ON CAUSAL DISCOVERY WITH EQUAL VARIANCE ASSUMPTION

WENYU CHEN, MATHIAS DRTON, AND Y. SAMUEL WANG

Abstract. Prior work has shown that causal structure can be uniquely identified from observational data when these follow a structural equation model whose error terms have equal variances. We show that this fact is implied by an ordering among (conditional) variances. We demonstrate that ordering estimates of these variances yields a simple yet state-of-the-art method for causal structure learning that is readily extendable to high-dimensional problems.

1. Introduction

A structural equation model for a random vector $X = (X_1, \ldots, X_p)$ postulates causal relations in which each variable X_j is a function of a subset of the other variables and a stochastic error ε_j. Causal discovery/structure learning is the problem of inferring which of other variables each variable X_j depends on. We consider this problem where only observational data, that is, a sample from the joint distribution of X, is available. While in general only an equivalence class of structures can then be inferred (Pearl, 2009; Spirtes et al., 2000), recent work stresses that unique identification is possible under assumptions such as non-linearity with additive errors, linearity with non-Gaussian errors, and linearity with errors of equal variance; see the reviews of Drton and Maathuis (2017) and Heinze-Deml et al. (2018) or the book of Peters et al. (2017).

This note is concerned with the equal variance case treated by Peters and Bühlmann (2014) and Loh and Bühlmann (2014) who prove identifiability of the causal structure and propose greedy search methods for its estimation. Our key observation is that the identifiability is implied by an ordering among certain conditional variances. Ordering estimates of these variances yields a fast method for estimation of the causal ordering of the variables. The precise causal structure can then be inferred using variable selection techniques for regression (Shojaie and Michailidis, 2010). Specifically, we develop a top-down approach that infers the ordering by successively identifying sources. The method is developed for low- as well as high-dimensional problems. Simulations show significant gains in computational efficiency when compared with greedy search and increased accuracy when the number of variables p is large.

An earlier version of this note also included a bottom-up method which identified the causal ordering by successively finding sinks via minimal precisions. However, after the note was finished, we became aware of Ghoshal and Honorio (2018) who proposed a similar bottom-up approach. We emphasize that our top-down approach only requires control of the maximum in-degree as opposed to the bottom-up approach which requires control of the maximum Markov blanket. This is discussed further in Section 4.2 and a direct numerical comparison is given in Section 5.2.

2. Structural Equation Models and Directed Acyclic Graphs

Suppose, without loss of generality, that the observed random vector $X = (X_1, \ldots, X_p)$ is centered. In a linear structural equation model, X then solves an equation system

$$X_j = \sum_{k \neq j} \beta_{jk} X_k + \varepsilon_j, \qquad j = 1, \ldots, p, \tag{1}$$

where the ε_j are independent random variables with mean zero, and the coefficients β_{jk} are unknown parameters. Following Peters and Bühlmann (2014), we assume that all ε_j have a common unknown variance $\sigma^2 > 0$. We will write $X \sim (B, \sigma^2)$ to express the assumption that there indeed exist

independent errors $\varepsilon_1, \ldots, \varepsilon_p$ of equal variance σ^2 such that X solves (1) for coefficients given by a real $p \times p$ matrix $B = (\beta_{jk})$ with zeros along the diagonal.

The causal structure inherent to the equations in (1) is encoded in a directed graph $\mathcal{G}(B)$ with vertex set $V = \{1, \ldots, p\}$ and edge set $E(B)$ equal to the support of B. So, $E(B) = \{(k, j) : \beta_{jk} \neq 0\}$. Inference of $\mathcal{G}(B)$ is the goal of causal discovery as considered in this paper. As in related work, we assume $\mathcal{G}(B)$ to be a directed acyclic graph (DAG) so that B is permutation similar to a triangular matrix. Then (1) admits the unique solution $X = (I - B)^{-1}\varepsilon$ where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)$. Hence, the covariance matrix of $X \sim (B, \sigma^2)$ is

$$\Sigma := \mathbb{E}(XX^T) = \sigma^2 (I - B)^{-1}(I - B)^{-T}. \tag{2}$$

We will invoke the following graphical concepts. If the considered graph $\mathcal{G}$ contains the edge $k \to j$, then k is a parent of its child j. We write $\mathrm{pa}(j)$ for the set of all parents of a node j. Similarly, $\mathrm{ch}(j)$ is the set of children of j. If there exists a directed path $k \to \ldots \to j$, then k is an ancestor of its descendant j. The sets of ancestors and descendants of j are $\mathrm{an}(j)$ and $\mathrm{de}(j)$, respectively. Here, $j \in \mathrm{an}(j)$ and $j \in \mathrm{de}(j)$. A set of nodes C is ancestral if $\mathrm{an}(j) \subseteq C$ for all $j \in C$. If $\mathcal{G}$ is a DAG, then it admits a topological ordering of its vertices. In other words, there exists a numbering σ such that $\sigma(j) < \sigma(k)$ only if $k \notin \mathrm{an}(j)$. Finally, every DAG contains at least one source, that is, a node j with $\mathrm{pa}(j) = \emptyset$. Similarly, every DAG contains at least one sink, which is a node j with $\mathrm{ch}(j) = \emptyset$.

3. Identifiability by Ordering Variances

The main result of Peters and Bühlmann (2014) shows that the graph $\mathcal{G}(B)$ and the parameters B and σ^2 are identifiable from the covariance in (2). No faithfulness assumptions are needed.

Theorem 1. *Let $X \sim (B_X, \sigma_X^2)$ and $Y \sim (B_Y, \sigma_Y^2)$ with both $\mathcal{G}(B_X)$ and $\mathcal{G}(B_Y)$ directed and acyclic. If $\mathrm{var}(X) = \mathrm{var}(Y)$, then $\mathcal{G}(B_X) = \mathcal{G}(B_Y)$, $B_X = B_Y$, and $\sigma_X^2 = \sigma_Y^2$.*

In this section we first give an inductive proof of Theorem 1 that proceeds by recursively identifying source nodes for $\mathcal{G}(B)$ and subgraphs. We then clarify that alternatively one could identify sink nodes. Our first lemma clarifies that the sources in $\mathcal{G}(B)$ are characterized by minimal variances. We define

$$\zeta \equiv \zeta(B) \;=\; \min_{(k,j) \in E(B)} \beta_{jk}^2. \tag{3}$$

Lemma 1. *Let $X \sim (B, \sigma^2)$ with $\mathcal{G}(B)$ directed and acyclic. If $\mathrm{pa}(j) = \emptyset$, then $\mathrm{var}(X_j) = \sigma^2$. If $\mathrm{pa}(j) \neq \emptyset$, then $\mathrm{var}(X_j) \geq \sigma^2(1 + \zeta) > \sigma^2$.*

Proof. Let $\Pi = (\pi_{jk}) = (I - B)^{-1}$. Each total effect π_{jk} is the sum over directed paths from k to j of products of coefficients β_{ab} along each path. In particular, $\pi_{jj} = 1$. From (2), $\mathrm{var}(X_j) = \sigma^2 \sum_{k=1}^p \pi_{jk}^2$. Hence, if $\mathrm{pa}(j) = \emptyset$, then $\mathrm{var}(X_j) = \sigma^2$ because $\pi_{jk}^2 = 0$ for all $k \neq j$. If $\mathrm{pa}(j) \neq \emptyset$ then by acyclicity of $\mathcal{G}(B)$ there exists a node $\ell \in \mathrm{pa}(j)$ such that $\mathrm{de}(\ell) \cap \mathrm{pa}(j) = \{\ell\}$. Then $\pi_{j\ell}^2 = \beta_{j\ell}^2 \geq \zeta$ and

$$\mathrm{var}(X_j) = \sigma^2 \left(1 + \sum_{k \neq j} \pi_{jk}^2\right) \geq \sigma^2 \left(1 + \pi_{j\ell}^2\right) \geq \sigma^2 \left(1 + \zeta\right).$$

□

The next lemma shows that by conditioning on a source, or more generally an ancestral set, one recovers a structural equation model with equal error variance whose graph has the source node or the entire ancestral set removed. For a variable X_j and a vector $X_C = (X_k : k \in C)$, we define $X_{j.C} = X_j - \mathbb{E}(X_j | X_C)$.

Lemma 2. *Let $X \sim (B, \sigma^2)$ with $\mathcal{G}(B)$ directed and acyclic. Let C be an ancestral set in $\mathcal{G}(B)$. Then $(X_{j.C} : j \notin C) \sim (B[-C], \sigma^2)$ for submatrix $B[-C] = (\beta_{jk})_{j,k \notin C}$.*

Algorithm 1: Topological Ordering: General procedure with criterion f

Input : $\hat{\Sigma} \in \mathbb{R}^{p\times p}$ (estimated) covariance of X
Output: Θ

$\Theta^{(0)} \leftarrow \emptyset$;
for $z = 1, \dots, p$ **do**
  $\theta \leftarrow \arg\min_{j \in V \setminus \Theta^{(z-1)}} f(\hat{\Sigma}, \Theta^{(z-1)}, j)$;
  Append θ to $\Theta^{(z-1)}$ to form $\Theta^{(z)}$
return the ordered set $\Theta^{(p)}$.

Proof. Let $j \notin C$. Since C is ancestral, X_C is a function of ε_C only and thus independent of ε_j. Hence, $\mathbb{E}(\varepsilon_j|X_C) = \mathbb{E}(\varepsilon_j) = 0$. Because it also holds that $X_{k.C} = 0$ for $k \in C$, we have from (1) that

$$X_{j.C} = \sum_{k \in \mathrm{pa}(j)\setminus C} \beta_{jk} X_{k.C} + \varepsilon_j.$$

□

The lemmas can be combined to identify a topological ordering of $\mathcal{G}(B)$ and prove Theorem 1.

of Theorem 1. The claim is trivial for $p = 1$ variables, which gives the base for an induction on p. If $p > 1$, then Lemma 1 identifies a source c by variance minimization. Conditioning on c as in Lemma 2 reduces the problem to size $p - 1$. By the induction assumption, σ^2 and $B[-\{c\}]$ can be identified. The regression coefficients in the conditional expectations $\mathbb{E}(X_j|X_c)$ for $j \neq c$ identify the missing first row and column of B; see e.g. Drton (2018, §7). □

Next, we show that alternatively one may minimize precisions to identify a sink node. We state analogues of Lemma 1 and 2 which can also be used to prove Theorem 1.

Lemma 3. *Let $X \sim (B, \sigma^2)$ with $\mathcal{G}(B)$ directed and acyclic. Let Σ be the covariance matrix of X, and $\Phi = \Sigma^{-1}$ the precision matrix. If $\mathrm{ch}(j) = \emptyset$, then $\Phi_{jj} = 1/\sigma^2$. If $\mathrm{ch}(j) \neq \emptyset$, then $\Phi_{jj} \geq \{1 + \zeta|\mathrm{ch}(j)|\}/\sigma^2 > 1/\sigma^2$.*

Proof. The diagonal entries of $\Phi = \frac{1}{\sigma^2}(I - B)(I - B)^T$ are $\Phi_{jj} = \frac{1}{\sigma^2}(1 + \sum_{k\in\mathrm{ch}(j)} \beta_{kj}^2)$. So $\Phi_{jj} = 1/\sigma^2$ if $\mathrm{ch}(j) = \emptyset$, and $\Phi_{jj} \geq \{1 + |\mathrm{ch}(j)|\zeta\}/\sigma^2$ if $\mathrm{ch}(j) \neq \emptyset$. □

Marginalization of a sink is justified by the following well-known fact (e.g. Drton, 2018, §5).

Lemma 4. *Let $X \sim (B, \sigma^2)$ with $\mathcal{G}(B)$ directed and acyclic. Let C be an ancestral set in $\mathcal{G}(B)$. Then $X_C \sim (B[C], \sigma^2)$ for submatrix $B[C] = (\beta_{jk})_{j,k\in C}$.*

4. Estimation Algorithms

4.1. **Low-dimensional Problems.** The results from Section 3 naturally yield an iterative top-down algorithm for estimation of a topological ordering for $\mathcal{G}(B)$. In each step of the procedure we select a source node by comparing variances conditional on the previously selected variables, so the criterion in the minimization in Algorithm 1 is the variance

$$f_1(\hat{\Sigma}, \Theta, j) \;=\; \hat{\Sigma}_{j,j} - \hat{\Sigma}_{j,\Theta}\hat{\Sigma}_{\Theta,\Theta}^{-1}\hat{\Sigma}_{\Theta,j} \;=\; \frac{1}{\{(\hat{\Sigma}_{\Theta\cup\{j\},\Theta\cup\{j\}})^{-1}\}_{j,j}}, \tag{4}$$

where $\hat{\Sigma}$ is the sample covariance matrix. Alternatively, and as also observed by Ghoshal and Honorio (2018), a bottom-up procedure could construct the reverse causal ordering by successively minimizing precisions (or in other words, full conditional variances).

To facilitate theoretical statements about our top-down procedure, we assume that the errors ε_j in (1) are all sub-Gaussian with maximal sub-Gaussian parameter $\gamma > 0$. We indicate this by writing

$X \sim (B, \sigma^2, \gamma)$. Our analysis is restricted to inference of a topological ordering. Shojaie and Michailidis (2010) give results on lasso-based inference of the graph given an ordering.

Theorem 2. *Let $X \sim (B, \sigma^2, \gamma)$ with $\mathcal{G}(B)$ directed and acyclic. Suppose the covariance matrix $\Sigma = \mathbb{E}(XX^T)$ has minimum eigenvalue $\lambda_{\min} > 0$. If*

$$n \;>\; p^2 \left\{\log(p^2+p) - \log(\epsilon/2)\right\} 128 \left(1 + 4\frac{\gamma^2}{\sigma^2}\right)^2 \left(\max_{j \in V} \Sigma_{j,j}\right)^2 \left(\frac{\zeta\lambda_{\min} + 2\sigma^2}{\zeta\lambda_{\min}^2}\right)^2,$$

then Algorithm 1 using criterion criterion (4) *recovers a topological ordering of $\mathcal{G}(B)$ with probability at least $1 - \epsilon$.*

The result follows using concentration for sample covariances (Ravikumar et al., 2011, Lemma 1) and error propagation analysis as in Harris and Drton (2013, Lemma 5). We give details in Appendix A, which is found in the supplementary materials.

4.2. **High-dimensional Problems.** The consistency result in Theorem 2 requires the sample size n to exceed a multiple of $p^2 \log(p)$ and only applies to low-dimensional problems. If $p > n$, method will stop at the nth step when the conditional variance in (4) becomes zero for all $j \notin \Theta$.

However, in the high-dimensional setting if $\mathcal{G}(B)$ has maximum in-degree bounded by a small integer q, we may modify the criterion from (4) to

$$f_2(\hat{\Sigma}, \Theta, j) \;=\; \min_{C \subseteq \Theta, |C| = q} f_1(\hat{\Sigma}, C, j) \;=\; \min_{C \subseteq \Theta, |C| = q} \hat{\Sigma}_{j,j} - \hat{\Sigma}_{j,C}(\hat{\Sigma}_{C,C})^{-1}\hat{\Sigma}_{C,j}. \tag{5}$$

The intuition is that in the population case, adjusting by a smaller set $C \subseteq \Theta^{(z)}$ with $\mathrm{pa}(j) \subseteq C$ yields the same results as adjusting by all of $\Theta^{(z)}$. The next lemma makes the idea rigorous.

Lemma 5. *Let $X \sim (B, \sigma^2)$ with $\mathcal{G}(B)$ directed and acyclic with maximum in-degree at most q. Let $\Sigma = \mathbb{E}(XX^T)$, and suppose $S \subseteq V \setminus \{j\}$ is an ancestral set. If $\mathrm{pa}(j) \subseteq S$, then $f_2(\Sigma, S, j) = \sigma^2$. If $\mathrm{pa}(j) \not\subseteq S$, then $f_2(\Sigma, S, j) \geq \sigma^2(1 + \zeta)$.*

Proof. The conditional variance of X_j given X_S is the variance of the residual $X_{j.S}$. By Lemma 2, $X_{j.S}$ has the same distribution as X'_j when $X' \sim (B[-S], \sigma^2)$. Now, j is a source of $\mathcal{G}(B[-S])$ if and only if $\mathrm{pa}(j) \subseteq S$. Lemma 1 implies that $\mathrm{var}(X_j | X_C) = \sigma^2$ if $\mathrm{pa}(j) \subseteq S$ and $\mathrm{var}(X_j | X_C) \geq \sigma^2(1 + \zeta)$ otherwise. The claim about $f_2(\Sigma, S, j)$ now follows. □

Based on Lemma 5, we have the following result whose proof is analogous to that of Theorem 2. The key feature of the result is a drop from p^2 to $(q+1)^2$ in the sample size requirement.

Theorem 3. *Let $X \sim (B, \sigma^2, \gamma)$ with $\mathcal{G}(B)$ directed and acyclic with of maximum in-degree at most q. Suppose all $(q+1) \times (q+1)$ principal submatrices of $\Sigma = \mathbb{E}\left(XX^T\right)$ have minimum eigenvalue at least $\lambda_{\min} > 0$. If*

$$n \;>\; (q+1)^2 \left\{\log(p^2+p) - \log(\epsilon/2)\right\} 128 \left(1 + 4\frac{\gamma^2}{\sigma^2}\right)^2 \left(\max_{j \in V} \Sigma_{j,j}\right)^2 \left(\frac{\zeta\lambda_{\min} + 2\sigma^2}{\zeta\lambda_{\min}^2}\right)^2,$$

then Algorithm 1 using criterion (5) *recovers a topological ordering of $\mathcal{G}(B)$ with probability at least $1 - \epsilon$.*

We contrast our guarantees with those for the bottom-up method of Ghoshal and Honorio (2018) which selects sinks by minimizing conditional precisions that are estimated using the CLIME estimator (Cai et al., 2011). Because CLIME requires small Markov blankets, the bottom-up procedure has sample complexity $\mathcal{O}\left(d^8 \log(p)\right)$ where d is the maximum total degree. This implies that the procedure cannot consistently discover graphs with hubs, i.e., nodes with very large out-degree, in the high dimensional setting. This said, the computational complexity of the bottom-up procedure is polynomial in d, while our top-down procedure is exponential in the maximum in-degree. In practice, we use a branch-and-bound procedure (Lumley, 2017) to efficiently select the set which minimizes the conditional variance; see Section 5.2.

TABLE 1. Low-dimensional dense settings

		Kendall's τ			Recall %			Flipped %			FDR %		
p	n	TD	BU	GDS	TD	BU	GDS	TD	BU	GDS	TD	BU	GDS
5	100	0.85	0.82	0.88	91	89	91	7	8	6	17	18	9
	500	0.98	0.97	0.98	99	98	99	1	1	1	4	4	2
	1000	0.99	0.98	0.99	99	99	99	1	1	1	3	3	1
20	100	0.92	0.85	0.61	85	83	62	3	5	13	32	35	43
	500	0.99	0.97	0.75	99	98	81	1	1	11	28	29	35
	1000	1.00	0.99	0.82	100	100	88	0	0	8	26	26	28
40	100	0.96	0.91	0.53	71	69	44	2	3	11	41	43	58
	500	0.99	0.98	0.59	96	96	63	0	1	14	41	42	57
	1000	1.00	0.99	0.64	97	97	71	0	0	14	40	41	57

TABLE 2. Low-dimensional sparse settings

		Kendall's τ			Recall %			Flipped %			FDR %		
p	n	TD	BU	GDS	TD	BU	GDS	TD	BU	GDS	TD	BU	GDS
5	100	0.87	0.84	0.88	91	89	90	6	7	6	16	17	9
	500	0.98	0.96	0.98	98	98	99	1	2	1	5	5	2
	1000	0.99	0.98	0.99	99	99	99	1	1	1	3	4	1
20	100	0.77	0.59	0.60	85	79	77	9	13	15	35	40	39
	500	0.96	0.88	0.77	98	96	89	2	4	10	19	22	26
	1000	0.99	0.94	0.81	100	98	90	0	2	9	14	16	23
40	100	0.72	0.44	0.47	81	72	72	10	16	20	38	46	54
	500	0.96	0.80	0.58	98	94	81	2	5	18	24	31	47
	1000	0.99	0.91	0.61	99	98	82	1	2	17	17	22	48

5. NUMERICAL RESULTS

5.1. **Low-dimensional Setting.** We first assess performance in the low-dimensional setting. Random DAGs with p nodes and a unique topological ordering are generated by: (1) always including edge $v \to v+1$ for $v < p$, and (2) including edge $v \to u$ with probability p_c for all $v < u-1$. We consider a sparse setting with $p_c = 3/(2p-2)$ and a dense setting with $p_c = 0.3$. All linear coefficients are drawn uniformly from $\pm[.3, 1]$. The error terms are standard normal. Performance is measured using Kendall's τ between rankings of variables according to the true and estimated topological orderings. Although the true graph admits a unique ordering by construction, the graph estimated by the greedy search may not admit a unique ordering. Nevertheless, the ranking of variables according to the estimated graph is unique if we allow ties, and Kendall's τ remains a good measure for all the methods. We also compute the percentage of true edges discovered (Recall), the percentage of estimated edges that are flipped in the true graph (Flipped), and the proportion of estimated edges which are either flipped or not present in the true graph (false discovery rate; FDR). Tables 1 and 2 show averages over 500 random realizations for our top-down procedure (TD), the bottom-up procedure (BU) of Ghoshal and Honorio (2018), and greedy DAG search (GDS). For the bottom up procedure in the low-dimensional setting, we may in fact simply invert the sample covariance to estimate precisions. For GDS, we allow for 5 random restarts using the same procedure as Peters and Bühlmann (2014).

In both dense and sparse settings, when $p = 5$, greedy search performs best in all metrics. However, for $p = 20$ and 40, the top-down approach does best, followed by bottom-up, and finally greedy search. The top-down and bottom-up method both have a substantially higher average Kendall's τ than greedy search.

In our experiments, the proposed methods are roughly 50 to 500 times faster than greedy search as graph size and density increases. On our personal computer, the average run time in the dense setting with $p = 40$ and $n = 1000$ is 8 seconds for the top-down and bottom-up methods, but 4,500 seconds for the greedy search.

5.2. **High-dimensional Setting.** We now test the proposed procedures in a high-dimensional setting with $p > n$ in two scenarios. Random DAGs with p nodes and a unique topological ordering are generated by: (1) always including edge $v \to v+1$ for $v < p$, and either (2a) for each $v > 2$, including $u_1, u_2 \to v$, where $u_i < v$, and u_i has out-degree $d_{\text{out}}(u_i) < 4$, or (2b) for each $v > 2$, including $u_1, u_2 \to v$, where $u_i < \min(v, 10)$. In both scenarios, the maximum in-degree is fixed to be $q = 3$. In the first scenario, it is also guaranteed that the maximum Markov blanket size is small, bounded by $k \leq 15$. In the second scenario when there exists hubs in the graph, the maximum Markov blanket size grows with p, with $k \geq 0.2p$. The errors are standard normal.

Algorithm 1 with (5) as HTD (high-dimensional top-down) and to the bottom-up method of Ghoshal and Honorio (2018) as HBU. The best subset search step in HTD is carried with subset size $q = 3$; increasing q beyond the true maximum in-degree does not change performance substantially. The HBU is tuned with $\lambda_n = 0.5\sqrt{\log(p)/n}$. Results for greedy search are not shown as computation becomes intractable when $p > 100$. Performance is measured by Kendall's τ to provide direct comparison.

TABLE 3. High-dimensional setting with maximum in-degree $q = 3$

		Small k		Hub graph	
n	p	HTD	HBU	HTD	HBU
80	$0.5n$	0.99	0.89	1.00	0.70
	$0.75n$	0.98	0.89	0.99	0.52
	n	0.95	0.87	0.95	0.39
	$1.5n$	0.84	0.83	0.77	0.25
	$2n$	0.72	0.73	0.55	0.16
100	$0.5n$	1.00	0.93	1.00	0.70
	$0.75n$	0.99	0.92	1.00	0.50
	n	0.97	0.87	0.97	0.38
	$1.5n$	0.86	0.84	0.74	0.26
	$2n$	0.73	0.78	0.63	0.12
200	$0.5n$	1.00	0.95	1.00	0.77
	$0.75n$	1.00	0.90	1.00	0.61
	n	0.99	0.79	0.99	0.48
	$1.5n$	0.87	0.74	0.80	0.20
	$2n$	0.74	0.64	0.65	0.13

Table 3 demonstrates that in the first scenario, both methods perform reasonably well when the considered graph has small Markov blanket. The HTD procedure performs the best in low-dimensional and moderately high-dimensional settings, and both methods have similar performance in very high-dimensional settings. However, when there exists nodes with very large Markov blanket, the top-down method substantially outperforms the bottom-up method.

On our personal computer, the average run time for problems of size $p = 200$ is 10 minutes for the HTD method with $q = 3$. The computational complexity of HBU is determined by the choice of tunning parameter in the precisions estimation step.

Additional simulation settings are presented in Appendix B-E in the supplement including a setting with Rademacher errors as considered by Ghoshal and Honorio (2018).

6. Discussion

In this note, we proposed a simple method for causal discovery under a linear structural equation model with equal error variances. The procedure consistently estimates a topological ordering of the underlying graph and easily extends to the high-dimensional setting where $p > n$. Simulations demonstrate that the procedure is an attractive alternative to previously considered greedy search methods in terms of both accuracy and computational effort. The advantages of the proposed procedures become especially salient as the number of considered variables increases.

In comparison to the related work of Ghoshal and Honorio (2018), our approach is computationally more demanding for graphs with higher in-degree but requires only control over the maximum in-degree of the graph as opposed to the maximum degree. We also note that as shown in simulations in Appendix E a hybrid method in which greedy search is initialized at estimates obtained from our variance ordering procedures can yield further improvements in performance.

Finally, we note that all discussed methods extend to structural equation models where the error variances are unequal, but known up to ratio. Indeed, if $\mathrm{var}(\varepsilon_j) = a_j^2\sigma^2$ for some unknown σ^2 but known $a_1, \ldots, a_p$, we may consider $\tilde{X}_j = X_j/a_j$ instead of the original variables.

Acknowledgements

This work was supported by the U.S. National Science Foundation (Grant No. DMS 1712535).

References

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):594–607.

Drton, M. (2018). Algebraic problems in structural equation modeling. In Hibi, T., editor, *The 50th Anniversary of Gröbner Bases*, Advanced Studies in Pure Mathematics. Mathematical Society of Japan. arXiv:1612.05994.

Drton, M. and Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annu. Rev. Stat. Appl.*, 4:365–393.

Ghoshal, A. and Honorio, J. (2018). Learning linear structural equation models in polynomial time and sample complexity. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1466–1475. PMLR.

Harris, N. and Drton, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.*, 14:3365–3383.

Heinze-Deml, C., Maathuis, M. H., and Meinshausen, N. (2018). Causal structure learning. *Annu. Rev. Stat. Appl.*, 5:371–394.

Loh, P.-L. and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *J. Mach. Learn. Res.*, 15:3065–3105.

Lumley, T. (2017). *leaps: Regression Subset Selection*. R package version 3.0.

Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge, second edition. Models, reasoning, and inference.

Peters, J. and Bühlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference*. MIT Press, Cambridge, MA.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980.

Shojaie, A. and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition.

Supplementary material for On Causal Discovery with Equal Variance Assumption

Appendix A. Proof of Theorem 2

We first give a lemma that addresses the estimation error for inverse covariances.

Lemma 6. *Assume $X \sim (B, \sigma^2, \gamma)$. Suppose all $(q+1)\times(q+1)$ principal submatrices of $\Sigma = \mathbb{E}(XX^T)$ have minimum eigenvalue at least $\lambda_{\min} > 0$. If for $\epsilon, \eta > 0$ we have*

$$n \;\geq\; (q+1)^2 \left\{\log(p^2+p) - \log(\epsilon/2)\right\} 128 \left(1 + 4\frac{\gamma^2}{\sigma^2}\right)^2 \left(\max_{j\in V} \Sigma_{j,j}\right)^2 \left(\frac{\eta\lambda_{\min}+1}{\eta\lambda_{\min}^2}\right)^2. \tag{6}$$

then

$$\max_{C\subseteq V, |C|\leq q+1} \|(\Sigma_{C,C})^{-1} - (\hat{\Sigma}_{C,C})^{-1}\|\infty \leq \eta$$

with probability at least $1-\epsilon$.

Proof. Let $\delta = \frac{\eta\lambda_{\min}^2}{(q+1)(\eta\lambda_{\min}+1)}$. Because $\delta < \frac{\lambda_{\min}}{q+1}$, by Lemma 5 from Harris and Drton (2013), we have

$$\max_{C\subseteq V, |C|\leq (q+1)} \|(\Sigma_{C,C})^{-1} - (\hat{\Sigma}_{C,C})^{-1}\|_\infty \;\leq\; \frac{(q+1)\delta/\lambda_{\min}^2}{1-(q+1)\delta/\lambda_{\min}} \;=\; \eta$$

provided $\|\widehat{\Sigma} - \Sigma\|_\infty \leq \delta$. The proof is thus complete if we show that $\mathbb{P}\|\widehat{\Sigma} - \Sigma\|_\infty > \delta \leq \epsilon$.

Note that $X_j = \varepsilon_j + \sum_{k\in \mathrm{an}(j)} \pi_{jk}\varepsilon_k$ has variance $\sigma^2(1 + \sum_{k\in\mathrm{an}(j)} \pi_{jk}^2)$. Since γ is a bound on the sub-Gaussian parameters of all ϵ_l, it follows that $X_j/\sqrt{\mathrm{var}(X_j)}$ is sub-Gaussian with parameter at most γ/σ. Lemma 1 of Ravikumar et al. (2011) applies and gives

$$\mathbb{P}\{|\widehat{\Sigma}_{i,j} - \Sigma_{i,j}| > \delta\} \leq 4\exp\left\{-\frac{n\delta^2}{128(1+4\gamma^2/\sigma^2)^2 \max_j(\Sigma_{j,j})^2}\right\} \leq \frac{2}{p(p+1)}\epsilon.$$

A union bound over the entries of Σ yields that indeed $\mathbb{P}\left(\|\widehat{\Sigma} - \Sigma\|_\infty > \delta\right) \leq \epsilon$. □

of Theorem 2. Our assumption on n is as in (6) with $\eta = \zeta/(2\sigma^2)$. Lemma 6 thus implies that, with probability at least $1-\epsilon$, we have for all subsets $\Theta \subseteq V$ with $|\Theta| < q+1$ that

$$\|(\hat{\Sigma}_{\Theta,\Theta})^{-1} - (\Sigma_{\Theta,\Theta})^{-1}\|_\infty \leq \frac{\zeta}{2\sigma^2}. \tag{7}$$

Let j be a source in $\mathcal{G}(B)$, and let k be a non-source. Note that variance of j conditional on some set C_1 is

$$\sigma^2_{j|C_1} = \frac{1}{\left\{(\Sigma_{C_1\cup\{j\},C_1\cup\{j\}})^{-1}\right\}_{j,j}}.$$

By Lemma 5, for any $C_1, C_2 \subseteq \Theta \subseteq V\setminus\{j,k\}$ such that Θ is an ancestral set and $\mathrm{pa}(j) \subseteq C_1$

$$\left\{(\Sigma_{C_1\cup\{j\},C_1\cup\{j\}})^{-1}\right\}_{j,j} - \left\{(\Sigma_{C_2\cup\{k\},C_2\cup\{k\}})^{-1}\right\}_{k,k} \geq \frac{1}{\sigma^2} - \frac{1}{\sigma^2(1+\zeta)} \geq \frac{\zeta}{\sigma^2} \tag{8}$$

Using (7), when $|C_1|$ and $|C_2|$ are both at most q, we obtain that

$$\left\{(\hat{\Sigma}_{C_1\cup\{j\},C_1\cup\{j\}})^{-1}\right\}_{j,j} - \left\{(\hat{\Sigma}_{C_2\cup\{k\},C_2\cup\{k\}})^{-1}\right\}_{k,k} - \frac{\zeta}{\sigma^2} \;>\; 0. \tag{9}$$

Thus $\hat{\sigma}^2_{j|C_1} - \hat{\sigma}^2_{k|C_2} > 0$ which implies that Algorithm 1 correctly selects a source node at each step. On the first step, $\Theta = \emptyset$ which is trivially an ancestral set. By induction, each subsequent step then correctly adds a sink to Θ so Θ remains ancestral and a correct ordering is recovered. □

Appendix B. Simulations as in Peters and Bühlmann (2014)

We revisit the simulation study of Peters and Bühlmann (2014). DAGs are generated by first creating a random topological ordering, then between any two nodes, an edge is included with probability p_c. We simulate a sparse setting with $p_c = 3/(2p-2)$ and a dense setting with $p_c = 0.3$. The linear coefficients are drawn uniformly from $[-1,-.1] \cup [.1,1]$ and the errors are drawn from a standard Gaussian distribution. Since there may not be a unique ordering for the true graph, we compute the Hamming distance between the true and estimated adjacency matrix rather than Kendall's τ.

Tables 4 and 5 demonstrate that in both settings, the greedy algorithm performs better when p is small. However, when $p = 40$ the proposed algorithms infer the graph more accurately. In the dense setting, the proposed methods have similar FDR to greedy search, but substantially higher recall. In the sparse setting, the proposed methods have lower recall than greedy search, but also substantially lower FDR.

Table 4. Dense setting

		Hamming Dist.			Recall %			Flipped %			FDR %		
p	n	TD	BU	GDS	TD	BU	GDS	TD	BU	GDS	TD	BU	GDS
5	100	1.3	1.3	1.1	73	73	78	7	7	7	16	15	18
	500	0.7	0.7	0.5	80	80	88	4	4	5	8	7	9
	1000	0.5	0.5	0.4	85	84	92	3	3	5	5	5	7
20	100	31	32	30	73	73	74	4	3	6	27	28	25
	500	22	22	14	91	91	91	2	3	4	24	24	13
	1000	28	28	8	94	94	96	2	2	2	21	21	10
40	100	170	174	215	66	65	54	2	3	8	36	37	45
	500	152	155	186	93	93	76	2	2	9	38	39	42
	1000	136	137	168	96	95	83	1	1	8	36	36	38

Table 5. Sparse setting

		Hamming Dist.			Recall %			Flipped %			FDR %		
p	n	TD	BU	GDS	TD	BU	GDS	TD	BU	GDS	TD	BU	GDS
5	100	1.6	1.7	1.4	74	73	78	8	8	8	18	18	17
	500	0.8	0.9	0.6	85	84	91	3	4	5	7	7	9
	1000	0.6	0.6	0.4	88	88	94	3	4	5	6	6	7
20	100	7	7	12	69	69	81	4	4	6	16	17	43
	500	3.5	3.5	4.5	85	84	93	4	4	4	9	8	21
	1000	2.2	2.2	2.8	90	90	97	3	2	3	5	5	14
40	100	14	15	45	64	63	78	3	4	8	16	18	62
	500	7	7	16	84	84	94	3	3	3	8	7	33
	1000	5	5	10	90	89	97	3	3	3	6	6	24

Appendix C. Simulations as in Ghoshal and Honorio (2018)

We construct random graphs as in Section 5.2, but we follow the data sampling procedure as used in Ghoshal and Honorio (2018). All linear coefficients are drawn uniformly from $\pm[.5,1]$, and errors are drown from the Rademacher distribution and scaled to have $\sigma_i^2 = 0.8$. Table 6 demonstrates that both methods performs reasonably well when Markov blankets are restricted to be small, and the top-down approach performs substantially better when there are hubs.

TABLE 6. High-dimensional setting with Rademacher noise and maximum in-degree $q = 3$

		Small k		Hub graph	
n	p	HTD	HBU	HTD	HBU
80	$0.5n$	0.99	0.95	0.98	0.73
	$0.75n$	0.98	0.90	0.89	0.46
	n	0.96	0.90	0.76	0.36
	$1.5n$	0.84	0.86	0.52	0.23
	$2n$	0.71	0.80	0.35	0.10
100	$0.5n$	0.99	0.97	0.99	0.69
	$0.75n$	0.99	0.95	0.92	0.46
	n	0.96	0.93	0.76	0.34
	$1.5n$	0.84	0.88	0.52	0.26
	$2n$	0.72	0.82	0.39	0.13
200	$0.5n$	1.00	0.99	1.00	0.79
	$0.75n$	1.00	0.98	0.98	0.59
	n	0.98	0.97	0.86	0.47
	$1.5n$	0.86	0.84	0.61	0.20
	$2n$	0.73	0.77	0.48	0.10

APPENDIX D. SIMULATIONS OF FULLY CONNECTED GRAPHS

We run simulations with fully connected graphs, as suggested by a reviewer. The linear coefficients are drawn uniformly from $\pm[.3, 1]$ and the errors are drawn from a standard Gaussian distribution. The results confirm the advantages of the proposed methods and are shown in Table 7. In general, the estimated graphs from the top-down and bottom-up procedure differ only slightly, and the values reported in the table differ in the 3rd or 4th digit.

TABLE 7. Fully connected setting

		Kendall's τ			Recall %			Flipped %			FDR %		
p	n	TD	BU	GDS	TD	BU	GDS	TD	BU	GDS	TD	BU	GDS
5	100	0.92	0.93	0.83	91	92	80	4	3	7	4	4	9
	500	0.99	0.99	0.97	98	98	98	1	1	1	1	1	1
	1000	1.00	1.00	0.99	99	100	99	0	0	1	0	0	1
20	100	0.98	0.98	0.62	74	74	45	1	1	9	1	1	17
	500	1.00	1.00	0.73	90	90	66	0	0	8	0	0	12
	1000	1.00	1.00	0.81	92	92	76	0	0	7	0	0	8
40	100	0.99	0.99	0.55	42	42	33	0	0	7	1	1	17
	500	1.00	1.00	0.62	50	50	49	0	0	8	0	0	14
	1000	1.00	1.00	0.67	52	52	59	0	0	8	0	0	12

APPENDIX E. AS INITIALIZER FOR GREEDY SEARCH

As suggested by a reviewer, we explore the performance of the greedy DAG search (GDS) algorithm initialized with the estimates from the proposed procedures. We run simulations with the same data as in Section 5.1. Tables 8 and 9 show averages over 500 random realizations for the top-down procedure (TD), the greedy DAG search with random initialization (GR), and the greedy DAG search with warm initialization (GW). The GR procedure is identical to the GDS procedure described in Section 5.1 and Peters and Bühlmann (2014). In the GW procedure, we initialize with the output from the top-down

TABLE 8. Low-dimensional dense settings

		Kendall's τ			Recall %			Flipped %			FDR %		
p	n	TD	GR	GW	TD	GR	GW	TD	GR	GW	TD	GR	GW
5	100	0.85	0.88	0.88	91	91	91	7	6	6	17	9	10
	500	0.98	0.98	0.99	99	99	99	1	1	1	4	2	2
	1000	0.99	0.99	0.99	99	99	99	1	1	1	3	1	1
20	100	0.92	0.61	0.94	85	62	90	3	13	3	32	43	15
	500	0.99	0.75	0.99	99	81	99	1	11	0	28	35	3
	1000	1.00	0.82	1.00	100	88	100	0	8	0	26	28	2
40	100	0.96	0.53	0.96	71	44	84	2	11	2	41	58	20
	500	0.99	0.59	1.00	96	63	100	0	14	0	41	57	4
	1000	1.00	0.64	1.00	97	71	100	0	14	0	40	57	2

TABLE 9. Low-dimensional sparse settings

		Kendall's τ			Recall %			Flipped %			FDR %		
p	n	TD	GR	GW	TD	GR	GW	TD	GR	GW	TD	GR	GW
5	100	0.87	0.88	0.87	91	90	91	6	6	6	16	9	10
	500	0.98	0.98	0.98	98	99	99	1	1	1	5	2	2
	1000	0.99	0.99	0.99	99	99	99	1	1	1	3	1	1
20	100	0.77	0.60	0.82	85	77	90	9	15	7	35	39	25
	500	0.96	0.77	0.98	98	89	99	2	10	1	19	26	8
	1000	0.99	0.81	0.99	100	90	100	0	9	0	14	23	4
40	100	0.72	0.47	0.79	81	72	89	10	20	7	38	54	36
	500	0.96	0.58	0.98	98	81	99	2	18	1	24	47	13
	1000	0.99	0.61	0.99	99	82	100	1	17	0	17	48	8

method, then search through a large number of graph neighbors ($k = 300$) at each greedy step. Since the GW procedure is supplied with a good initializer, we do not restart the greedy search after it terminates, while 5 random restarting with $k = p, 2p, 3p, 5p, 300$ is used in GR to insure performance. For simplicity, we omitted the experiment with the bottom-up procedure (BU).

Tables 8 and 9 shows that in all the settings, GW performs better than the other two methods, especially when p is large. For reference, the average run time in the dense setting with $p = 40$ and $n = 1000$ is 8 seconds for the top-down method, 4,500 seconds for GR, and 400 seconds for GW.

DEPARTMENT OF STATISTICS, UNIVERSITY OF WASHINGTON
E-mail address: wenyuc@uw.edu

DEPARTMENT OF MATHEMATICAL SCIENCES, UNIVERSITY OF COPENHAGEN
E-mail address: md5@uw.edu

BOOTH SCHOOL OF BUSINESS, THE UNIVERSITY OF CHICAGO
E-mail address: swang24@uchicago.edu