continuous constrained optimization for structure learning, Nov 2020. arXiv: 2011.11150, presented at NeurIPS 2020 Workshop on Causal Discovery and Causality-Inspired Machine Learning.

Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer series in operations research. Springer, New York, 2nd ed edition, 2006. ISBN 978-0-387-30303-1.

Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, September 2009. ISBN 978-1-139-64398-6.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachuestts, 2017. ISBN 978-0-262-03731-0.

Alexander Shapiro. On the asymptotics of constrained local M-estimators. *The Annals of Statistics*, 28(3):948–960, May 2000. doi: 10.1214/aos/1015952006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12(33):1225–1248, 2011. URL `http://jmlr.org/papers/v12/shimizu11a.html`.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*, volume 81 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 1993. doi: 10.1007/978-1-4612-2748-9.

P. Stoica and C.B. Ng. On the Cramer-Rao bound under parametric constraints. *IEEE Signal Processing Letters*, 5(7):177–179, July 1998. doi: 10.1109/97.700921.

A. W. van der Vaart. M- and Z-Estimators. In *Asymptotic Statistics*. Cambridge University Press, 1 edition, October 1998. doi: 10.1017/CBO9780511802256.

Jinde Wang. Asymptotics of least-squares estimators for constrained nonlinear regression. *The Annals of Statistics*, 24(3):1316–1326, June 1996. doi: 10.1214/aos/1032526971.

Dennis Wei, Tian Gao, and Yue Yu. Dags with no fears: A closer look at continuous optimization for learning bayesian networks. In Larochelle et al. [2020].

Jeffrey M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, Cambridge, Mass, 2nd ed edition, 2010. ISBN 978-0-262-23258-6.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7154–7163. PMLR, 09–15 Jun 2019. URL `http://proceedings.mlr.press/v97/yu19a.html`.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32 proceedings (NeurIPS 2018)*, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Learning sparse nonparametric DAGs. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020*, volume 208, pages 3414–3425. PMLR, 2020.

Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, December 1997. doi: 10.1145/279232.279236.